LREC 2018 Workshop

The First Financial Narrative Processing Workshop (FNP 2018)

PROCEEDINGS

Edited by

Mahmoud El-Haj, Paul Rayson, Andrew Moore

ISBN: 979-10-95546-23-8
EAN: 9791095546238

7 May 2018

Proceedings of the LREC 2018 Workshop
"The First Financial Narrative Processing Workshop (FNP 2018)"

7 May 2018 – Miyazaki, Japan

Edited by Mahmoud El-Haj, Paul Rayson, Andrew Moore

https://wp.lancs.ac.uk/cfie/fnp2018

Organising Committee

- Mahmoud El-Haj, Lancaster University, UK*
- Paul Rayson, Lancaster University, UK*
- Steven Young, Lancaster University, UK
- Andrew Moore, Lancaster University, UK
- Catherine Salzedo, Lancaster University, UK
- Stefan Evert, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany

*: Main editors and chairs of the Organising Committee

Programme Committee

- Andrew Moore, Lancaster University, UK
- Antonio Moreno Sandoval, Universidad Autónoma de Madrid, Spain
- Catherine Salzedo, Lancaster University, UK
- Denys Proux, Naver Labs, France
- George Giannakopoulos, N.C.S.R. Demokritos, Greece
- Mahmoud El-Haj, Lancaster University, UK
- Marina Litvak, Sami Shamoon College of Engineering, Israel
- Martin Walker, University of Manchester, UK
- Paul Rayson, Lancaster University, UK
- Scott Piao, Lancaster University, UK
- Simonetta Montemagni, Istituto di Linguistica Computazionale, Italy
- Stefan Evert, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
- Steven Young, Lancaster University, UK
- Vasiliki Athanasakou, London School of Economics, UK

Preface

Welcome to the First Financial Narrative Processing Workshop (FNP 2018) held at LREC 2018 in Miyazaki, Japan. The workshop will focus on the use of Natural Language Processing (NLP), Machine Learning (ML), and Corpus Linguistics (CL) methods related to all aspects of financial text mining and financial narrative processing (FNP). There is a growing interest in the application of automatic and computer-aided approaches for extracting, summarising, and analysing both qualitative and quantitative financial data. In recent years, previous manual small-scale research in the Accounting and Finance literature has been scaled up with the aid of NLP and ML methods, for example to examine approaches to retrieving structured content from financial reports, and to study the causes and consequences of corporate disclosure and financial reporting outcomes. One focal point of the proposed workshop is to develop a better understanding of the determinants of financial disclosure quality and the factors that influence the quality of information disclosed to investors beyond the quantitative data reported in the financial statements. The workshop will also encourage efforts to build resources and tools to help advance the work on financial narrative processing (including content retrieval and classification) due to the dearth of publicly available datasets and the high cost and limited access of content providers. The workshop aims to advance research on the lexical properties and narrative aspects of corporate disclosures, including glossy (PDF) annual reports, US 10-K and 10-Q financial documents, corporate press releases (including earning announcements), conference calls, media articles, social media, etc.

We received 11 submissions and selected 9 (81% acceptance rate) for presentation in the workshop. All papers were reviewed by two reviewers on average. As this is the first workshop of its kind we are delighted to have received enough interest to help increase awareness and build a network group of researchers working on financial narratives processing. The nine papers will be presented orally. The papers cover a diverse set of topics in financial narratives processing reporting work on financial reports from different stock markets around the globe presenting analysis of financial reports written in Arabic, English, Chinese and Portuguese. The quantity and quality of the contributions to the workshop are strong indicators that there is a continued need for this kind of dedicated Financial Narrative Processing workshop. We would like to acknowledge all the hard work of the submitting authors and thank the reviewers for the valuable feedback they provided. We hope these proceedings will serve as a valuable reference for researchers and practitioners in the field of financial narrative processing and NLP in general.

Mahmoud El-Haj, General Chair, on behalf of the organizers of the workshop May 2018

Programme

	Session 1
14.00 – 14.10	Opening Remarks by Workshop Chair (Mahmoud El-Haj)
14.10 – 14.30	Duygu Altinok The Ontology-Based Banking Chatbot
14.30 – 14.50	Oi Yee Kwong Analysis and Annotation of English-Chinese Financial Terms for Benchmarking and Language Processing
14.50 – 15.10	Mohammed Alshahrani, Fuxi Zhu, Mohammed Alghaili, Eshrag Refaee and Mervat Bamiah BORSAH: An Arabic Sentiment Financial Tweets Corpus
15.10 – 15.30	Matthew Purver, Aljoša Valentinčič, Marko Pahor and Senja Pollak Diachronic Lexical Changes In Company Reports
15.30 – 15.50	Damir Cavar and Matthew Josefy Mapping Corporate Filings to Knowledge Graphs using Deep NLP
16.00 – 16.30	**Coffe Break**
	Session 2
16.30 – 16.50	Chung-Chi Chen, Hen-Hsen Huang and Hsin-Hsi Chen NTUSD-Fin
16.50 – 17.10	Matthew Lamm, Arun Chaganty, Dan Jurafsky, Christopher D. Manning and Percy Liang QSRL: A Semantic Role-Labeling Schema for Quantitative Facts
17.10 – 17.30	Mahmoud El-Haj, Paul Rayson, Paulo Alves and Steven Young Multilingual Financial Narrative Processing
17.30 – 17.50	Martin Žnidaršič, Jasmina Smailović, Jan Gorše, Miha Grčar, Igor Mozetič and Senja Pollak Trust and Doubt Terms in Financial Communication
17.50 – 18.00	Closing Remarks by Workshop Chair (Paul Rayson)

Table of Contents

An Ontology-Based Dialogue Management System for Banking and Finance Dialogue Systems

Duygu Altinok
4Com Innovation Center
Berlin, Germany
duygu.altinok@4Com.de

Abstract

Keeping the dialogue state in dialogue systems is a notoriously difficult task. We introduce an ontology-based dialogue manager (OntoDM), a dialogue manager that keeps the state of the conversation, provides a basis for anaphora resolution and drives the conversation via domain ontologies. The banking and finance area promises great potential for disambiguating the context via a rich set of products and specificity of proper nouns, named entities and verbs. We used ontologies both as a knowledge base and a basis for the dialogue manager; the knowledge base component and dialogue manager components coalesce in a sense. Domain knowledge is used to track Entities of Interest, i.e. nodes (classes) of the ontology which happen to be products and services. In this way we also introduced conversation memory and attention *in a sense*. We finely blended linguistic methods, domain-driven keyword ranking and domain ontologies to create ways of domain-driven conversation. Proposed framework is used in our in-house German language banking and finance chatbots. General challenges of German language processing and finance-banking domain chatbot language models and lexicons are also introduced. This work is still in progress, hence no success metrics have been introduced yet.

Keywords: ontology, knowledge base, finance ontology, dialogue management, ontology based dialogue management, ontology based conversation, chatbot, virtual personal assistant, banking and finance virtual assistant, banking and finance chatbot

1. Challenges with German Language

German language processing is inherently challenging in general, independent of what the specific NLP task is. The main challenge is high variability in word forms due to inflections and compound words.

Finance domain lexicons include many compound words just like other technical domains in German. Here are some examples from our in-house banking and finance lexicon:

Verfügungsberechtigung	power to draw from an account
Sparkonto	savings account
Girokonto	checking account
Steuernummer	tax ID
Zahlungsverkehr	payment transactions
Zahlungsverkehrsraum	payment transactions area
Onlinebanking	online banking
Hypothekendarlehen	mortgage

Nouns, adjectives and verbs can be inflected according to gender, number and person. Rich word forms can pose a challenge language understanding components. In this paper, we focus on dialogue management. However, one should keep in mind that input to dialogue management components are provided by natural language understanding components.

Another practical issue in everyday written language is the *umlaut* (mutated vowels). Everyday informal written text includes umlauts replaced by their plain counterparts i.e. “Madchen, uber, schon” rather than “Mädchen, über, schön” etc. Especially in conversational interfaces, usage of umlauts reduce significantly due to English layout keyboards or just being lax about punctuation while typing quickly on a smartphone. In our opinion, umlaut-to-plain vowel replaced words are also a part of chatbot language models.

Morphologically rich languages have received considerable attention from many researchers. Many technical papers have been published to highlight the inherent technical difficulties in statistical methods e.g. MT, ASR-TTS, language models, text classification; practical solutions are offered in (Mikolov et al., 2016).

We overcome the challenges of rich German morphology using DEMorphy, an open source German morphological analyzer and recognizer[1]. Throughout our work, all lemmatizing and morphological analysis tasks are done by DEMorphy.

2. Introduction

Keeping dialogue state in conversational interfaces is a notoriously difficult task. Dialogue systems, also known as chatbots, virtual assistants and conversational interfaces are already used in a broad set of applications, from psychological support[2] to HR, customer care and entertainment.

Dialogue systems can be classified into goal-driven systems (e.g. flight booking, restaurant reservation) vs open-domain systems (e.g. psychological support, language learning and medical aid). As dialogue systems has gained attention, research interest in training natural conversation systems from large volumes of user data has grown. Goal-driven systems admit slot-filling and hand crafted rules, which is reliable but restrictive in the conversation (basically the user has to choose one of available options). Open domain conversational systems, based on generative probabilistic models attracted

[1] https://github.com/DuyguA/DEMorphy
[2] https://www.wysa.io

attention from many researchers, due these limits for goal-oriented systems (Serban et al., 2015; Li et al., 2017).

One problem with conversation is maintaining the dialogue state. This comprises of what the user said and how the chatbot answered, *what we're talking about* and which pieces of information are relevant to generating the current answer. Kumar et al. (2017) introduced neural networks with memory and attention (DMN). Done up to here DMN includes episodic memory and an attention module plus to a recurrent encoder decoder. DMN first computes question representation. Then the question representation triggers the attention process. Finally the memory module can reason the answer from all relevant information.
However, purely statistical approach has some drawbacks:

- statistical frameworks need **huge** training sets. Especially frameworks with many statistical components such as DMN, have a great number of parameters and are vulnerable to sparseness problems.
- Anaphora resolution is implicit. The anaphora resolutions go into neural network as implicit parameters, there's no direct easy way to see how the resolutions worked. Answers come through at least two distinct statistical layers, one encoder and one decoder at least. Thus there is no easy way to understand why a specific answer is generated and how the anaphora resolution contributed to the generation.

This paper addresses the dialog management. We will describe domain-driven ways to

- keep the conversation memory, both the user and the bot side
- make the anaphora resolution
- generate knowledge-based answers
- possibly contribute to *what to say next*
- integrate linguistic features into the context

NLU and answer generation modules will not be considered in detail in this paper. The focus is on how ontologies can be used to generate natural conversations. However we will present the outputs and presentations for clarity. The goal is here to improve quality of conversations via domain knowledge. This work is still in progress, hence we were not able to include performance metrics yet, given the difficult nature of evaluation of dialogue systems in general.

3. Proposed Framework

3.1 The Domain

This paper describes the methodology that is used in our in-house banking and finance chatbots. We chose the banking and finance domain due to rich set of products and specificity of proper nouns, named entities and verbs; high potential for disambiguate the context and drive the conversation. Though development was made on banking and finance domain, framework is applicable to other highly specific domains such as medicine, law and online-shopping.
Banking and finance conversations, either seeking financial advice or inquiring banking products; aim to *get information* rather than to accomplish a goal. Hence throughout this work, conversations are not goal-oriented but rather domain-driven. Users usually do not aim to achieve a well-defined goal, i.e. book a table at a restaurant or book a flight. The financial chat is mostly about getting information about rates, prices, investment instruments and sometimes about picking a suitable option i.e. *purchase advice*. Purchase rarely happens immediately after the advice. The banking chat can include both asking for information about account types, credit card types, their yearly fees and rates i.e. the banking products; or making money transfer, asking for account balance, viewing account activity i.e. achieving a certain goal. We will address these issues in next chapters.

3.2 Ontology-based Dialogues

Ontology-based conversating is indeed a way of domain-driven conversation. We used ontologies in our work for two purposes:

- to store the knowledge
- to navigate through the domain

The knowledge base component takes part in many dialogue systems. After the NLU component turns queries into a logical form, next step is to interrogate the knowledge base for the answer generation. In banking and finance domain ontologies, one potentially needs to store

- a range of banking products: credit, credit card, debit card…
- attributes of these products: general conditions, rates and fees
- range of banking services: ATM, money transfer...
- attributes of the banking services: ATM points, branch addresses...

For instance, in following conversations (Fig.1 & 2) it is not possible to generate an answer *without* knowing the product. The knowledge base module provides information to the answer generation module.

User: Was kostet die Bestellung im Internet? (How much an online purchase costs?)
Bot: Es entstehen keine Kosten für Sie.. (It costs you nothing.)

Figure 1: Dialogue Example

User: Haben Sie einen Geldautomaten in Berlin-Tiergarten? (Do you have an ATM in Berlin-Tiergarten?)
Bot: Ja, …….. (Yes, ….)

Figure 2: Dialogue Example

In our work the ontology also drives the conversation: *keeps the context*, provides a basis to the *anaphora resolution* and possibly produces *what to say next*. The knowledge base component and the dialogue management components coalesce.

User: Ich möchte eine Kreditkarte bestellen. (I want to get a credit card.)
Bot: Wir bieten Ihnen eine Mastercard an. (We have a Mastercard offer.)
User: Was kostet die? (What does that cost?)
Bot: …….
User: Ist eine Internetbestellung möglich? (Is online purchase possible?)
Bot: Ja, klicken Sie einfach auf "Mastercard beantragen" unter … (Yes, please click…)

Figure 3: Dialogue Example

In Figure 3, die refers to the Mastercard product. One needs to remember from the previous line that we were speaking about the Mastercard. The sentence "Was/PWS kostet/VVFIN die/PDS ?" is a wh-question sentence; also does not have any noun phrases, includes a substitution demonstrative pronoun and *nothing else*. Worse, the verb attaches to both the interrogative pronoun and the demonstrative pronoun. From the linguistic information so far, it is clear that one needs to resolve what die refers to. "*Ist eine Internetbestellung möglich*?" involves a more difficult kind of anaphora, an *implicit* one. There is a noun phrase "Internetbestellung", but it's a property (in semantic and ontological sense) of the Mastercard indeed; not any other banking product. If one *knows* that current *Entity of Interest*[3] is the Mastercard and the Mastercard has a property related to purchasing channels, then the resolution becomes clear.
We took FIBO as our guideline and created our own banking and finance domain ontologies in German. Including the whole ontology is not possible, thus here in Figure 4 we present a small section devoted to the credit products, generated by VOWL[4] within Protégé [5]:

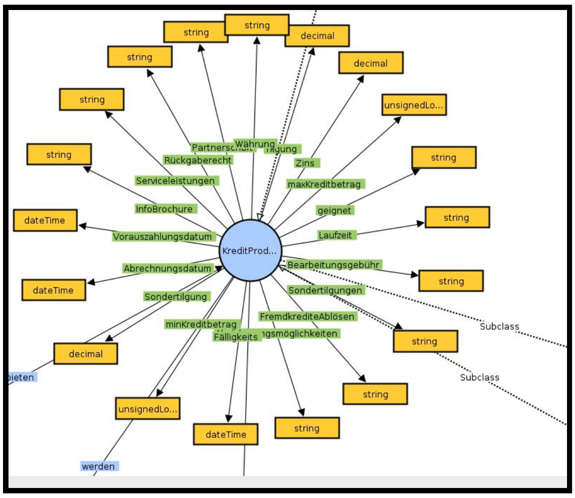


Figure 4: Ontology Example generated by VOWL within Protégé

[3] will be explained soon
[4] http://vowl.visualdataweb.org/protegevowl.html
[5] https://protege.stanford.edu/

Nodes i.e. the classes, consist of noun phrases (corresponding to abstract terms) and attributes (object and data properties) can be either noun phrases or verb phrases.
Domain ontology is an abstract schema. When we want to store information about a specific financial institute and deploy the corresponding chatbot, we instantiate the ontology with corresponding individuals. For instance, if we want to chat about 4Bank and its products, we instantiate the *KreditProduct* class by the *4Kredit* individual and the *Zins* with the *0.23* individual.
With the domain knowledge, we track significant entities in the conversation. The Entity of Interest notion includes *all classes* in the domain ontology (products and services). Noun phrases in the user inquiries fall into two categories, either a term from the ontology or not. Ontology terms are used to drive the conversation, the rest is kept as the user's conversation memory. This way we unite the linguistic knowledge, output of the NLU module with the domain knowledge. In every step of the conversation, the bot keeps track of user speaks about which product/service explicitly. Difference between the neural dialogue managers and the ontology-based approach becomes clear here, ontology-based DM keeps conversation memory **explicitly** throughout the conversation. In our framework, conversation memory and attention is provided via pointers to significant EOIs and keeping track of already visited nodes.

4. Implementation Details

Methods of resolution and disambiguation include blending linguistic information, several layers of semantic similarity and statistical methods together with ontologies. We will include the semantic representations of the user inputs to clarify the methodology. Throughout this paper, we will use 4Bank Ontology, our in-house banking and finance ontology.

4.1 The Domain Ontology

4Bank ontology contains 76 distinct classes, 37 object properties and 534 data properties. The ontology classes correspond to products, services, legal terms, contracts, customers as well as more abstract concepts such as "finance product". Abstract concepts usually serve as superclasses to more specific concepts. The object property (relation) names are verbs, the data property (attribute) names can be both names are verbs. The class names are always noun phrases; either proper nouns correspond to the products/services, or common nouns for more general concepts.
Reason for relatively less number of relations is that, many relations map superclasses of products/services to again superclasses. Figure 5 shows some examples of the classes from the ontology, together with their subclasses and identical classes:

Figure 5: Classes of the Domain Ontology

Underlying undirected graph is dense and connected. Average degree of the underlying undirected graph is 3.1, maximum out- and in-degrees of underlying directed graph are 4 and 4, respectively (only the object properties are counted).

4.2 Architecture Overview

An overall architecture of our finance-banking FAQ chatbot is shown in Figure 6. First, the system transforms user's text message into a semantic representation. Then the semantic information is processed together with the current dialogue state to generate an appropriate answer.
We define a chat session as $S=(C, (Q_1, A_1, Q_2, A_2, ...))$; a finite sequence of user messages (Q_i) and bot answers (A_i), together with the current context (C).
During runtime, every new user message creates a new query object (throughout this paper we will use the user message and the corresponding query object interchangeably). The query object contains linguistic, semantic and syntactic information parsed from the user message. Only one context object is alive per chat session and the context object changes state with user queries.

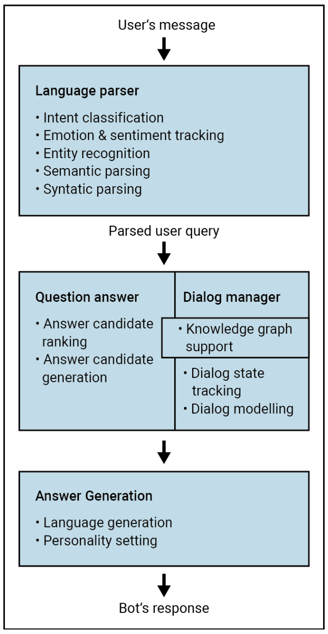


Figure 6: Architecture of the Banking FAQ Bot

4.2.1 Natural Language Understanding

The NLU component parses semantic and syntactic information from user queries. For the sake of simplicity, we will provide examples of single sentence user messages. Multi-sentence user messages are handled similarly. A query object has the following attributes: the sentence type, the intent, keyphrases, noun phrases, verb phrases, POS tags and the length as the number of tokens.
There are four sentence types recognized by our system: Greeting, Chitchat, Action and Ordinary. Each type corresponds to a subclass of the query object. Some examples for each sentence types are given in Figure 7.

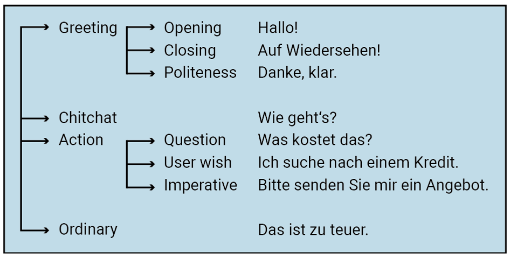


Figure 7: Sentence Types

Different subclasses might admit more fields. For instance, Action sentence has an "agent" and Question sentence has a "question type" (one of the following: yesno_q, wh_q or misc.)
Action sentences represent sentences that corresponds to an action, either from the user or from the bank. Action sentences also admits "agent"s. In our work, agent has a different meaning than its usual linguistic meaning. The agent is the performer of the action; either the user, the bank or a third person. See the examples:

User: Hallo!
{ 'GTYPE': 'opening',
'IS_UNINFORMATIVE': True,
'LENGTH': 1,
'SENTENCE': Hallo,
'SENTENCE_TYPE': 'Greeting'}

Figure 8: Example Greeting Sentence

User: Können Sie mir Unterlagen zukommen lassen?
{
'AGENT': '2nd',
'INTENTS': ['send_document'],
'IS_UNINFORMATIVE': False,,
'KPHRASES': ['unterlagen', 'unterlagen zukommen lassen'],
'LENGTH': 6,
'NOUN_PHRASES': ['unterlagen'],
'POS_TAGS':"VMFIN PPER PPER NN VVINF VVINF $.",
'QTYPE': 'yesno_q',
'QWORD': 'können',
'SENTENCE': Können Sie mir Unterlagen zukommen lassen?,
'SENTENCE_TYPE': 'Action',
'TYPE_ATTR': 'question',
'VERB_PHRASES': ['zukommen lassen']}

Figure 9: Example Question Sentence

Greeting and Chitchat sentences short-circuit to the answer generation module without further processing.
All syntactic parsing tasks were done by SpaCy. The semantic parsing comes in several layers, the question type, the keyphrases, the intent and the semantic similarity score based on word vectors during the graph search.
Keyphrase Ranking Keyphrase candidates are extracted in the NLU module by linguistic features. In our work we included noun phrases and noun phrases followed by verb phrases as keyphrase candidates. The IR module ranks candidates and generates the keyphrases. We designed an Okapi BM25 based keyword ranking algorithm in an unsupervised fashion. The ranking algorithm is trained on our in-house synonyms corpus. The corpus contains 120 questions in total, annotated with their rephrased (semantically similar) counterparts and keyphrases. See an example entry from the dataset:

Ich möchte meinen Kredit gerne erhöhen. Können Sie mir Unterlagen zukommen lassen?
Kredit erhöhen, Kredit aufstocken, Krediterhöhung, Kreditaufstockung, Unterlagen zur Kreditaufstockung, Unterlagen zur Krediterhöhung
Wie kann man den Kredit erhöhen
Wie kann ich meinen Kredit aufstocken
Wie kann ich den bestehenden Kredit aufstocken?
Ich möchte mein Darlehen erhöhen, Wie erhöhe ich mein Darlehen
Wie stockt man sein Darlehen auf
Senden Sie mir Unterlagen zur Krediterhöung
Schicken Sie mir Unterlagen zur Kreditaufstockung
Können Sie mir Formulare zur Krediterhöhung zukommen lassen

Intent Classifier The intent classifier is also trained on the synonyms dataset. The classifier is again subword based, we used fastText library[6] for the classification.

[6] https://pypi.python.org/pypi/fasttext

4.2.2 Context Object

The context object represents the current state of the conversation. The context object class has the following fields: curr_prod, curr_prod_indiv, curr_inode, curr_leaf and message_index.
The first four fields are pointers to the ontology nodes. The current product always refers to the product that has been the current chat topic and the corresponding curr_prod always points to one of subclasses of the Product or Service nodes. The current product individual references a specific instance of the current product, for instance if the current product class is the credit card, the current product individual can be Mastercard Gold or Mastercard Standard. The current inode points to the attribute of the current product class or individual spoken, the current leaf points to the most recently fetched leaf of the current inode if applicable. The current leaf can be both a class or an individual of the ontology. All four pointers reference significant EOIs of the chat; during the chat's lifetime we hold pointers to one individual, two conceptual classes as well as keeping track of the visited nodes.
A context object begins its lifetime with a null curr_prod and curr_prod_indiv. In every step of the conversation, if the current product or the current product instance is ambiguous or not present; the bot asks user for clarification or lists the available options to direct the conversation. See for instance:

U: Ich interesse für mich für ein Konto curr_prod←Konto
B: Wir bieten Girokonten und Sparkonten an.
curr_prod has 2 subclasses Girokonto and Sparkonto. The bot lists all.
U: Was kostet das Girokonto? curr_prod←Girokonto
curr_prod is now Girokonto and has 2 individuals. The bot lists both.
B: Das Superkonto kostet 60 Euro pro Jahr und das Standard4Konto is für Sie kostenlos.

Figure 10: Dialogue Example

This way we keep track of which product class, if applicable which specific product and which attributes are being discussed. It is also possible to conversate without referencing a specific individual, for instance:

U: Bieten Sie Leasing an? (Do you provide leasing?)
B: Nein, wir bieten nur Privatkredite an. (No, we provide only personal kredits.)

Figure 11: Dialogue Example

4.3. Context Resolution Algorithm

In this, section we introduce our graph-based context resolution algorithm. The CRO algorithm is combination of depth-first search and breadth-first search on the ontology as a directed graph.

4.3.1 Graph Search

We recall the example from Figure 12 in Section 3.2:

User: Ich möchte eine Kreditkarte bestellen. (I want to get a credit card.)
Bot: Wir bieten Ihnen eine Mastercard an. (We offer Mastercard.)

Figure 12: Dialogue Example

A short version of the output of NLU module for the first user input is below. "*Ich möchte eine Kreditkarte bestellen*" has only one noun phrase and "*Kreditkarte*" ranks as a keyphrase. Please see the output below:

ActionS(type="user-wish",agent="customer",kphrases=["Kredit-karte", "Kreditkarte bestellen"], noun_phrases=["Kreditkarte"])

Figure 13: Output Example

Next step is to semantically match the noun phrases to the ontology. We search for the *Kreditkarte* in our 4Bank ontology and locate the class in the schema (see Figure 14):

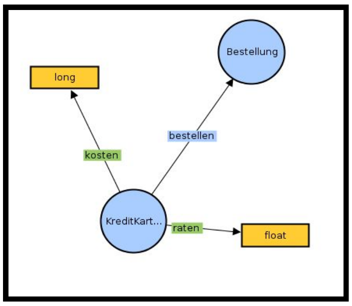


Figure 14: Matching Keyword *Kreditkarte* in the Ontology

Then *Kreditkarte* immediately becomes the curr_prod as it is a subclass of Finanzprodukt class. *Kreditkarte* has only one instance and it is called *Mastercard*. We feed this information to the answer generation module to generate an answer. Dialogue manager holds a pointer to this node to keep the context.
The graph search also includes word vector based semantic distance to match the synonyms correctly. For instance the keyphrase "Online" correctly matches to the node "Internet". We used fastText vectors[7] due to their superior performance on the German word forms.

4.3.2 Resolution

The context resolution algorithm resolves which specific product/service, concept or attribute the user is speaking about.
In every step of the context resolution algorithm,
(a) if a product individual name appears in the user query, we set curr_prod_indiv to this node and fetch the questioned attribute. We can directly fetch the answer from the ontology without any more resolution as the context is straightforward. Also the curr_prod is set the product individual's class.

User: Was ist der Zinssatz bei 4Kredit? (What is the interest rate of 4Kredit?) curr_prod←Kredit, curr_prod_indiv←4Kredit

Figure 15: Dialogue Example

(b) if the user mentions a product class, either it is the class name for a product individual that (s)he mentioned before or (s)he changes the context to speak more about this product class. In the latter case, we set curr_prod to the class name. In the first case, the class name refers to an already mentioned individual; hence curr_prod is already set to this class name.

User: Hallo curr_prod = null
Bot: Hallo
User: Ich bekomme einen Kredit. curr_prod←Kredit
Bot: Wir bieten Ihnen den 4Kredit an. (We provide 4Kredit.)

User: Hallo curr_prod = null
Bot: Hallo
User: Was ist der Zinssatz bei 4Kredit? (What 4Kredit's interest rate?)
curr_prod←Kredit, curr_prod_indiv←4Kredit
Bot: 0.23
User: Was kostet der Kredit? (Here, Kredit refers to the individual 4Kredit. No updates to curr_prod or curr_prod_indiv are necessary)
Bot: 4Kredit kostet ….

User: Moin moin
Bot: Hallo!
User: Was ist der Zinssatz bei 4Kredit?
curr_prod←Kredit, curr_prod_indiv←4Kredit
Bot: 0.23
User: Was kostet eine Kreditkarte? (What does a credit card cost?)
(Here a context change occurs as the curr_prod_indiv 4Kredit is not an instance of Kreditkarte class.) curr_prod←Kreditkarte

Figure 16: Dialogue Examples

(c) the user can query without specifying a product name or a product individual name. This usually happens at the beginning of the conversation:

Bot: Hallo curr_prod = null
User: Hallo, was kostet das? (Hello, what does that cost?)
curr_prod = null, curr_prod_indiv = null
Bot: Wir bieten Kredite, Konten, … an. (We offer credits, accounts..)

Figure 17: Dialogue Example

In this case we list the available subclasses of the product/service classes and ask the user which one(s) (s)he is interested in.
(d) the user not always refers to a product/service explicitly. In this case the intent classifier comes into the play:

User: Ich möchte meinen Hausbau finanzieren. Kannst du mir bitte helfen? (I want to finance my house construction. Would it be possible to help me?) curr_prod←Hypothek

Figure 18: Dialogue Example

[7] https://github.com/facebookresearch/fastText/blob/master/pretrained-vectors.md

4.3.3 Pronominal Anaphora Resolution

We will continue with the example from section 4.3.1. We already make a graph search to locate the Kreditkarte node in the ontology and generated an answer to the first user question.

User: Ich möchte eine Kreditkarte bestellen. (I want to get a credit card.)
Bot: Wir bieten Ihnen eine Mastercard an. (We offer Mastercard.)
User: Was kostet die? (What does that cost?)
Bot: 80 Euro jährlich (80 euros annually.)

Figure 19: Dialogue Example

Next user input is "*Was/PWS kostet/VVFIN die/PDS ?*". Please see the semantic representation and the result of the dependency parse[8] in Figure 20 below:

Q(type="wh-q", agent="no", qword="was")

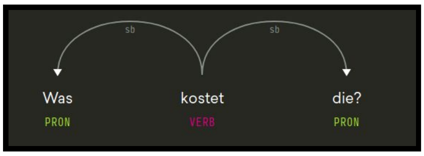


Figure 20: Example of the User Input

There are two pronouns in the sentence, an interrogative pronoun *Was*, one substitution demonstrative pronoun *die* and a verb. There are no noun phrases in the sentence. It becomes clear that this question does not carry much *information* itself and die needs to be resolved. In the current context, we have only one past EOI i.e. the "*Kreditkarte*". After lemmatizing with DEMorphy, "kostet" becomes "kosten" and resolves to the "kosten" relation in the domain ontology, please see Figure 20. The KB module fetches the individual related, "*80 Euro jährlich*" and inputs it to the rest of the pipeline.

4.3.4 Implicit Anaphoras

It is not always clear when one should do any anaphora resolution at all. Following conversation in Figure 19 includes two such user inputs, where necessity for any anaphora resolution is not immediately recognizable.

User: Kann ich einen Kredit aufnehmen? (Can I get a loan?)
Bot: Die 4Bank bietet Ihnen mit 4Kredit einen schnellen und unkomplizierten Weg zu Ihrem Kredit. (4Bank offers 4Kredit, a fast and uncomplicated way to obtain a credit.)
User: Ist eine Internetbestellung möglich? (Is online purchase possible?)
Bot: Ja, klicken Sie …. (Yes, please click…)
User: Am Telefon? (Purchase by phone?)
Bot: Rufen Sie uns unter 05113003990 an und vereinbaren Sie einen Termin für Ihre telefonische Kreditbestellung. (Please phone us....)

Figure 21: Dialogue Example with Ambiguous Anaphora

Question in Line 3 contains *information*, indeed EOI that appears in our banking and finance lexicon:

Q(type="yesno-q",agent="no",qword="ist",kphrases=["Internet-bestellung"])

Figure 22: Semantic Parse of the User Input

Obviously this question is about purchasing online but it is still not clear only from this line, user is interested in purchasing *which* product online. Should it be a credit or an insurance? If there are several products that can be purchased online, it is not possible to generate an answer only based on this question. Here the context memory i.e. the current product ("*Kredit*") comes into play. The context resolution algorithm first performs a BFS to locate Bestellung node. Head of a German compound is its last word, previous words might qualify the head word in different ways. Internetbestellung has Bestellung as the head word. The current inode pointer is set to the Bestellung node and another BFS is performed to locate Internet subclass to mark Internet as the current leaf. Then the answer generator can take into account that (1) it is a yes-no question and (2) Bestellung class has a subclass called Internet to generate the answer "yes".

At first glance it is not clear that one needs a context disambiguation at all. However, these type of ambiguities might appear even at the very final stages of the context disambiguation computations.

It is also possible that there are more than one EOIs in the current context memory. Line 5 is an example for such situation, current EOIs are the *Kredit* and the Internetbestellung. Telefon indeed is a "shorthand" for the *Telefonbestellung* (purchase by phone). Here Telefon semantically attaches to the both EOIs, the *Kredit* (as an attribute) and the Internetbestellung (not directly, via analogy) (see Fig. 22).

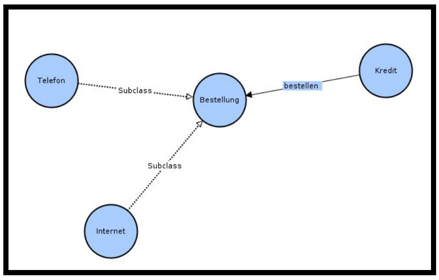


Figure 23: Example of EOI Ambiguity

The CRO handles this case as follows: The current leaf is the Internet, the current inode is the Bestellung. Previous keyword which resolved to Internet node was Internetbestellung, the concatenation of the node itself (Internet) and its parent node (Bestellung). Then, the context resolution algorithm first looks for a Telefon subclass under the Bestellung node. If it was the case no Telefon node was located under the current inode, the CRO would perform another BFS under the current

[8] generated by spaCy and displaCy, https://spacy.io/

product node, the *Kredit*. The resolution algorithm favors more specific anaphoras and the most recent EOI.
This way the CRO resolves the implicit semantic analogy that Telefon in Line 5 refers to Telefonbestellung and finally fetch the correct node.
We already pointed out that our framework introduces memory and attention *in a sense*. Dynamic memory networks with attention *selects* relevant previous episodic memory members, the *relevancy* is computed by the attention module. Our framework offers a more explicit solution than the statistical methodology, since at every step of the conversation the DM stores EOIs explicitly.

4.3.5 Long Sentences with Compound Intents

Multi-sentence user messages and long sentences may contain more than one intents and EOIs. For instance:

> Sind meine Unterlagen schon bei Ihnen eingegangen? Falls ja, wo erfahre ich den aktuellen Bearbeitungsstand? (Have you already received my documents? If so, where do I get the current processing status?)

> Wo und wie bekomme ich eine Kredit? (Where and how can I get a credit?)

> Was kostet eine Kredit und zu welcher Laufzeit? (What does a credit cost and what is the loan period?)

Figure 24: Dialogue Examples

In the first sentence, there is one EOI, the *Unterlagen* and two intents: to check out if the documents arrived and to learn the processing status. The current product or the current product individual does not need any updates. The CRO sets the curr_inode to Unterlagen and answer generation module generates two answers for the two different intents and concatenate them to generate a composite answer.
The second sentence involves a product class name, *Kredit*. The CRO sets the current product to *Kredit* and the answer generation module generates an answer for the two intents jointly unlike the first sentence.
A product class name (*Kredit*) and two different attribute names of this class (*kosten*, *Laufzeit)* comes together in the third sentence. The context resolution rule (b) from the Section 4.3.2 applies here, we fetch two attributes the *kosten* and the *Laufzeit* iteratively and pass to the answer generation module.
However, even the sentence is long, it is very unusual to refer to two or more product classes and instances. For instance, such a sentence

> Was kostet ein Kredit und eine Kreditkarte? (What is the cost of a credit and a credit card?)

Figure 25: Dialogue Example

with two distinct product class names occurred almost none times in our evaluation corpus.

4.3.6 Cycle Prevention

While traversing the graph, the context resolution algorithm marks the visited nodes to prevent the future cycles. Users may ask the same questions that they previously asked, however the bot shouldn't make the same offerings that it already made. For instance:

> Bot: Hallo!
> User: Was bieten Sie an? (What services do you offer?)
> Bot: Wir bieten Kredite, Konto und Kreditkarte für Sie an.
> User: Was kostet der Kredit? (How much does the credit cost?)
> (Here the CRO marks Kredit node as visited.)
> Bot: …….
> User: …..
> Bot: ….
> User: Welche anderen Produkte haben Sie? (What else do you offer?)
> (Kredit node is marked as visited. The BFS skips this node.)
> Bot: Möchten Sie unseren Kreditkarten oder Kontobereich besuchen?
> User: ….

Figure 26: Dialogue Example

The CRO also marks the edges (data properties) that have already lead to fetching an answer:

> B: Hallo
> U: Ich brauche einen Kredit. (I need a credit.)
> B: Wir bieten Ihnen 4Kredit an. (We offer 4Kredit.)
> U: Was ist die Laufzeit? (What is the loan period?)
> (Here Laufzeit edge is marked as used)
> B: Wir bieten Laufzeiten von 12 bis 84 Monaten an. Wie hoch ist Dein Kreditwunsch? (The loan period is between 12 and 84 months. What amount are you asking for?)
> (Since we already fetched the Laufzeit edge, we do not speak about this attribute anymore. The BFS skips Laufzeit edge and the DM generates *what to say next* from another attribute: the kredit amount).

Figure 27: Dialogue Example

The current product node and previous curr_prod nodes are always marked as visited. Marked product nodes and attributes together prevents back edges, hence cycles in the chat session.

5. Conclusion and Future Work

This paper presents methods of ontology-based ways of dialogue management in the banking and finance area. Though this work has not been tested extensively yet, the current achievements are promising. The future work includes finishing the framework and the chatbot development. Besides, it is even more important that we will evaluate the whole chatbot success by the BLEU score. We also aim to introduce a metric for the evaluation of the dialogue manager module. Considering the fact that the quality of the ontology directly affects the success of the DM, introducing an evaluation metric is a separate application on its own.

6. Bibliographical References

Asri L.E., Schulz H., Zumer J., Harris J., Fine E., Mehrotra R. and Suleman K. (2017). Frames: A Corpus for Adding memory to Goal-Oriented Dialogue Systems. CoRR abs/1704.00057

Bordes A. and Weston J. (2016). Learning End-to-End Goal Oriented Dialogue. CoRR abs/1605.07683

Kumar, A., Irsoy O., Su J., Bradbury J., English, R., Pierce, B., Ondruska P., Gulrajani I. and Socher R. (2015). Ask Me Anything: Dynamic Memory Networks for Natural Language Processing. CoRR abs/1506.07285

Li X., Chen Y.-N., Li L. and Gao J. (2017). End-to-End Task Completion Neural Dialogue Systems. CoRR abs/1703.01008

Mikolov T., Bojanowski, P., Grave E. and Joulin A. (2016). Enriching Word Vectors with Subword Information. CoRR abs/1607.04606

Serban, I.V., Sordoni, A., Bengio, Y., Courville A.C. and Pineau, J. (2015). Hierarchical Neural Network Generative Models for Movie Dialogues. CoRR abs/1507.04808

Schulz H., Zumer J. and Sharma S. (2017). A frame Tracking Model for Memory-Enhanced Dialogue Systems. CoRR abs/1706.01690

Vinyals O. and Quoc V. Le (2015). A Neural Conversation Model. CoRR abs/1506.05869

Analysis and Annotation of English-Chinese Financial Terms for Benchmarking and Language Processing

Oi Yee Kwong
Department of Translation
The Chinese University of Hong Kong
oykwong@arts.cuhk.edu.hk

Abstract

This paper reports on our ongoing work in annotating bilingual (English-Chinese) terminology in the financial domain. Arising from a larger project to produce a benchmarking dataset for evaluating term extraction tools, the resulting language resource is expected to be of use in financial narratives processing. The study will make available a gold standard to translators and researchers, especially for the former to leverage in the evaluation of commercial term extraction tools. To accommodate the diverse interests of both end users and researchers from multiple perspectives, an analysis of existing terminological resources for translators was done. Based on the linguistic properties observed, a set of term annotation guidelines was formulated for marking up bilingual financial terms in a corpus, for systematic selection of terms according to various criteria and expectation. The resulting dataset will fill the gap for term extractor evaluation for which bilingual data are lacking, and serve as a shared and transparent evaluation standard to help enhance the mutual understanding between computational terminologists and translators. In a wider context of natural language processing, bilingual terms extracted from a variety of financial texts are anticipated to be of help for information mining especially from regularly updated and structurally repetitive documents.

Keywords: financial terminology, bilingual term extraction, benchmarking dataset, term annotation

1. Introduction

Terminology management is a core function of computer-aided translation, and automatic term extraction is a commonly affiliated component of it. Off-the-shelf commercial term extraction systems available to translators, as products of software development instead of prototypes in academic research, are often packaged with many user-friendly features, but the opaque operation mechanism (or the algorithm) unknown to the users often leaves them with unrealistic expectation and misunderstanding. Such software tools thus do not always invite positive comments when they enter into the market. This is especially evident when a tool claims to be language independent, but when it is applied to a distant language pair, such as English and Chinese, it often turns out to perform much more poorly than its promotional demo may show, for instance, between English and French. User feedback found in classrooms and on the internet often echoes this observation. The situation is therefore not quite in concord with the many encouraging results reported from studies on automatic term extraction all along. Such an undesirable scenario, to a certain extent, is the result of little mutual understanding between computational terminologists and translators, and the lack of a shared and transparent evaluation standard. These issues merit attention, reflection, and resolution.

The project starts with term annotation in the financial domain for various reasons. Despite the availability of English-Chinese bilingual financial glossaries from various sources, like those mentioned in Section 4.1 below, they might not be ideal for benchmarking in term extraction especially for translators, as they are not tailored for specific sets of source and target texts that come with the translators as end users of the term extraction systems. Second, the bilingual terms in existing glossaries may not cover the precise translation equivalents used in the bilingual texts of specific companies and organisations (such as different banks). Third, the "termhood" as considered in existing glossaries is mostly from domain practitioners' perspective, which may or may not be the same as that from terminologists and translators. In addition, financial terminology is particularly crucial in the Hong Kong context given its long established status as one of the world's leading financial centres, and its continuing economic development with increasingly closer relations with Mainland China. New concepts and thus new terms keep emerging, and their timely and accurate mining would be of great help for professional translators.

2. Users' Concerns and Expectations

As end users of commercial term extraction systems, most translators could only judge a system by their own first-hand experience. Of most relevance to them is probably the preparation of bilingual input text for the extraction process and the validation of the term candidates subsequently returned by the system.

2.1. Input Text

As Blancafort et al. (2013) pointed out, computer-assisted translation suffers from the terminology bottleneck, and bilingual terminologies generated with statistical machine translation toolkits require parallel corpora. Scarcity of domain-specific parallel corpora is the major hindrance. As they noted, commercial tools only handle parallel corpora. Even though most off-the-shelf tools nowadays often allow a variety of file formats such as .rtf, .doc, .pdf and .xml, in addition to the conventional plain text files, it is impractical to expect a professional translator or translation student to supply a very large corpus, not even for monolingual materials, which may only be more readily available to computational linguists. Nevertheless, the extraction performance would depend on the size of the input text to a certain extent, especially if the extraction is statistically based, although the actual impact is often opaque to the user. For bilingual term extraction, users often have extra work in preparing bilingually aligned text for input. The aligned text will have to be in a specific format such as the Translation Memory eXchange (.tmx) format. For instance, bilingual texts in other formats like .doc will only be treated as monolingual ones by SDL MultiTerm Extract. To this end, the general users might need to make use of other tools (e.g. SDL Trados) to do the alignment and save/export the aligned sentences in the required format beforehand, which is a time-consuming step, while the more automatic alignment toolkits like GIZA++ (Och and Ney, 2003) would simply be deterrent to them. The extra pre-processing steps thus tend to keep users away, especially those who are less comfortable with computer-aided work.

2.2. Extraction Algorithm

Term extraction approaches are generally categorised as linguistic (e.g. Bourigault, 1992), statistical (e.g. Daille and Morin, 2005), or hybrid (e.g. Daille, 1996; Drouin, 2003). For bilingual term extraction, parallel corpora would be most preferred, but given the scarcity of parallel corpora, often it might have to make do with comparable corpora (e.g. Laroche and Langlais, 2010). The TTC platform, for instance, provides a whole pipeline of tools for terminology mining from comparable corpora for seven languages, including English and Chinese (Blancafort et al., 2013). Terms are separately extracted from monolingual corpora first and then bilingually aligned based on context of occurrence and compositionality (Daille, 2012).

To most translators as end users, the extraction step in commercial systems is just a click of button. The details of the algorithm used by a particular tool are usually unknown to them, that is, a black box. Although many a time users might be told that a certain tool makes use of a statistical algorithm to come up with the term candidates, which is almost the norm of modern term extractors, the algorithms adopted in individual systems could have different degrees of sophistication, and this is often at least partially disclosed from the results they generate. The more computer-literate users are often able to get a clue from the output to reverse engineer the mechanism by which the tools work. For example, some tools relying primarily on simple n-gram frequencies without paying much attention to linguistic validity (e.g. phrasal structures) may output incomplete or ungrammatical word strings among the suggested terms. Users' evaluation of the systems is usually impressionistic, based on their overall experience with the user interface, functionalities, effort needed for validation, and compatibility with their working translation environment, amongst other criteria. For example, Xu and Sharoff (2014) evaluated various term extraction tools working on comparable corpora, and although their performance was at most mediocre, especially on Chinese, student interpreters still found the low precision tolerable.

3. Modes of Evaluation

It is this last point regarding evaluation measures that we find mostly responsible for the gap between users and researchers. For computational terminologists, apart from qualitative comparisons among term extraction systems (e.g. Cabré et al., 2001), evaluation may also rely on human judges to go through the system-generated term candidate list (e.g. Fulford, 2001) or compare the term list against an existing term bank (e.g. Drouin, 2003). Nowadays it is often preferred to have system performance to be objectively measured with reference to some benchmarking data, by precision and recall, as is popularly done for many other natural language processing tasks. However, the reliability and validity of such quantitative measures are based on the assumption that a clear task definition exists. As noted by Bernier-Colborne and Drouin (2014): "Whereas other natural language processing tasks have well-defined evaluation schemes and benchmarks, the question of how to evaluate TEs [term extractors] remains unresolved. Evaluations are regularly reported in work on term extraction, yet the methodology varies from one work to the next, such that comparisons are hard to establish." (p.51) For the term extraction task to be well-defined, one must state precisely what counts as a term and thus which expressions should or should not be extracted. In fact, given the different backgrounds and expectations, evaluation criteria also vary, and many reference standards in different studies may only be ad hoc term lists drawn up without adequate systematic control. It is apparent that a translator's evaluation of a term extraction tool usually has no reference to benchmarking data. For instance, in a preliminary study on terms extracted by SDL MultiTerm Extract from a small amount of financial texts (from the annual reports of a bank), we compared the so-called unmatched items against a translation student's so-called "gold standard". It was observed that among the "noise", many are grammatical linguistic expressions (e.g. noun phrases, verb phrases, prepositional phrases) and some are obviously genuine financial terms; and among the expected but unmatched items, some are actually partially extracted already while others might be considered semi-technical terms (Kwong, in press). While translators' concerns are multifarious, including but not limited to system performance (e.g. accuracy of candidates), software design (e.g. user-friendliness), and very importantly compatibility with their own expectation, which lead to an overall perception of a system, access to a gold standard is nevertheless desirable for end users. After all, the validation process is often the decisive factor for whether a translator will find the term extractor a help or a nuisance.

4. Toward a Gold Standard for Bilingual Term Extraction

Defining the gold standard for term extraction can sometimes be tricky. The main problem has to do with spelling out the criteria for the selection and annotation of terms systematically. As Vivaldi and Rodríguez (2007) remarked, "there is low agreement between terminologists and domain experts on what term candidates should be treated as terms" (p.244). Estopà (2001) also reported great difference in the type and number of terms manually selected by terminologists, domain experts, translators and information scientists. Hence one important element that any gold standard for term extraction should consider is the diverse expectations from different stakeholders.

To reconcile the considerably varied interests and concerns, a better understanding and thus consensus of the distinction between terms and non-terms in any given domain is most important. Terminology and phraseology should be sufficiently distinguished as far as practicable. Measures should be established to enable us to better define and differentiate along the gradation from common expressions to core technical terms for a certain domain. This calls for a more thorough analysis of a whole range of expressions deemed important by translators for a specialised domain, based on well-defined criteria, linguistic or otherwise. Secondly, instead of just black-box testing according to their own subjective judgement, translators should also be entitled to more objective testing on their side, as much as computational terminologists. A gold standard based on a term-annotated corpus, which is obviously lacking for bilingual English-Chinese terminology, will therefore be necessary. Such a benchmarking dataset should annotate a full range of expressions deemed relevant by translators, as well as terminologists and computational linguists, with the type and domain specificity indicated. Moreover, researchers need to re-consider the corresponding linguistic and statistical criteria in automatic term extraction, based on the linguistic description of domain-specific terms, to accommodate a more comprehensive set of concepts and their expressions. The relevance and applicability of compositional approaches need to be studied in more details, and new approaches need to be devised for the non-compositional cases. Testing and evaluation of newly developed tools, of course, should refer to the benchmarking data available.

4.1. Bilingual Term Analysis

Linguistic insights are often helpful for extracting terms from one language as well as identifying translation equivalents in another language. Syntactic structures, variant forms and compositionality are particularly relevant considerations (e.g. Baldwin and Tanaka, 2004; Hippisley et al., 2005; Daille, 2005; Bartels and Speelman, 2014). Sometimes regional variation may also be an issue.

We first collected various existing bilingual financial glossaries or term lists available in Hong Kong. These resources include: terms listed in two textbooks on financial translation[1] and the glossary from the Education Bureau for secondary school education[2], as well as glossaries from the Hong Kong Exchanges and Clearing Ltd[3], the Securities and Futures Commission[4], and the Hong Kong Monetory Authority[5]. The data sizes range from a few hundred to over 10,000 term pairs. While the more official glossaries are obviously more comprehensive than the lists given in textbooks, a collection like this can reveal a broader spectrum of what different groups of people, including translators, educators and domain experts, view the nature and scope of terminology in the financial domain.

Samples of English-Chinese term pairs were selected from the various sources and analysed with respect to the following aspects[6]:

- Word classes (e.g. nominal or verbal): As expected, the majority of the terms are nominal. In general, there is less than 1% of the terms in our samples which are not nominal. It was nevertheless observed that the Chinese equivalents are not necessarily in the same word class as the English terms (e.g. "dilution" is a noun while its Chinese equivalent 攤薄 is apparently verbal), although the vast majority of the pairs do have compatible word classes (e.g. "brokerage" 券商).

- Constituent compatibility: The lexicalisation of concepts could be quite different between English and Chinese, although most term pairs are in fact multi-word English and Chinese expressions, as illustrated in Table 1.

#words (E:C)	%	Examples
1:1	7.73	underwriter 包銷商 volatility 波幅
1:N	12.84	exclusions 不受保項目 jumbomize 將股票化零為整
N:1	8.84	resumption of trading 復牌 bad and doubtful debts 呆壞帳
N:N	70.59	backdoor listing 借殼上市 rateable value 應課差餉租值

Table 1: Lexicalisation among E-C Bilingual Term Pairs

- Syntactic structures (for multi-word English terms): As also suggested by Table 1, about 80% of the selected English terms are multi-word expressions. Considering the nominal expressions, the majority take a modifier-head structure, while some have post-modifiers and others have both pre- and post-modifiers, as shown in Table 2.

- Compositionality (for multi-word Chinese equivalents): For the multi-word English terms, it is important to see whether the corresponding Chinese equivalents are also formed compositionally. For example, the Chinese equivalent for "interim dividend" can be compositionally formed as 中期/interim

[1] 周兆祥、范志偉 (2004) 《財經翻譯精要》香港：商務印書館 and 李德鳳 (2007) 《財經金融翻譯：闡釋與實踐》香港大學出版社 (Both books are written in Chinese.)

[2] http://www.edb.gov.hk/attachment/en/curriculum-development/kla/technology-edu/whats-new/bafs_glossary_071130.pdf

[3] http://www.hkex.com.hk/eng/global/documents/glossary_ec.pdf

[4] http://www.sfc.hk/web/doc/EN/inutilbar/glossary/2006/full_list.pdf

[5] http://www.hkma.gov.hk/gdbook/chi/main/index_c.shtml

[6] The quantitative figures reported here are based on some 1,900 term pairs selected from two of the above sources.

Structure	%	Examples
Head-Mod	5.17	merger by absorption, statement of capital
Mod-Head	88.74	accrued expenses, bare trustee
Mod-Head-Mod	4.50	straightline method of depreciation, carrying amount of an asset
Others	1.59	stores and spares, delivery vs payment

Table 2: Syntactic Structures of English Multi-word Terms

股息/dividend, while the correspondence is not as straightforward between "initial margin requirement" and 開倉保證金, or between "evening evaluations" and 最後收盤價. Nevertheless, at least two-third of the samples are the compositional type.

4.2. Devising Annotation Guidelines

Guidelines are then to be drawn up to explain the what and how for selecting bilingual terms from a corpus for our gold standard. In addition to the setting-based criteria, linguistic criteria, and formal criteria discussed in Bernier-Colborne and Drouin (2014) governing term selection for their test corpus on automotive engineering, we include further considerations. First, more attention will be paid to translators' expectations, which might include not only terms that are likely to be found in specialised glossaries but also other semi-technical expressions. Second, the linguistic criteria will be adjusted with the observations from our own analysis in the project, especially noting the variety of syntactic structures, term variants and compositionality relevant to English-Chinese financial terms.

4.2.1. Scope of Terms

Bernier-Colborne and Drouin's (2014) setting-based criteria took on relatively strict terminological considerations. Given the domain of automotive engineering, only expressions denoting tangible and intangible objects or products directly related to the understanding of the subject matter are considered valid terms, while those referring to more generally associated entities are excluded.
<u>*Our consideration*</u>: Given the nature of the financial domain, the concepts are possibly more intangible than tangible (e.g. "insurance" 保險 is a relatively intangible concept), and very often involve processes (e.g. "closing transaction" 平倉交易) other than objects. In addition, from the perspectives of translators and domain experts, the boundary between terms and non-terms is more fuzzy than that as viewed by terminologists. For instance, organisation names are often included in professional glossaries, while common fixed phrases are found in translators' term lists. Hence it is less easy to limit the scope to "pure" terms, and that would not meet most users' expectation. For our purpose, the scope of terms is therefore less restricted, and annotators are asked to distinguish among four types of expressions. Type A will contain the core financial terms, for which expressions (single-word or multi-word) could be selected as long as they carry self-contained domain-specific meanings relevant to the financial context. Examples are "profit before tax" 除稅前利潤, "derivatives" 衍生工具, etc. Type B will be the semi-relevant expressions, such as "Board of Directors" 董事會, which is not only relevant to banking and finance, but also to a wider business context or even other non-commercial settings. Type C may include frequent phrases and jargon which appeal to translators as warranting specific translations. However, these expressions, despite being found frequently in financial documents, are not necessarily terms. The inclusion of this type inevitably loosens the scope of terms to a large extent, but it also allows us to accommodate broader views and to make better distinction for various situations. For example, the phrase "top and emerging risks" 首要及新浮現風險 has appeared more than 10 times in the annual report of one particular bank, but "top" and "emerging" do not refer to any intrinsic quality of "risks" or indicate a specific kind of risks (unlike "credit risk" 信貸風險, for instance), so the phrase should not be considered a term. But naturally translators are tempted to include it in their glossary to facilitate translation. Annotators are thus not to go simply by frequency, as a less frequent expression like "mitigating action" 緩減行動 may have a more specific meaning (in risk management for this example) and thus merit a different categorisation. The fourth type of expressions, which does not need to be annotated, contains general words and phrases like "location" 所在地, "information" 資料, "charitable donations" 慈善捐款, etc. They are not considered for the current purpose, whether or not a translator finds them difficult or special.

4.2.2. Form of Terms

The linguistic criteria in Bernier-Colborne and Drouin (2014) stipulate that only nouns and noun phrases should be annotated. Base terms and variants are included but distinguished. Following L'Homme (2004), morphologically related forms of a selected term and expressions which are paradigmatically related to a selected term could also be included. There was no restriction on compositionality, but syntactic variants are not considered.
<u>*Our consideration*</u>: As observed in the analysis above, although most terms listed in translation textbooks and professional glossaries are nominal, a small amount of verbal terms are also found. In fact, it is not unusual for the processes taking place in the financial domain to be expressed by verbs. Hence our annotation is not limited to nouns and noun phrases, but also covers verbs and verb phrases as appropriate. For example, while "underwriter" 包銷商 would be found in many glossaries, the verb "underwrite" 承辦 may not be a less important term especially in real financial corpora. In addition, we do not really distinguish base terms and variants, as our analysis shows that some terms do have post-modifiers and they are not exactly variants of a base form (e.g. "statement of capital" 股本說明 is not really the variant of "capital statement").

4.2.3. Span of Terms

According to their formal criteria, no limit was placed by Bernier-Colborne and Drouin (2014) on the length of terms as they appear in the corpus, and only maximum-length terms were to be annotated as far as they fulfilled the other criteria. Otherwise shorter terms embedded therein and satisfying all term selection criteria could be annotated.

Our consideration: We basically also follow the maximum-length principle. We do not ask annotators to consider the shorter embedded terms, for if such shorter terms appear elsewhere in the corpus, they would be selected anyway. For example, where "financial system abuse risks" 金融系統濫用風險 is the longest term found in a context, there is no need to annotate "financial system" and "risks" in the same context. "Financial system" would have been annotated in another context when it stands on its own, such as "... to the stability and effective working of the financial system of Hong Kong", if the annotator chooses to mark it. One important point which is specific to our guidelines is the consideration of the Chinese equivalents. All along, the guidelines imply starting from the English text. Nevertheless, any English term fulfilling all selection criteria should only be annotated if a Chinese equivalent in its full form can be located from the corresponding Chinese text.

4.3. Term Annotation in Corpus

We started term annotation, based on the above principles, with the Annual Report and Accounts 2016 of the Hongkong and Shanghai Banking Corporation Limited (香港上海滙豐銀行有限公司2016年報及賬目), available in English and Chinese[7]. The corpus size of the various sections in the annual report is shown in Table 3.

Section	English (words)	Chinese (chars)
Report of the Directors 董事會報告	3,917	6,749
Financial Review 財務回顧	1,511	2,693
Risk Report 風險報告	14,383	27,013
Capital 資本	1,825	3,426
Notes on the Financial Statements 財務報表附註	18,360	31,350
Total	39,996	71,231

Table 3: Corpus Size (HSBC Annual Report 2016)

4.3.1. Preliminary Comparisons

For training, four annotators, all undergraduate translation students, were instructed with the guidelines and asked to mark up bilingual term pairs from the section on Report of the Directors. The results, at least quantitatively, turn out to be quite varied. As shown in Table 4, the first three annotators (AC, MY and JT) apparently form a more lenient group. The number of expressions they selected almost doubles the number of expressions selected by the fourth annotator (CC). Moreover, the distribution of types (A, B or C) looks very different across annotators. Such discord on number and classification reflects that despite being instructed with the same guidelines, individual annotators could still differ considerably in their perception of termhood as well as their understanding of the task requirements.

Annotator	Type A	Type B	Type C	Total
AC	86	52	4	142
MY	29	71	37	137
JT	37	62	39	138
CC	49	19	4	72

Table 4: Number of term pairs annotated

Ignoring the term classification for the time being, Table 5 gives a sketch of the agreement among the annotators. For a total of 246 distinct expressions selected, over half were actually selected by two or more annotators.

Selected by	N (%)
4 annotators	28 (11.38%)
3 annotators	46 (18.70%)
2 annotators	67 (27.24%)
1 annotator	105 (42.68%)

Table 5: Agreement among annotators

Expressions selected by all four annotators include some of the relatively standard financial terms and fixed phrases, although not as comprehensively as one might expect, and the classification is not always uniform. Some examples are shown below:

Selected by all annotators:

financial statements	財務報表
material risk takers	承受重大風險人員
ordinary shares	普通股
risk appetite	承受風險水平
share capital	股本
subsidiary	附屬公司
terms of reference	職權範圍
trade corridors	貿易走廊

It happens that there are more cases where the same expressions have not been selected by all annotators but only some of them. This is probably where the translators' (or users') expectation can be shown as a salient factor. Although the annotators have received some basic translation training, they are still students. The more capable ones or those with better language proficiency may choose to ignore the more common expressions and the straightforwardly compositional ones. As they may not really think of including such expressions if they are to keep their own glossaries manually, they may not expect or strongly wish them to be extracted by automatic term extraction systems either. Here are some examples:

[7] https://www.personal.hsbc.com.hk/1/2/hk/regulatory-disclosures

Selected by 3 annotators:

Banking Ordinance	銀行業條例
base salary	基本薪金
consolidated profit	綜合利潤
debentures	債券
dividends	股息
material interest	重大利益
Nomination Committee	提名委員會
remuneration policy	薪酬政策

Selected by 2 annotators:

auditor	核數師
business strategy	業務策略
economic capital	經濟資本
funding structure	資金架構
priority growth markets	優先發展市場
risk environment	風險環境
shareholders	股東
transactions	交易

Notwithstanding the maximum-length principle discussed in Section 4.2.3, it is also noted that individual annotators may from time to time tend to select phrases longer than necessary. For example, Annotator AC has selected "income tax and social security", while others tend to select "income tax" as one unit, and optionally with "social security" as a separate one.

4.3.2. Implications on Training

The annotation work is in progress, and further analysis of the annotators' work is being done. Feedback is given to the annotators regularly, to gradually bring their understanding of the annotation guidelines closer to one another, especially regarding the selection principles as well as basic linguistic awareness. It is not intended, and in fact not possible, for uniform annotation from everyone, as individual variation is what we are interested in, to see how translators vary in their perception of terms in financial translation, and thus their expectation of automatic term extraction systems. On the other hand, notwithstanding individual differences, the annotators need to follow a similar practice in their selection of terms, and in the end we will take the majority of vote to arrive at the benchmarking data. In addition, the classification of term types may need to be verified with existing professional financial glossaries, instead of relying entirely on the annotators' judgement. After all, they are still novice translators.

Annotation of more bilingual financial texts has been planned, including but not limited to prospectuses, annual reports of listed companies, banking information, as well as government documents relating to economics and finance.

5. Potential NLP Applications

Although the primary objectives of the annotation at the outset mainly focus on the evaluation of automatic term extraction from a more translator-oriented perspective, the resulting bilingual terminology dataset is expected to be beneficial to the processing of financial narratives in the NLP context as well.

While the actual usage of the language resource in concrete applications has to await the completion of the resource on the one hand and further in-depth research on the other, a potential example is portrayed here for some preliminary idea of the benefits that the resource might offer in practice. Instead of a general financial glossary, the annotated term dataset is grounded on authentic texts from the financial domain. The annual reports, for instance, are known to follow a standard format and structure every year, with changes mostly on the numerical data and certain qualitative details. Hence, the terms and fixed expressions obtained from the annual report in a certain year could provide a useful resource for language processing systems to anchor at various parts of the reports in other years. Leveraging the structural and linguistic similarities, information mining from reports of different years for comparison should be facilitated. The fact that the resource is document-specific means that the differences in individual series of documents could be taken into account, especially when bilingual processing is concerned. For example, given the same financial terms in English, it has been observed that annual reports of different banks may use different Chinese equivalents. The annotated dataset could thus offer more than existing general glossaries for language processing systems to gauge important information from financial documents of different organisations with individual linguistic conventions.

6. Conclusion

This paper has presented our ongoing work which aims to develop a gold standard based on a term-annotated corpus, to offer a resource currently lacking for the development and evaluation of bilingual English-Chinese terminology extraction systems.

The production of the intended benchmarking dataset consists of two major tasks. The first is a thorough linguistic classification and analysis of single-word and multi-word terms (and term pairs). The second task is, based on the linguistic analysis, to devise a set of term annotation guidelines and build up an annotated parallel corpus. Third, and more importantly, we are annotating bilingual terms in the corpus. While previous attempts have been for monolingual term annotation only, our term selection criteria cover different scenarios of English-Chinese correspondence.

Tools in translation technology are intended to assist translators, and the whole purpose will be defeated if translators fail to fully utilise and appreciate them. With a carefully cultivated benchmarking resource, we hope to enable a more translator-oriented perspective for the evaluation of automatic term extraction systems so that they can fulfill their roles in translation technology better and embrace more appreciation from their target users. As Agirre et al. (2000) stated, "... tools for translation cannot be satisfactorily designed without the cooperation of human translators" (p.296). Although our current work deals primarily with English-Chinese financial terms, the rationale and significance underlying a gold standard to accommodate multiple perspectives, including users and researchers, apply to terminology work on other languages and domains alike.

7. Acknowledgements

The work described in this paper was partially supported by grants from the Faculty of Arts of the Chinese University of Hong Kong (Project No. 4051094) and the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. 14616317).

8. Bibliographical References

Agirre, E., Arregi, X., Artola, X., de Illarraza, A. D., Sarasola, K., and Soroa, A. (2000). A methodology for building translator-oriented dictionary systems. *Machine Translation*, 15:295–310.

Baldwin, T. and Tanaka, T. (2004). Translation by machine of complex nominals: Getting it right. In *Proceedings of the Second ACL Workshop on Multiword Expressions: Integrating Processing*, pages 24–31, Barcelona, Spain.

Bartels, A. and Speelman, D. (2014). Clustering for semantic purposes: Exploration of semantic similarity in a technical corpus. *Terminology*, 20(2):279–303.

Bernier-Colborne, G. and Drouin, P. (2014). Creating a test corpus for term extractors through term annotation. *Terminology*, 20(1):50–73.

Blancafort, H., Bouvier, F., Daille, B., Heid, U., and Ramm, A. (2013). TTC web platform: from corpus compilation to bilingual terminologies for MT and CAT tools. In *Proceedings of TRALOGY II*, Paris.

Bourigault, D. (1992). Surface grammatical analysis for the extraction of terminological noun phrases. In *Proceedings of the Fourteenth International Conference on Computational Linguistics (COLING '92)*, pages 977–981, Nantes, France.

Cabré Castellví, M. T., Bagot, R. E., and Palatresi, J. V. (2001). Automatic term detection: A review of current systems. In Didier Bourigault, et al., editors, *Recent Advances in Computational Terminology*, pages 53–87. John Benjamins Publishing Company, Amsterdam/Philadelphia.

Daille, B. and Morin, E. (2005). French-English terminology extraction from comparable corpora. In Robert Dale, et al., editors, *Natural Language Processing – IJCNLP 2005. Lecture Notes in Artificial Intelligence, Volume 3651*, pages 707–718. Springer-Verlag.

Daille, B. (1996). Study and implementation of combined techniques for automatic extraction of terminology. In Judith L. Klavans et al., editors, *The Balancing Act: Combining symbolic and statistical approaches to language*, pages 49–66. MIT Press, Cambridge, MA.

Daille, B. (2005). Variations and application-oriented terminology engineering. *Terminology*, 11(1):181–197.

Daille, B. (2012). Building bilingual terminologies from comparable corpora: TheTTC termsuite. In *Proceedings of the 5th Workshop on Building and Using Comparable Corpora*, pages 29–32, Istanbul, Turkey.

Drouin, P. (2003). Term extraction using non-technical corpora as a point of leverage. *Terminology*, 9(1):99–115.

Estopà, R. (2001). Les unités de signification spécialisées: élargissant l'objet du travail en terminologie [units of specialised meaning: Broadening the scope of terminology work]. *Terminology*, 7(2):217–237.

Fulford, H. (2001). Exploring terms and their linguistic environment in text: A domain-independent approach to automated term extraction. *Terminology*, 7(2):259–279.

Hippisley, A. R., Cheng, D., and Ahmad, K. (2005). The head-modifier principle and multilingual term extraction. *Natural Language Engineering*, 11(2):129–157.

Kwong, O. Y. (in press). Evaluating term extraction tools: System performance vs user perception. In Sin-Wai Chan, editor, *The Human Factor in Machine Translation*. Routledge.

Laroche, A. and Langlais, P. (2010). Revisiting context-based projection methods for term-translation spotting in comparable corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, pages 617–625, Beijing, China.

L'Homme, M.-C. (2004). *La terminologie: principes et techniques [Terminology: Principles and Techniques]*. Presses de l'Université de Montréal, Montréal.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Vivaldi, J. and Rodríguez, H. (2007). Evaluation of terms and term extraction systems: A practical approach. *Terminology*, 13(2):225–248.

Xu, R. and Sharoff, S. (2014). Evaluating term extraction methods for interpreters. In *Proceedings of the 4th International Workshop on Computational Terminology*, pages 86–93, Dublin, Ireland.

BORSAH: An Arabic Sentiment Financial Tweets Corpus

Mohammed Alshahrani[1,2], Fuxi Zhu[1*], Mohammed Alghaili[3], Eshrag Refaee[4], Mervat Bamiah[5]
Computer School, Wuhan University[1], College of Computer Science and IT, Albaha University [2], Computer School, Hunan University[3], Computer Faculty, Jazan University[4], Faculty of computer science, Prince Sultan University[5], Wuhan, China[1], Albaha, Saudi Arabia[2], Changsha, China [3], Jazan, Saudi Arabia[4], Riyadh, Saudi Arabia[5]
*Corresponding author: fxzhu@whu.edu.cn

Abstract

Impact of social media networks such as Twitter, Facebook, Instagram, etc. on business is vital since people opinions and attitudes may affect the success or failure of a product or a service. This study is a part of continues research project entitled "Evaluating the influence of Twitter on the Saudi Arabian Stock market (TADAWUL)[1] to investigate the impact of Twitter financial tweets on the Saudi Arabia stock market. This paper presented BORSAH an Arabic financial sentiment analysis dataset (corpus) crawled from Twitter. The collected dataset consists of (41,455) Arabic gold-standard annotated Twitter feeds gathered from (118,283) tweets tagged manually from total crawled dataset that consists of (277,453) tweets. The experiment went through three steps, Firstly, we labeled the corpus for Subjectivity and Sentiment Analysis (SSA). Secondly, we applied three machine learning algorithms on part of the corpus. Thirdly, we calculated the accuracy rate of each algorithm. A first sub-corpus will be released via the European Language Resources Association (ELRA) repository with this submission. As far to our knowledge, this is the largest manual annotated Arabic tweets corpus for SSA and the first Arabic financial corpus that will be available for the research community.

Keywords: Arabic Corpus, Sentiment Analysis, Stock Market, Tweets Dataset Analysis, Twitter

1. Introduction

Researchers have been investigating the impact of Twitter on diverse fields such as politics, healthcare, public opinion and stock markets among others. The main challenge was understanding the behavior of users and trends accurately. Several research works were performed to identify users' preferences to predict stock markets prices trends (Fama et al., 1969; Qiu and Song, 2016). This paper presented BORSAH an Arabic financial sentiment analysis dataset crawled from Twitter. It is the Arabic synonym of "Souq Alashom" which means stock exchange market. The word BORSAH has inherited its name from the word "Bourse" that refers to the stock exchange. BORSAH is part of a previous research project for investigating the influence of Twitter on TADAWUL All Shares Index (TASI)[1].

In this study, we used Twitter Application Programming Interface (API) for crawling (277,453) Arabic tweets related to the stock market. We implemented three types of correlations, Pearson correlation coefficient, Kendall rank correlation, and Spearman rank correlation to prove the correlation between Twitter and the Saudi stock market. Furthermore, we considered also the variable mention role for identifying twitter accounts, whose tweets contributed towards market trends. The experiment deployed three machine learning algorithms on a training (test) dataset for crawling (14,000) tweets from the gold-standard annotated feeds. However, during the annotation process, we observed a phenomenon of Twitter selling the Twitter[2] followers tweets to increase the count of followers and illustrate popularity for a specific Twitter account.

This paper is organized as follows: section 2 discusses the related works in Arabic sentiment analysis for the Saudi stock market, and Twitter. Section 3 presents the corpus collection steps including data collection, pre-processing, annotation, and analysis. Section 4 discusses the results. Section 5 views the correlation between BORSAH and TASI, whereby Section 6 describes the release format. Finally, Section 7 concludes the study and the findings.

2. Related Works

Researchers have applied machine learning techniques in Natural Language Processing (NLP) for Subjectivity and Sentiment Analysis (SSA). They annotated their corpora by disregarding the grammar or the lexicon-based methods. Moreover, based on their observations they stated that there is an urgent need to build an Arabic corpus by harvesting different types of Arabic texts on the web to fulfill the demands of scientists in the Arabic NLP research field (Al-Sabbagh and Girju, 2012; Abdul-Mageed and Diab, 2012). Refaee et al. (2014) have crawled and annotated a general Twitter Arabic corpus for SSA. Duwairi et al. (2014) stated that sentiment analysis was implemented in financial industry to illustrates the sentiment data of targeted companies which enables better decision in real time.

AL-Rubaiee et al. (2015) have classified Arabic text in stock trading using Support Vector Machine (SVM), K-Nearest Neighbors (KNN) and Naive Bayes algorithms. They presented the relations between Arabic tweets and stock market movements based on observation of both stock market closing price and daily sentiment tweets graph. However, their dataset was small, and they did not prove the correlation.

In their extended research AL-Rubaiee et al. (2016) studied MUBASHER the leading stock analysis software provider in Gulf Cooperation Council (GCC) region using Twitter for opinions mining purposes. They extracted feedbacks from MUBASHER by designing a model for the Saudi Arabic tweets sentiment analysis. Their model

[1]https://www.tadawul.com.sa/wps/portal/tadawul/markets/equities/indices/today

[2]https://www.theatlantic.com/technology/archive/2018/01/all-followers-are-fake-followers/551789/

combined machine learning and NLP approaches for classifying the Arabic tweets into sentiment polarity classes: Positive, Neutral and Negative.
This research aims to mitigate the gaps in related works at the Arabic NLP community regarding finance by proving the correlation between Twitter and TASI as continuous work of a previous project. Moreover, this paper provides the largest gold annotated Arabic manual available that contains (41,455) Arabic tweets corpus for SSA, also it provides the first Arabic tweets dataset for SSA experts in the stock market for mining other objects of tweets such as Mention, Retweets, Following Count, Followers Count and its impact on stock market trends.

3. Twitter Financial Corpus

Twitter API[3] is used by developers and researchers for retrieving or modifying data. The Twitter Search API provides relevant results to ad-hoc user queries from a limited corpus or recent tweets. The Representational State Transfer (REST) API allows access to Twitter texts for reading the timeline, tweeting and following. A study conducted by Alsing and Bahceci (2015) stated that there are several issues when using Twitter Search API including the complexity of restricted tweets quires and the availability of data that cannot be older than seven days. Since the Search API only uses indices which contains most recent or popular tweets. Another issue is that Twitter Search API is used for relevance and not for completeness which results in some missing tweets in the query results.

The Twitter Search API Developers Page[4] states that Streaming API is more suitable for completeness-oriented queries. However, this is not the case of gathering data for sentiment analysis where high completeness is required to analyze the whole data. Twitter provides various types of streaming endpoints, each type is customized to a specific use such as a) Public streaming, b) User streaming and c) Site streaming. Public streaming refers to long-lived Hypertext Transfer Protocol (HTTP) and parsing requests incrementally. The streaming API allows requests for various parameters including language, location, follows, track, count, and delimited to define what data is supposed to be returned. However, the request is different for each language or framework based on HTTP library.

The used Twitter API supports UTF-8 for the Arabic language that may cause characters counting problems for the Arabic dialectal tweets. The main drawback of Twitter API is that a user may crawl tweets based on a country code or enters a location longitude and latitude using API search. Unfortunately, API streaming does not provide this option. Moreover, Twitter does not provide country code location of the user unless he activates the location tracking.

This study aims to identify the minority of users who have activated their locations, also to mitigate the geographical attribute challenge for API streaming, since Saudi Arabia geographical shape is semi-rectangular from north to south. The tweets were crawled within a radius of Saudi Arabia from central point covering the whole country, also the radius distance include some countries such as Jordan, Iraq, Egypt, Sudan, GCC, and Iran.

3.1 Data Collection

In this study several Twitter API's were used in parallel to crawl Arabic tweets regarding TADAWUL. However, only (10) keywords were accepted for each query, due to the limitation of crawlers which may lead to missing tweets about TADAWUL that do not contain the designed keywords. A set of search queries were generated to enhance the possibility of acquiring tweets which convey emotions, attitudes, and opinions towards a specified entity. Table 1 shows the keywords used to retrieve the Arabic financial tweets. Crawling live stream tweets was conducted between (27th August to 23rd December 2015), with total of (277,453) collected tweets.

Arabic	Transliteration	Translation
السوق	Assouq	Market
تاسي	TASI	TASI
تداول	TADAWUL	TADAWUL
الأسهم	Al-as-home	Shares
السوق السعودي	Assouq Assaoodi	Saudi Market
سوق الأسهم	Sooq Al-as-home	Stock Market
سعر السوق	Se'er Al-sooq	Market Price
سعر الإغلاق	Se'er Al-eGh-laaq	Closing Price
ارتفاع	Ertifaa	Growth
هبوط	Hoboot	Fall

Table 1: Arabic financial related tweets keywords

3.2 Data Pre-processing

The tweets were stored in MongoDB[5] which contains the objects of each tweet. During the crawling process, several filters were added to a) block spam, and b) to blacklist spammers' accounts, also c) to remove duplicate tweets from the same IDs, as well as e) to remove the tweets that contain long words. These steps were conducted to ensure that the source was safe from any automatic spam tweets. We created a blacklist for spammers IDs and keywords of their spamming tweets from the initial test. Moreover, (5000) tweets were collected and manually classified their contents based on the same keywords to detect random spammers IDs and spamming keywords, to reduce spam tweets in the main intended crawling process.

Figure 1. illustrates the process of crawling, filtering, inserting, tagging and exporting Arabic tweets regarding TADAWUL. The tweets contained public objects such as User ID, Screen Name, Name, created at, Text, Mention, Retweets, Following Count, Followers Count. Moreover, Figure 1. presents the platform built for annotators to tag each tweet based on its polarity, region, spam, and non-related meaning. It illustrates the mechanism for retrieving the tweets from the database for simple statistical analysis, and easier manual annotation process, whereby the annotator clicks on one of the annotation icons colored Green to choose his/ her preference, either Green, Red, or

[3] https://dev.twitter.com/rest/public/search
[4]https://developer.twitter.com/en/docs/tweets/filterrealtime/guides/connecting.html
[5] https://www.mongodb.com/

Orange. The Green is positive, Red is negative, and Orange is neutral. Furthermore, the “No go icon” indicates it is spam. The “3-connected-dots icon” represents non-related tweets that do not refer to the stock market. The “location icon” refers to stock market tweets related to another country besides Saudi Arabia.

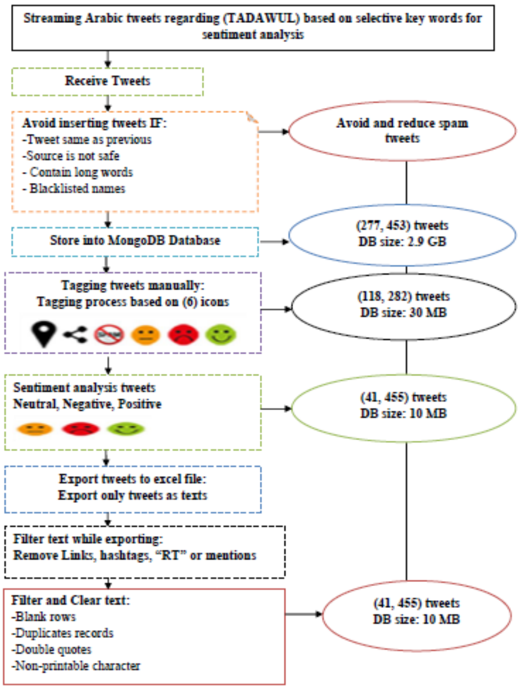


Figure 1: Arabic tweets streaming API.

3.3 Data Annotation

Table 2 illustrates the tweets statistics. Total of manual tweets tagged were (118,282), and total of gold-standard (41,455) tweets were in the dataset based on their sentiment polarity. Moreover, total of (3,756) spam tweets were tagged manually including spamming IDs and keywords lists, this process prevented thousands of spammers from storing their tweets in our database. Furthermore, (58,840) tweets did not have any related meaning even it contained the research keywords due to Arabic multiple meanings of the same word.

Type of Tweets	No. of Tweets
Total tweets crawled	277,453
Total tweets tagged manually	118,282
Positive tweets	5,449
Negative tweets	9,469
Neutral tweets	26,537
Spam	3,756
Non-related meaning tweets	58,840
Non-Saudi Stock tweets	14,231

Table 2: Tweets Statistics

However, a total of (14,231) tweets were excluded from the dataset due to their irrelevant content to stock market as they only contained some stock-market-related keywords. The annotated set of (14,231) tweets were excluded since they were discussing other major stock markets (e.g. GCC) with no mention of the Saudi stock market.

The first round of tagging process was carried out by finance graduate researcher as the tweets contained finance related vocabulary and daily stock market information. Afterward, a Saudi linguistic expert double checked the tagged tweets. When there was a conflict between the two annotators about such tweet, a third assessor had evaluated the tweet to which class should be assigned to. Moreover, since the tagging process is time-consuming we divided the dataset into two parts. The first dataset consists of (118,282) tweets which were tagged and analyzed in this paper. The rest will be released in the future after completing manual tagging and analysis.

Identifying the related tweets was costly in terms of time and efforts as it needed several months to tag the collected tweets manually since the required data has specific nature regarding the stock market in Saudi Arabia. Twitter has inserted enormous tweets from GCC countries, it was extremely tough in some cases to identify the tweet that refers to which stock market in the region Saudi or GCC. Additionally, the same keywords of the stock market can be used in diverse fields not only stock market.

We observed a phenomenon that anonymous users are selling the Twitter[6] followers to increase the number of followers and illustrate popularity for a Twitter account. This observed phenomenon prevented us to from storing such tweets by false popular accounts that may lead to a false analysis of this dataset. For example, a Twitter account following 180k accounts and 190k followers most probably this Twitter account user does not read the tweets of those accounts he follows due to the huge number of tweets that are generated hourly by the users. Additionally, to reduce the impact of such cases, we set a threshold of 5k accounts that each user can follow. A total of (38,432) tweets or IDs were excluded into a separate spam dataset. The Main attributes of the dataset are defined in Table 3.

Attribute Name	Attribute Description
User_ID	Unique ID assigned by Twitter to each user
Created_at	Date and time on which tweet as posted.
Screen_name	Account name displayed on twitter for each user
Followers_count	Number of other users following this account
Retweets_count	Total No of retweets posted for this tweet
Mentions	Total No of mentions of this tweet
Following_count	No of twitter accounts a user is following
Status	The polarity of the tweet text. It mainly consists of three emotions, negative, positive and neutral
Text	Arabic text 140 letters

Table 3: Main Attributes of Data Set

[6] https://www.theatlantic.com/technology/archive/2018/01/all-followers-are-fake-followers/551789/

Figure 3. illustrates the Arabic tweets daily volume and amount of positive, neutral and negative tweets. These tweets were gathered from (27th August to 20th October 2015). Negative tweets are presented in red, Positive tweets in green and daily volume in blue color.

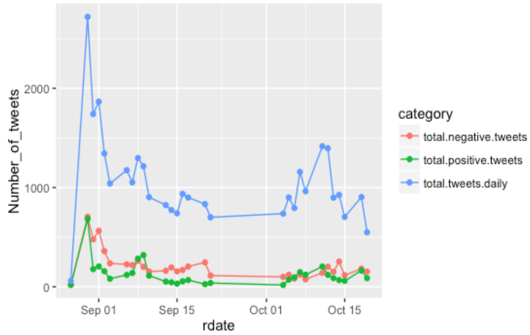


Figure 3: Volume of Daily Tweets

3.4 Data Analysais

A machine learning system was built for Arabic sentiment analysis. The used algorithms to explore the polarity of a given data were Naive-Bayes, SVM, and KNN, which provide the best accuracy in sentiment analysis for Arabic text in relation to the Saudi stock market. First, part of the gold-standard corpus of (14,000) tweets were used as a training dataset. After that, each one of the three machine learning algorithms was applied to the training dataset. Finally, the experiment was processed on (625) tweets test dataset to evaluate the accuracy rate of each one of the algorithms. The implementation of Naïve Bayes was divided into three stages.

1- *First stage:* representing sentences by a vector of words called “victory” for each feature in the dataset. Victory refers to a large array that contains each word with its frequency in the dataset denoted by $|VOC|$.

2- *Second stage:* learning the dataset by training the text test to estimate the probability of each word (w_k) in the text test with each feature in the dataset. To perform this a calculation process must be performed as follows:

- Calculating the average of each feature (F_j) in the dataset by dividing the number of each feature by the whole number all features in the dataset.

- Calculating the probability of each word (w_k) within each feature (F_j) in the dataset separately. By the following equation:

$$P(w_k|F_j) = \frac{n_k+1}{(n+|VOC|)} \quad (1)$$

Whereby n is the number of words in the a given feature F_j in the dataset. n_k is the number of times the word k occurs in each featureF_j in the dataset.

3- Third stage: Training the classifier by calculating the value of Naïve Bayes (V_{NB}) using the following equation:

$$V_{NB} = argmaxP(F_j)\prod_{w_k\in W} P(w_k|\, F_j) \quad (2)$$

KNN is the second machine learning algorithm that has been used for:

- Calculating the Euclidean distance between vectors and the dataset sentence.
- Sorting the distances in decreasing order.
- Taking the k nearest distances to the sentence that needs to be classified.
- Finding the majority class among the selected k distances. the majority class will be the chosen prediction for the given sentences.

Assuming k = 3 (where k is the closest distance to the given sentence).

The last machine learning algorithm used in the proposed system is SVM has the common library libsvm as most researchers are using it.

4. Results

We have examined the three algorithms with text test to find out the accuracy rate. The performance metrics that were widely used to evaluate the classification results were precision and recall (Khan et al., 2014). The results were summarized in Table 4. highlights the number of tweets annotated automatically.

Algorithm	No. of Test Sentences	No. of Positive Sentences	No. of Negative Sentences	No. of Neutral Sentences	No. of Errors
Naive Bayes	625	35	42	548	155
KNN	625	67	165	393	97
SVM	625	90	142	393	58

Table 4: Numbers of tweets annotated automatically

Algorithm	TP	FP	TN	FN	TP rate (Recall)	FP rate	Kappa	Accuracy rate	Precision
Naive Bayes	35	40	590	16	0.686275	0.011091	78.3077%	75.2%	0.466667
KNN	67	31	558	34	0.663366	0.052632	87.074%	84.48	0.683673
SVM	90	6	535	37	0.708661	0.063492	91.1878%	90.72	0.9375

Table 5: Statistical significance test for each algorithm

Moreover, statistical significance test was conducted to validate comparison of results. From the analysis, SVM has the highest accuracy rate with (90.72%) while Naive Bayes is the lowest with (75.2%).

This study is based on Cohen's methods (Cohen, 1960) which measure the degree of agreements among the assigned labels correcting for agreement by chance. We have tested (625) tweets that are not included in the training dataset to calculate the performance of each algorithm. All these test tweets annotated manually before testing them. We found out that the number of errors using Naive Bayes is larger than the number of errors in KNN and SVM. Table 5 shows the statistical significance test for each algorithm.

5. BORSAH and TASI Correlation

This study uses TADAWUL for finding the Saudi stock market trends. We collected the closing prices data for TASI through TADAWUL from (27th August to 20th October 2015). This data proved the correlations between the influential users, tweets and the stock market prices. Figure 4 shows the distribution of TASI performance, positive and negative tweets at the same time.

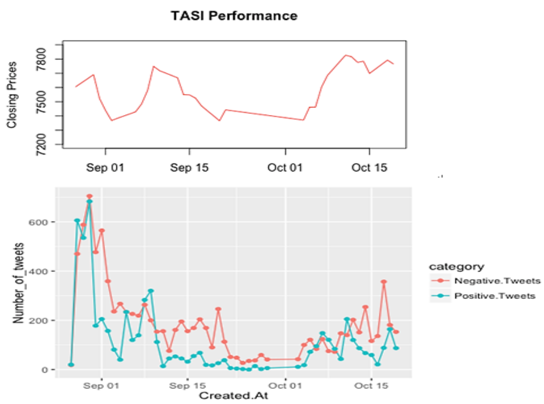


Figure 4 : Distribution of TASI performance

We illustrated in our previous work that three types of correlations were implemented, Pearson's correlation coefficient, Kendall rank correlation, and Spearman's rank correlation. Furthermore, we presented how the variable mention plays an important role in identifying twitter accounts, whose tweets contributed towards market trends. We emphasized that most influential users can be predicted in the future, who may have a significant impact on the stock market trends based on studying the followers count variable. In this paper, we observed a relationship between the daily volume of tweets and stock market index. The daily volume of tweets had increased and decreased in real time parallel with the stock market indicator during rising and falling phases. Figure 5. shows the distribution of TASI performance and daily volume at the same time.

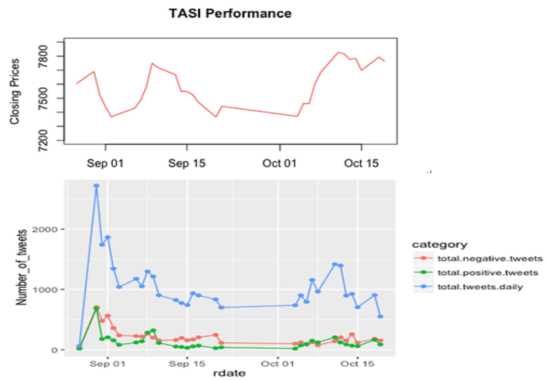


Figure 5: The daily distribution of TASI performance.

6. Release Format

The experiment consists of two datasets, the gold-standard annotated Arabic twitter feeds that consist of (41,455) tweets, and (118,283) manually tagged tweets dataset. Those two datasets will be released via the ELRA data repository which is saved in Comma Separated Values (CSV) file format. However, we removed Tweets text from the dataset due to Twitter privacy restrictions, our future submission will contain the text as encrypted. This is the first subset release among several planned releases. The aim of this study is to provide an annotated Financial Arabic Twitter dataset for the research community to investigate the influence of Twitter on the stock market.

7. Conclusion and Future Works

This paper presented BORSAH as an Arabic financial sentiment analysis dataset crawled from Twitter. We illustrated a method for harvesting Twitter Arabic finance tweets and presented the annotation process, we studied the properties and the statistics of the corpus. Moreover, we applied an Arabic text classification in finance through different algorithms such as SVM, KNN, and Naive Bayes. The SVM algorithm showed the best result.

This corpus is part of a project evaluating the influence of Twitter on the Saudi Arabian Stock market indicators TASI, whereby the correlation between TASI and Twitter was previously proved using three types of coefficients correlations. Moreover, we extended current corpus by annotating manually and automatically extra (159,171) tweets. A first sub-corpus will be released via the ELRA repository that focused only on TADAWUL Arabic tweets.

In future, we aim to investigate an observation we obtained about specific group users who tend to post many tweets targeting or mentioning specific companies. Thus, more investigation is needed to assess the impact of varying the threshold of number of accounts followed by a specific user, i.e. rather than the 5k threshold used in this study.

8. Acknowledgment

This research is supported by The National Natural Science Foundation of China with Grant No: 61272277.

9. Bibliographical References

Abdul-Mageed, M. and Diab, M. (2012). Awatif: A Multi-Genre Corpus of Modern Standard Arabic Subjectivity and Sentiment Analysis. *In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

AL-Rubaiee, H., Renxi, Q. and Dayou L. (2016). Identifying Mubasher Software Products through Sentiment Analysis of Arabic Tweets. *Industrial Informatics and Computer Systems (CIICS), 2016 International Conference on* IEEE. DOI: 10.1109/ICCSII.2016.7462396.

AL-Rubaiee, H., Renxi, Q. and Dayou L. (2015). Analysis of the Relationship Between Saudi Twitter Posts and the Saudi Stock Market. *Intelligent Computing and Information Systems (ICICIS), 2015 IEEE Seventh International Conference on.* IEEE. DOI: 10.1109/IntelCIS.2015.7397193

Al-Sabbagh, R. and Girju, R. (2012). Yadac: Yet another dialectal Arabic corpus. *In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Alsing, O. and Bahceci, O. (2015). Stock Market Prediction using Social Media Analysis. *Degree Project, In Computer Science*, First Level Stockholm, Sweden.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement,* 20(1):37–46.

Duwairi, RM. (2015). Sentiment Analysis for Dialectical Arabic. *6th international conference on Information and communication systems(ICICS)*. IEEE, pp. 166–170. DOI: 10.1109/IACS.2015.7103221.

Fama, E., Fisher, F., Jensen, M., and Roll, R. (1969). The Adjustment of Stock Prices to New Information. *International Economic Review*. 10(1):1–21. DOI: 10.2307/2525569.

Qiu, M. and Song, Y. (2016). Predicting the Direction of Stock Market Index Movement Using an Optimized Artificial Neural Network Model. *PLoS ONE*, 11(5): 0155133. https://doi.org/10.1371/journal.pone.0155133.

Refaee, E. and Rieser, V. (2014). An Arabic Twitter Corpus for Subjectivity and Sentiment Analysis. *In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Diachronic Lexical Changes In Company Reports: An Initial Investigation

Matthew Purver,*[*] Aljoša Valentinčič,[†] Marko Pahor,[†] Senja Pollak[‡]
[*]School of Electronic Engineering and Computer Science, Queen Mary University of London, UK
[†]Faculty of Economics, University of Ljubljana, Ljubljana, Slovenia
[‡]Jožef Stefan Institute, Department of Knowledge Technologies, Ljubljana, Slovenia
m.purver@qmul.ac.uk, {aljosa.valentincic,marko.pahor}@ef.uni-lj.si, senja.pollak@ijs.si

Abstract

We present initial investigations for a diachronic study of lexical changes in financial reporting, looking at methods suitable for analysing semantic associations between financial terms and how these change across time. Our corpus consists of US 10-K annual reports of 30 companies included in the Dow Jones Industrial Average stock index over the years 1996-2015. We grouped the reports by the reported fiscal year and derived word embedding models for each year using both GloVe and a count-based PPMI method; these vectors were then used to calculate cosine similarity between pairs of words. We expect the resulting diachronic patterns of lexical contexts of financial terms to vary with the economic cycle; here we select pairs of terms with strong increasing association over time (e.g. *dividend* and *shareholder*) or strong decreasing association over time (e.g. *dividend* and *gain*), and suggest some qualitative explanations for these changes due to the economic crisis.

Keywords: lexical changes, word embeddings, distributional semantics, financial reporting, 10-K

1. Introduction

The main goal of reporting in the financial system is to ensure that high-quality, useful information about the financial position of firms, their performance and changes in their financial position is available (IASB Framework 2015) to a wide range of users, including existing and potential investors, financial institutions, employees, the government, etc. The central element of the formal system of financial reporting is accounting standards. Common accounting standards increase transparency and comparability of the information that firms communicate to users (investors). Transparency decreases uncertainty about the future prospects of the firm, and the information asymmetry between a firm and external stakeholders; and better understanding of the reporting process assures higher transparency. Here, as part of an ongoing project (FORMICA, 2017), we propose a study of diachronic lexical changes in annual reports; by examining how key terms are used and how this usage changes over time, we hope to gain more insight into how language used in reporting reflects and is affected by the financial cycle.

We collected a corpus of 10-K forms from 30 companies from the Dow Jones Industrial Average (DJIA) index for the period from 1996 to 2015, thus including the period c.2007-8 of the most severe economic and financial crisis since the Great Depression in the 1930s. Here, our initial study focuses on developing suitable methods to automatically characterise word usage and meaning, and track changes over time. We focus initially on a small set of words expected to vary with the economic cycle, and apply methods from distributional semantics, checking the suitability of these methods by deriving year-specific word embeddings and examining diachronic changes in the lexical associations they represent, investigating pairs of terms with strong increases or decreases in association over time.

The paper is structured as follows. After the related work in Section 2, we describe our corpus of annual reports in Section 3. Section 4 presents the selected financial terms for this initial study, and Section 5 the methodology used to discover diachronic changes. After the discussion of results and some tentative qualitative explanations in Section 6), we conclude and present ideas for future work in Section 7.

2. Related Work

2.1. Analysing financial reports

Formal reports contain both strictly regulated, financial sections and unregulated, narrative parts. While the financial aspects have seen a large amount of academic research, studies on narrative parts are relatively scarce. Non-financial information from reports has been used for prediction of financially relevant events (Qiu et al., 2006), such as next year performance (through indicators such as return on equity) (Qiu et al., 2006; Butler and Kešelj, 2009; Kogan et al., 2009; Balakrishnan et al., 2010; Hájek and Olej, 2013; Leung et al., 2015), contemporaneous returns around filing dates (Feldman et al., 2008), stock return volatility (Li, 2010; Loughran and McDonald, 2011), earnings forecast dispersion (Kothari et al., 2009; Loughran and McDonald, 2011), costs of capital (Kothari et al., 2009), financial distress (Hajek et al., 2014), credibility of reports (Athanasakou and Hussainey, 2014) or fraud detection (Goel and Uzuner, 2016).

2.2. Linguistic analysis

Several studies have explored more linguistic aspects, often using a corpus linguistics analysis approach. For example, genre analysis of corporate annual report narratives (U.K. Operating and Financial Review) is proposed by Rutherford (2005), where the authors pay special attention to the "Pollyanna effect" (language biased towards the positive terms). Aiezza (2015) studied the use of verbal markers of forward-looking statements and their contribution to the creation of an ethical image in corporate social responsibility (CSR) reports. Impression management in chairman's statements has been analyzed by Merkl-Davies et al. (2011) and a corpus analysis of stance expression is proposed in Fuoli (2017). Very relevant for our work are analyses with a diachronic aspect, including a diachronic analysis of persuasive language in earnings calls (Camiciottoli, 2017).

2.3. Distributional semantics

One of the most visible trends in the field of natural language processing in recent years is the use of distributed lexical representations in the form of vectors or *word embeddings* learned from observed distributions in raw text. The vectors may be derived directly from observed co-occurrence probabilities, or learned (usually with neural networks) to capture this information implicitly; see e.g. (Baroni et al., 2014; Clark, 2015) for overview and comparison of methods. These representations capture many aspects of word meaning (Firth, 1959), including not only judgements of semantic similarity and relatedness but higher-level regularities including limited kinds of analogy (Mikolov et al., 2013). Word embeddings have been also applied to analyze diachronic semantic changes. For example, Hamilton et al. (2016) use neural network-based embeddings to detect shifts in meaning of words in the Google Books corpus, while Kenter et al. (2015) use similar methods for monitoring shifts in vocabulary over time.

3. Corpus of Annual Reports

We focus on companies from the Dow Jones Industrial Average 30 (DJIA) and use their annual (10-K) reports (FORMICA, 2017). The reports cover the period from 1996 to 2015, but the entire period is not covered for all the companies (depending on the availability of the reports in the EDGAR database). We do not consider the amendments (forms of type 10-K/A and 10-K405/A).

Formal reports contain both strictly regulated, financial sections, and less regulated, narrative parts. In our work we focus on the latter, as our interest is in changes in language used in the reporting process, and therefore extract from the 10-K reports only Part I and Items 7 and 7A from the Part II. For example, Item 7 (Management's Discussion and Analysis (MD&A)), discloses company operations and management in a way that is easy for investors and other interested parties to understand and includes information on what the company does in the face of risks, legislation, competition

For extraction of the selected parts, and cleaning of the dataset, we follow Smailović et al. (2017). In short, the desired document parts are detected by searching for the titles of the sections (e.g., Part I), but taking care that the references to these parts are not considered as titles; we also skip potential .pdf, .xls, .jpg, .zip, .gif objects and tables, and remove html/xml tags to leave plain text (see Smailović et al. (2017) for full details). In total, the dataset contains 528 annual reports, as it can be seen from Table 1.

4. Financial Terms

In this initial methodological investigation for our diachronic study, we manually defined a set of financial terms for examination, rather than attempting to extract them automatically (e.g. on the basis of term relevance or change) so as to avoid domain- or sector-specific terms. Some were very general (*'risk'*, *'profit'*, *'loss'*, *'cash'*); some more specific (*'impairment'*, *'dividend'*, *'repurchase'*, *'residual'*, *'capitalization'*, *'development'*, *'expenditure'*, *'discount'*), and selected as expected to vary with the economic cycle. Our dataset covers arguably the most severe period of economic and financial crisis since the 1930s Great Depression. During this period, past investment mistakes on the part of firms had to be recognized in financial statements, via an accounting procedure called asset *impairment*. Firms must compare the values at which their investments are recorded in statements of financial position (balance sheet) with the value in use and the replacement value; during the crisis, these comparisons result in reporting bottom-line losses. While the procedure is highly discretionary — managers may exploit the resulting write-offs for benefits other than shareholder value maximization — research shows that the signal is viewed as credible by the market in general (Riedl, 2004). Even in empirical environments where the discretionary component may be large, write-offs still indicate declining future performance (Kosi and Valentincic, 2013).

Firms — even the good ones — consequently struggle with signalling their true state to the market. Those in good shape either may return cash to their shareholders, by increasing *dividends* (the financial sector tended to do this during the period in question) or by increasing share *repurchases*. Early research suggested that management prefers to keep dividends stable over time, believing that investors prefer stable dividends and themselves preferring to formulate dividend policies as a fixed percentage of net income (Lintner, 1956). There is a clear asymmetry in dividend increases and dividend decreases: Allen and Michaely (2003) report that only about 5% of dividend changes over 30 years were decreases. From this, various studies report evidence consistent with dividend signalling, e.g. Michaely et al. (1995) report that dividend initiations result in a 3.4% increase in share price, while dividend omissions result in a 7% decrease (note that dividend displacement theory predicts a one to one relation in the same direction (Rees and Valentincic, 2013)). Similar findings have been reported for various other settings, for example large dividend increases and large dividend cuts (Grullon et al., 2002). Share repurchases are also a form of payout and may also be used as signals (Brav et al., 2005). Typically, a firm might repurchase shares on the open market when the managers see the share as undervalued (Brav et al., 2005). However, the commitment to repurchases is less firm. Firms request shareholder permission for the maximum amount they intend to repurchase, but do not then necessarily use the full amount (see e.g. (Berk and DeMarzo, 2014), pp. 610-611). Over time, the prevalence of dividends has been declining both in frequency and in amount, while the importance of repurchases has increased. Before the financial crisis, repurchases for US industrials represented two thirds of firms' aggregate payout to shareholders (Floyd et al., 2015). However, this was reversed during the financial crisis and the importance of dividends has increased again.

5. Method

We divided the corpus into collections for each year, taking the stated fiscal year end in each 10-K report as the year of note; this resulted in the frequencies shown in Table 1. We then used a neural network-based method, GloVe (Pennington et al., 2014), to learn word embedding vectors. We performed simple sentence segmentation based on sentence-final punctuation (`.`/`!`/`?`), and tokenised into words on white space and any non-alphanumeric characters (including remaining punctuation). We used NLTK's WordNet-based lemmatiser to reduce all nouns to their singular version (our selected terms of likely interest were all nouns in this study – other parts of speech were left unchanged). All text was normalised into lower case, embeddings used 100 dimensions, and we trained the models for 50 epochs using a learning rate of 0.05.[1] As a comparison point (see below) we also built a count-based vector space based on positive pointwise mutual information (PPMI), following e.g. (Milajevs et al., 2014), using the 2,000 most common words as the vector dimensions. For both spaces, we experimented with a range of co-occurrence context window sizes of 5, 10 and 20 words; previous work has found that this can affect what is captured by word vector relations (with narrower windows sometimes more likely to capture semantic *similarity* while wider ones reflect semantic *relatedness* (Agirre et al., 2009; Turney et al., 2010)) although this seems dependent on corpus and corpus size (Kiela and Clark, 2014).

Year	N	Year	N	Year	N	Year	N
1996	12	2001	28	2006	29	2011	30
1997	17	2002	29	2007	29	2012	30
1998	24	2003	29	2008	30	2013	30
1999	25	2004	29	2009	30	2014	31
2000	27	2005	30	2010	30	2015	9

Table 1: Document-year counts

We learned word vectors for each year independently; note that this means that vectors cannot be compared directly between years (as the latent dimensions of a GloVe vector space are arbitrary). In future work, we plan to learn transformations to align the vector spaces between years, thus allowing direct comparison, following e.g. (Hamilton et al., 2016). Here, we examine only the similarity between pairs of vectors as measured by cosine distance: this can be compared between years, as GloVe learns vectors whose dot-products correspond to ratios of empirically observed co-occurrence probabilities, and the normalisation in the cosine distance calculation accounts for effects of overall word frequency changes. We confirm our observations by comparing with results from the count-based PPMI vector space; although sparser and harder to interpret without further smoothing, this space is directly comparable between years as dimensions are consistent (being derived from co-occurrence counts with a fixed set of context words).

Our initial method is now to look for diachronic changes in similarity (or association) between words that have a high degree of positive association at some point in time. (Searching for apparent changes in associations with consistently *low* absolute values is of course subject to issues of noise and estimation error, and is harder to interpret intuitively; examining changes in associations with high *negative* values – i.e. dissociations or dissimilarities – is potentially useful and may be investigated in future). Given a candidate word w for investigation (see next section), a lexical neighbourhood L can now be discovered, defined as the set of words which appear in $N_y^{10}(w)$, the set of 10 nearest neighbours of w in any year y:

$$L = \bigcup_{y \in Y} \{w' | w' \in N_y^{10}(w)\}$$

We can now examine changes in similarity S between w and members of L over time, by examining changes in the dot-product (or its length-normalised equivalent, cosine distance) between the vectors $\overline{w}$ and $\overline{w'}$ for any $w' \in L$.

$$S = \frac{\overline{w}.\overline{w'}}{|\overline{w}| \times |\overline{w'}|}$$

6. Results and Discussion

Figure 1 shows an example of the diachronic patterns that can be observed, here for two of our candidate words *'dividend'* and *'repurchase'*. Remembering that the similarity measure here is cosine distance between word vectors, with those vectors derived from observed co-occurrence patterns within 10-word windows, we can interpret these patterns as telling us about words which become more (or less) strongly associated with each other over time.

Changes in context window size make some difference to the measured associations between lexical items, but little difference to the patterns of change in associations. As Figure 1(a-c) show, the 10- and 20-word windows show very similar results, both in terms of level of similarity and pattern of changes over time; the smallest 5-word window diverges from the other two slightly, but shows a similar pattern. As the narrower window is more likely to suffer from data sparsity in this relatively small corpus, we use the wider windows hereafter. Comparing the results using the GloVe method (Figure 1(a-c)) with the equivalent using explicitly co-occurrence-based PPMI where the vector dimensions are fixed across years (Figure 1(d)) again shows similar patterns and magnitude of change over time, but with more noise (probabilities used in the PPMI calculation were not smoothed). We therefore take this as our general method for examining the similarity (or lack thereof) in the usages of words over time.

This is of particular interest in this case, as the words *'dividend'* and *'repurchase'* refer to alternative ways in which firms can distribute profits. Dividends tend to be "fixed" – not necessarily by amount but by a fixed percentage of growth, fixed percentage of profits, or declared to be a residual after investment has been taken care of. Changing this policy can therefore send a strong signal to investors, and is therefore often strenuously avoided. Repurchases, on the other hand, are more flexible — shareholders are not forced to give up their shares in exchange for cash, but only if they wish to do so — and this "un-fixedness" can

[1] Trials with 40 and 60 epochs show similar results; a more comprehensive test will be carried out in future.

make repurchases popular with companies as changing the amounts repurchased does not tend to send strong signals. The increase in association between these words over time is statistically significant (Spearman's R shown in Figure 1, $p < 0.05$ in all cases), and suggests that there is an increasing tendency for firms to use these words in similar ways (i.e. in similar lexical contexts) when reporting. Note that simple direct measures of association do not reveal these patterns: PMI between the two words directly (measured via co-occurrence in the same 20-word context window) shows no significant correlation over time – see Figure 4.

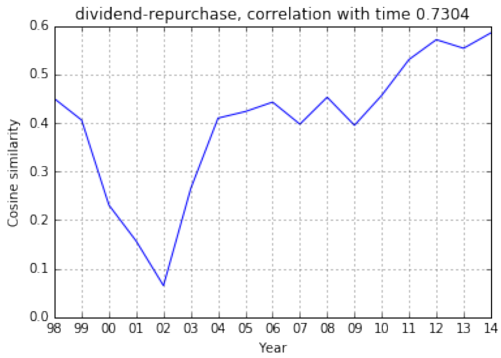
As Figures 2 and 3 show, we can use this method to look for the major diachronic changes across the lexical neighbourhood L more generally, by searching for the words in L whose similarities show large changes over time (here, we used Spearman's R to find the highest correlations with the year ordering). Each figure shows the top 6 positive correlations over time (increases in similarity) and top 4 negative correlations (decreases in similarity) with one of our words of interest; in Figure 2, these are changes in the similarity with the word *'dividend'*; in Figure 3, with *'impairment'*.

Inspecting Figure 2 (*'dividend'*), we can perhaps offer some tentative qualitative explanations. For (a),(b) and (f) (*'quarterly'*, *'shareholder'*, *'paying'*): before the financial crisis, we might expect the association between the two terms to be low, as dividends tended to be replaced by share repurchases (both in frequency and amount). During and after the financial crisis, this association increases significantly. This is possibly due to companies trying to signal to shareholders that their current and expected future profits are sound and can be thus distributed. The same most likely applies to (c) (*'declared'*), although it is unclear why. Dividends have always needed to be declared first with the on-record date, ex-dividend date and payment date (or interval) defined. The positive correlation is possibly due to dividends becoming more prominent in this period and hence the term *'declared'* becoming more frequent as a result.

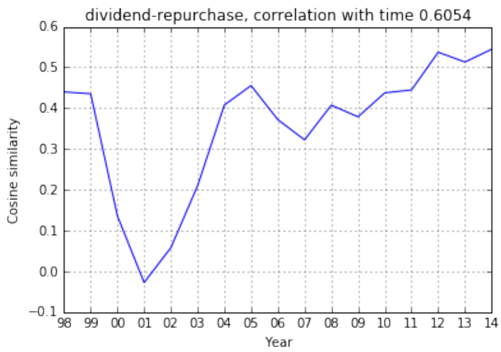
For some cases e.g. (e) (*'production'*) we offer no explanation. Whether production (of physical goods) has actually increased or decreased relative to providing services, or whether *'production'* in this case refers to other concepts will need to be investigated further. Similarly we currently have no insights into the pattern in (i) (*'conversion'*).

For (h) (*'gain'*) and possibly (g) (*'impact'*), as the importance of dividends increased during this period relatively to trends in previous periods (see (Floyd et al., 2015)), the reverse holds for capital gains. If a firm pays out a relatively high proportion of profits as dividends, then share prices will not increase as much as if a firm pays a low proportion of profits in the form of dividends. Hence, the higher the proportion of total return a shareholder receives in the form of dividends, the smaller the proportion of total return in the form of capital gain (in relative terms). Hence the decreasing correlation through time between dividends and capital gains. As dividends were discussed more, gains less.

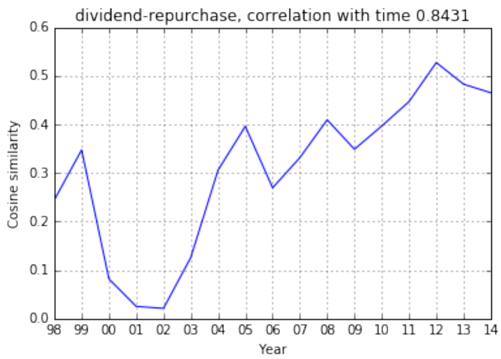
Inspecting Figure 3 (*'impairment'*), we can suggest similar explanatory background. For (a) (*'recognize'*) and (b) (*'testing'*), in the period under study, the first break was the Enron scandal and consequent introduction of the Sarbanes-Oxley act. This brought about an increase in con-



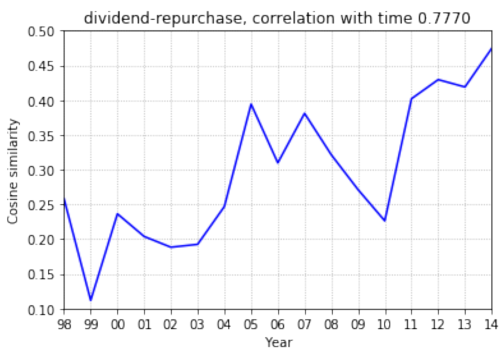
(a) GloVe, context window 20 words



(b) GloVe, context window 10 words

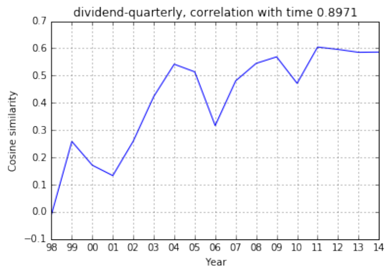


(c) GloVe, context window 5 words

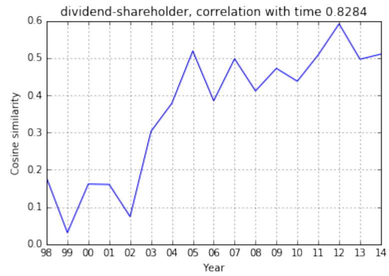


(d) PPMI, context window 20 words

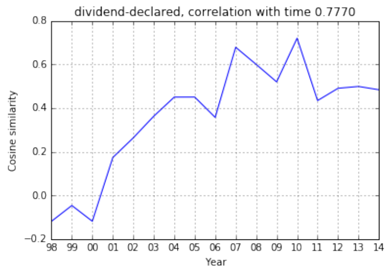
Figure 1: Cosine similarities for the word pair *'dividend'* vs *'repurchase'* over time, using a range of methods and lexical co-occurrence context window sizes.



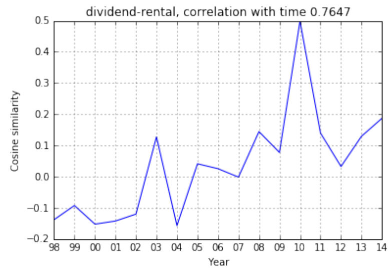
(a) ‘dividend’ vs. ‘quarterly’



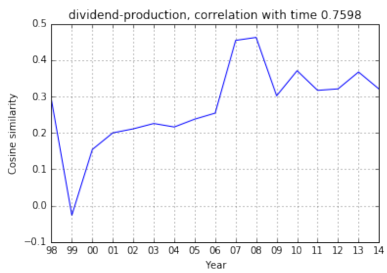
(b) ‘dividend’ vs. ‘shareholder’



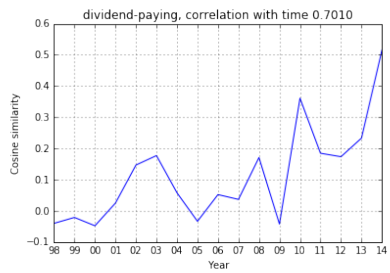
(c) ‘dividend’ vs. ‘declared’



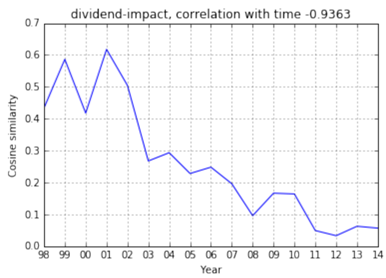
(d) ‘dividend’ vs. ‘rental’



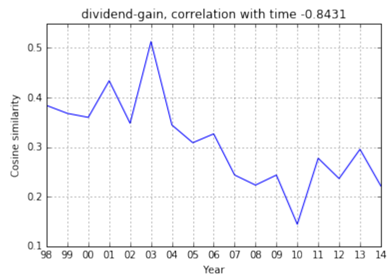
(e) ‘dividend’ vs. ‘production’



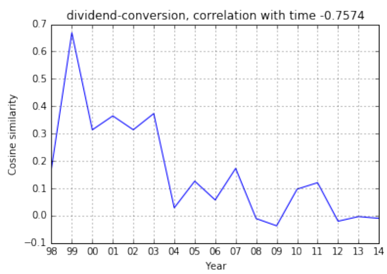
(f) ‘dividend’ vs. ‘paying’



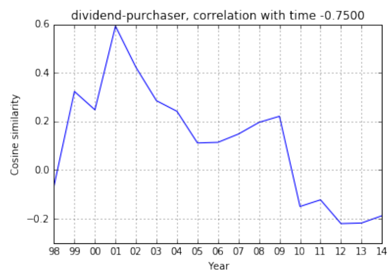
(g) ‘dividend’ vs. ‘impact’



(h) ‘dividend’ vs. ‘gain’

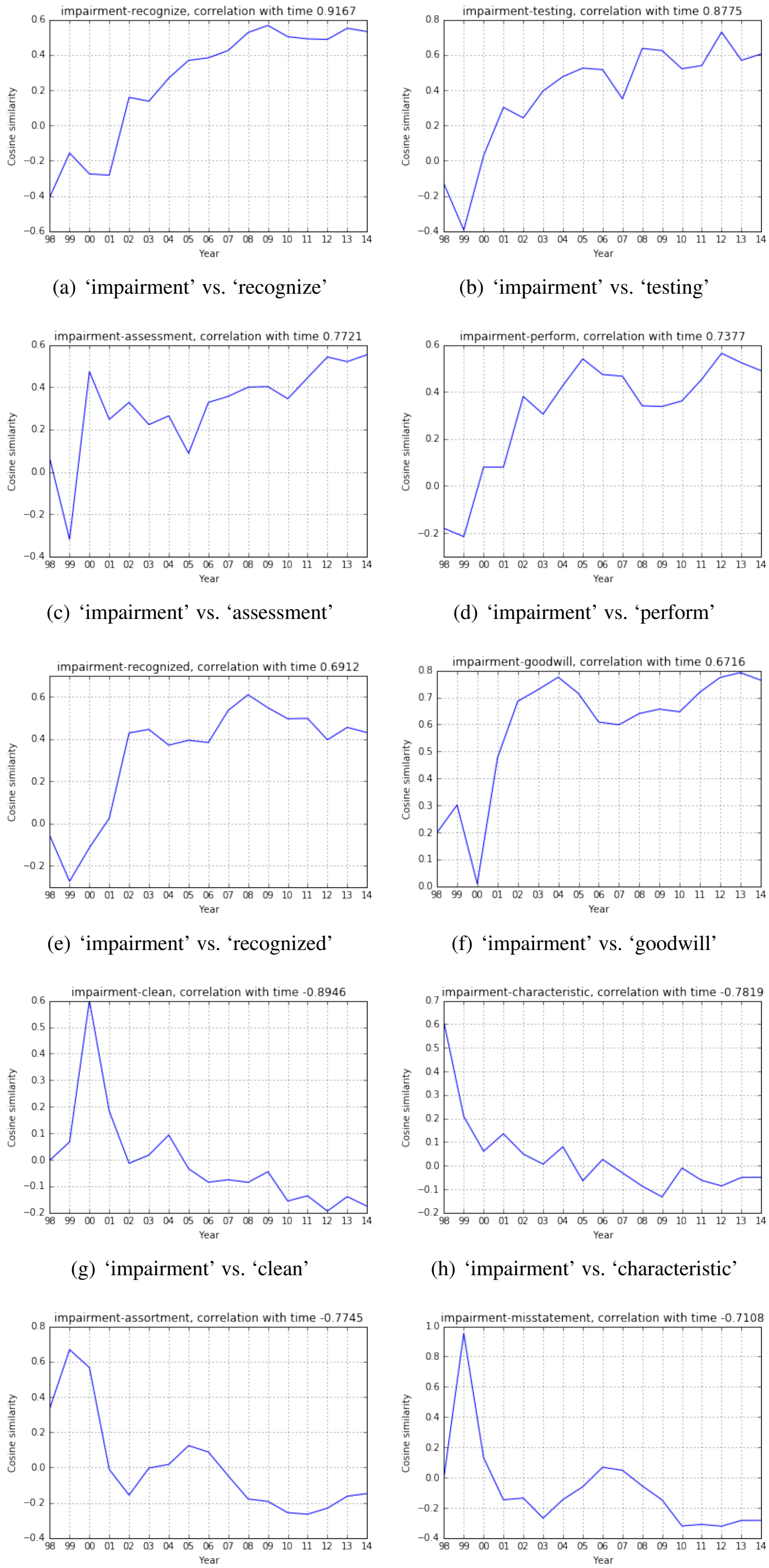


(i) ‘dividend’ vs. ‘conversion’



(j) ‘dividend’ vs. ‘purchaser’

Figure 2: Cosine similarities for word pairs with highest positive and negative correlations over time, for the lexical neighbourhood of ‘dividend’.



(a) ‘impairment’ vs. ‘recognize’

(b) ‘impairment’ vs. ‘testing’

(c) ‘impairment’ vs. ‘assessment’

(d) ‘impairment’ vs. ‘perform’

(e) ‘impairment’ vs. ‘recognized’

(f) ‘impairment’ vs. ‘goodwill’

(g) ‘impairment’ vs. ‘clean’

(h) ‘impairment’ vs. ‘characteristic’

(i) ‘impairment’ vs. ‘assortment’

(j) ‘impairment’ vs. ‘misstatement’

Figure 3: Cosine similarities for word pairs with highest positive and negative correlations over time, for the lexical neighbourhood of ‘impairment’.

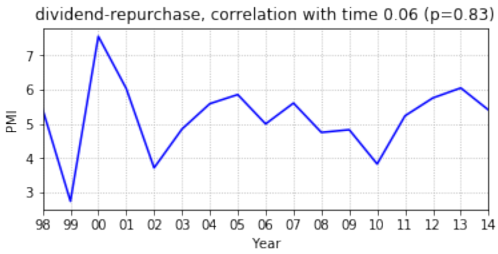


Figure 4: PMI for *'dividend'* vs *'repurchase'* over time.

servatism in preparation of financial statements, including recognizing all possible losses, but only realized gains. If a firm suspected a loss might occur in the future — and during the financial crisis they did suspect this a lot — they would have to recognize the present values of diminished expectations about future cash flows in current financial statements via recording (recognizing) an impairment in financial statements. Assets must be tested for impairment. This is generally done annually, although the regulation is more detailed and worded differently. However, some assets such as goodwill and other intangible assets must be tested for impairment rather than depreciated via "regular" depreciation expense in financial statements. The importance of these assets has generally increased through time, so financial crisis or otherwise, impairment and test would go hand in hand. The same explanation would account for (c) (*'assessment'*) which is an alternative term for *'testing'*; and (e) and (f) (*'recognized'* and *'goodwill'*) which are both related to goodwill and impairment testing.

The final term, (j) (*'misstatement'*) is particularly interesting. A "hump" can be observed with the term association increasing over the pre-crisis years, peaking with fiscal year ends 2006-7, and then decreasing down to a minimum in 2010. This could be due to firms correcting (possibly deliberate) mis-statements in financial statements from the pre-crisis years, as these were dug out by auditors, and recognised in the financial statements. Firms with mis-statements would also often record an impairment, as both are related to firms being too optimistic about their future prospects in the pre-crisis years. After 2010, this effect would therefore not be expected to be as pronounced as before.

7. Conclusion

Although only an initial methodological investigation, this study suggests that the use of word embeddings in a diachronic corpus can give some useful insights into terms used in financial reporting. Using GloVe provides a method to investigate changes in lexical associations which has revealed some intuitive relationships, while discovering others which warrant further investigation in the corpus data to understand the patterns. In future work we plan to extend this study in several ways: first, to use corpus analysis to explore the original context of the terms analysed to help understand the correlations more clearly; second, to explore more specific hypotheses from economic theory and financial research about term relations and changes; and third, to generalise the approach to automatically extend the list of terms of interest, discovering relationships in a more unsupervised fashion.

Our method is currently limited to analysing direct pairwise associations between words: more general properties of the word embeddings, including directions and magnitudes of movements in the general space of word meanings, cannot be derived when training GloVe models separately for each year as here. In future work, we will further investigate the use of the explicitly consistent spaces in the count-based PPMI variant by incorporating more appropriate smoothing, and the use of learned transformations between year-based spaces to make GloVe models consistent, following (Hamilton et al., 2016). Given a suitable model and dataset, it would ultimately be interesting to examine the relationship between terminological usage and companies' financial performance, via descriptive or predictive models.

8. Acknowledgements

The authors acknowledge the financial support from the Slovenian Research Agency for research core funding (No. P2-0103 and No. P5-0161), as well as for funding of the research project *Influence of formal and informal corporate communications on capital markets* (No. J5-7387). We also thank Jasmina Smailović and Martin Žnidaršič for the corpus preparation phase and Rok Spruk for the discussion during the design of the study.

9. Bibliographical References

Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Paşca, M., and Soroa, A. (2009). A study on similarity and relatedness using distributional and WordNet-based approaches. In *Proceedings of HLT-NAACL*, pages 19–27.

Aiezza, M. C. (2015). "We may face the risks"..."risks that could adversely affect our face." A corpus-assisted discourse analysis of modality markers in csr reports. *Studies in Communication Sciences*, 15(1):68–76.

Allen, F. and Michaely, R. (2003). Payout policy. In G M Constantinides, et al., editors, *Handbook of the Economics of Finance*. Elsevier.

Athanasakou, V. and Hussainey, K. (2014). The perceived credibility of forward-looking performance disclosures. *Accounting and Business Research*, 44(3):227–259.

Balakrishnan, R., Qiu, X. Y., and Srinivasan, P. (2010). On the predictive ability of narrative disclosures in annual reports. *Eur. J. Operational Research*, 202(3):789–801.

Baroni, M., Dinu, G., and Kruszewski, G. (2014). Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of ACL*, pages 238–247.

Berk, J. and DeMarzo, P. (2014). *Corporate Finance*. Pearson, 3rd edition.

Brav, A., Graham, J. R., Harvey, C. R., and Michaely, R. (2005). Payout policy in the 21st century. *Journal of Financial Economics*, 77(3):483–527.

Butler, M. and Kešelj, V. (2009). Financial forecasting using character n-gram analysis and readability scores of annual reports. In *Canadian Conference on Artificial Intelligence*, pages 39–51.

Camiciottoli, B. C. (2017). Persuasion in earnings calls: A diachronic pragmalinguistic analysis. *International Journal of Business Communication (preprint)*.

Clark, S. (2015). Vector space models of lexical meaning. In Shalom Lappin et al., editors, *Handbook of Contemporary Semantic Theory*, pages 493–522. Wiley-Blackwell.

Feldman, R., Govindaraj, S., Livnat, J., and Segal, B. (2008). The incremental information content of tone change in management discussion and analysis. *Available at SSRN: http://ssrn.com/abstract=1126962*.

Firth, J. R. (1959). A synopsis of linguistic theory, 1930-55. In F. R. Palmer, editor, *Selected Papers of J. R. Firth 1952-59*.

Floyd, E., Li, N., and Skinner, D. J. (2015). Payout policy through the financial crisis: The growth of repurchases and the resilience of dividends. *Journal of Financial Economics*, 118:299–316, November.

Fuoli, M. (2017). Building a trustworthy corporate identity: A corpus-based analysis of stance in annual and corporate social responsibility reports. *Applied Linguistics*, page amw058.

Goel, S. and Uzuner, O. (2016). Do sentiments matter in fraud detection? Estimating semantic orientation of annual reports. *Intelligent Systems in Accounting, Finance and Management*, 23(3):215–239.

Grullon, G., Michaely, R., and Swaminathan, B. (2002). Are dividend changes a sign of firm maturity? *The Journal of Business*, 75(3):387–424.

Hájek, P. and Olej, V. (2013). Evaluating sentiment in annual reports for financial distress prediction using neural networks and support vector machines. In *Intl. Conf. Engineering Applications of Neural Networks*, pages 1–10.

Hajek, P., Olej, V., and Myskova, R. (2014). Forecasting corporate financial performance using sentiment in annual reports for stakeholders' decision-making. *Technological and Economic Development of Economy*, 20(4):721–738.

Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. In *Proc. ACL*, pages 1489–1501.

Kenter, T., Wevers, M., Huijnen, P., and de Rijke, M. (2015). Ad hoc monitoring of vocabulary shifts over time. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 1191–1200.

Kiela, D. and Clark, S. (2014). A systematic study of semantic vector space model parameters. In *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC)*, pages 21–30.

Kogan, S., Levin, D., Routledge, B. R., Sagi, J. S., and Smith, N. A. (2009). Predicting risk from financial reports with regression. In *Proceedings of HLT-NAACL*, pages 272–280.

Kosi, U. and Valentincic, A. (2013). Write-offs and profitability in private firms: Disentangling the impact of tax-minimisation incentives. *European Accounting Review*, 22(1):117–150.

Kothari, S., Li, X., and Short, J. E. (2009). The effect of disclosures by management, analysts, and business press on cost of capital, return volatility, and analyst forecasts: A study using content analysis. *The Accounting Review*, 84(5):1639–1670.

Leung, S., Parker, L., and Courtis, J. (2015). Impression management through minimal narrative disclosure in annual reports. *British Accounting Review*, 47(3):275–289.

Li, F. (2010). The information content of forward-looking statements in corporate filings - a naïve bayesian machine learning approach. *Journal of Accounting Research*, 48(5):1049–1102.

Lintner, J. (1956). Distribution of incomes of corporations among dividends, retained earnings, and taxes. *The American Economic Review*, 46(2):97–113.

Loughran, T. and McDonald, B. (2011). Barron's red flags: Do they actually work? *Journal of Behavioral Finance*, 12(2):90–97.

Merkl-Davies, D. M., Brennan, N. M., and McLeay, S. J. (2011). Impression management and retrospective sense-making in corporate narratives: A social psychology perspective. *Accounting, Auditing & Accountability Journal*, 24(3):315–344.

Michaely, R., Thaler, R. H., and Womack, K. L. (1995). Price reactions to dividend initiations and omissions: Overreaction or drift? *The Journal of Finance*, 50(2):573–608.

Mikolov, T., Yih, W.-t., and Zweig, G. (2013). Linguistic regularities in continuous space word representations. In *Proceedings of NAACL-HLT*, pages 746–751.

Milajevs, D., Kartsaklis, D., Sadrzadeh, M., and Purver, M. (2014). Evaluating neural word representations in tensor-based compositional settings. In *Proceedings of EMNLP*, pages 708–719.

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of EMNLP*, pages 1532–1543.

Qiu, X. Y., Srinivasan, P., and Street, N. (2006). Exploring the forecasting potential of company annual reports. *Proceedings of the American Society for Information Science and Technology*, 43(1):1–15.

Rees, W. and Valentincic, A. (2013). Dividend irrelevance and accounting models of value. *Journal of Business Finance & Accounting*, 40(5-6):646–672.

Riedl, E. J. (2004). An examination of long-lived asset impairments. *The Accounting Review*, 79(3):823–852, July.

Rutherford, B. A. (2005). Genre analysis of corporate annual report narratives: A corpus linguistics-based approach. *The Journal of Business Communication*, 42(4):349–378.

Smailović, J., , Žnidaršič, M., Valentinčič, A., Lončarski, I., Pahor, M., Martins, P. T., and Pollak, S. (2017). Automatic analysis of financial annual reports: A case study. *Computación y Sistemas (Special Issue on Advances in Human Language Technologies)*, 21(4):809–818.

Turney, P. D., Pantel, P., et al. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37(1):141–188.

10. Language Resource References

FORMICA. (2017). *FORMICA Project Corpus (v0.9)*. Project dataset, currently not publicly released.

Mapping Deep NLP to Knowledge Graphs: An Enhanced Approach to Analyzing Corporate Filings with Regulators

Damir Cavar, Matthew Josefy
Indiana University
Bloomington, IN
{dcavar,mjosefy}@iu.edu

Abstract

Filings submitted by companies to the Securities and Exchange Commission provide a tremendous corpus for application of advanced natural language processing techniques. While business scholars actively utilize these texts, interdisciplinary efforts hold substantial promise to advance knowledge and techniques. In this study, we utilize deep natural language processing techniques to extract meaningful knowledge from SEC filings. We construct a comprehensive pipeline that extracts the original filings, processes them in order to recognize component segments for distinct analysis, feeds each text through multiple NLP processors to obtain optimal recognition of the linguistic properties, and ultimately seeks to construct a comprehensive knowledge graph of how companies, their executives and their directors are linked to one another, or how various risks are identified, weighted, and handled over long periods of time. We thus link advanced NLP techniques and knowledge graphing approaches, contributing greater domain specific knowledge to advance linguistic approaches and potentially discovering underlying networks that would be difficult to detect with other approaches.

Keywords: Deep NLP, Topic Modeling, Knowledge Graphs

1. Introduction

We describe our research activities in the domain of engineering and tuning Natural Language Processing (NLP) technologies for the analysis of business reports. The goal of this project is to provide data and detailed analyses of corporate reporting, focusing initially on the Securities and Exchange Commission (SEC) reports, while constructing our pipeline and process in such a way as to deploy in future to also analyze international financial reporting sources in similar fashion.

The SEC is the body in the U.S. charged with ensuring fair and transparent markets. To that end, the SEC requires publicly-listed firms in the U.S. (and other firms who meet certain thresholds) to comply with disclosure requirements. These reports become publicly-available through SEC EDGAR (`https://www.sec.gov/edgar.shtml`) which serves as a repository of over 21 million corporate filings. Information within these reports is often market-sensitive. Indeed, recent research found that certain subscribers who had access to the SEC's public dissemination system (PDS) were able to profit merely by having access to information 30 seconds before it was made available to the public on EDGAR (Jonathan L. Rogers and Zechman, 2017).

The SEC was founded in 1933 and its electronic repository dates back to 1984. Accordingly, longitudinal data is available dating back decades, allowing the potential to study networks of firms and the executives and board members associated with them. Thus, in our current work, one key aim is to map management and board members of firms in a network of relations over time. Thus our project uniquely integrates aspects of deep processing to harvest information with advanced network maps to ascertain previously obscured relationships between focal nodes, which in this case may be either individuals or companies. The filings contain detailed descriptions of individuals involved in the firms in specific roles, as well as the description of corporate relations between firms. Mapping of individuals and their roles and relations to other individuals and institutions over time allows us to study relations using complex network analyses (Newman et al., 2006). Ultimately, these networks may be useful in understanding the survival and failure of firms (Josefy et al., 2017).

Besides network relations between individuals over time, a specific interest area is the analysis and mapping of perceived risks and the corresponding risk management strategies, also along the time axis, as reported by the firms across different business types. The annual reports of firms filed on Form 10-K contain a detailed picture of the risks a firm faces at the time of the reporting. An analysis and classification of the particular risks and the arrangement along the time axis provide an extremely valuable analytical instrument for the understanding of the evolution of risks and risk management strategies in different industry sectors, as well as a correlation with national and geopolitical historical events and developments.

These main interest areas motivated our approach to utilize advanced Natural Language Processing (NLP) and Topic Modeling (Blei, 2012) methodologies for the analysis of the different types of SEC filings.

The filings are available in common accessible formats, e.g. raw HTML or the standardized eXtensible Business Reporting Language (XBRL) (Engel et al., 2013) XML format. XBRL is an XML language that is freely available as a global standard for exchanging business information. Independent of the formats, the document content is arranged in ways specific to the filing institution, and not in a strictly predefined order. This requires either manual processing and annotation of the content by section and paragraph, or automatic classification tools have to be developed that detect sections by content.

The lack of a uniform structure and clear semantic content specification in the filings implies that the target content (e.g. director biographies, work histories and board description, risk management sections) needs to be parsed and analyzed using Machine Learning, Document Classification, and Natural Language Processing (NLP) technologies. On the one hand, the relevant sections and paragraphs have to be identified. On the other hand it is necessary to map parsed and identified concepts, relations and other properties of concepts, and other semantic properties of the content and meta-information onto a structured data representation for subsequent analysis and processing. Common NLP technologies that are freely accessible come with serious limitations and unacceptable error rates when applied to complex business language in the specific document types. We developed various NLP components and pipelines to remedy this.

In the following we describe in more detail our approach to the problem of the identification of target content in semi-structured business reports of the SEC type, as well as the architecture and technologies that we developed and utilized to maximize the quality of the NLP results for knowledge mapping to graph representations and deep linguistic topic modeling.

2. Previous Work

When looking at previous work in the domain of interest, we should differentiate between a.) the analysis of SEC reports using NLP, b.) specific NLP technologies tuned for the particular task and business language, c.) processing of the particular information (e.g. extraction of relations between individuals and institutions mapped to time, analysis of risks and risk management over time), and d.) mapping the resulting information on knowledge graphs and advanced knowledge information systems.

While we are more than certain that many business make use of NLP technologies for processing of business documentation, financial reports, and other open or closed reports by firms, we can only comment on information accessible to us in form of publications or free and open toolkits and technologies.[1]

While there are commercial APIs that provide access to the SEC filings, these filings are also publicly available to any user. Therefore, we will not discuss details of such commercial systems here. There are numerous open modules and toolkits to process and access the SEC EDGAR data, a search online will reveal all the relevant sites and information. We will not go into detail here. There are a few resources and APIs that allow access to the SEC EDGAR repository or analyses of the bulk data.

The *CorpWatch API* (`http://api.corpwatch.org/`), for example, "uses automated parsers to extract the subsidiary relationship information from Exhibit 21 of companies' 10-K filings." There are various Python modules that facilitate crawling and downloading of the filings, as for example the Python module SEC-Edgar, which implements a Sphinx crawler.

[1]For example, *Intelligize* (`www.intelligize.com`), a LexisNexis company, offers services related to the SEC EDGAR data set, that we have not analyzed in detail.

Arelle (`arelle.org`) (Fischer and Mueller, 2011b; Fischer and Mueller, 2011a; Fischer, 2013) is a large project that provides a set of tools and applications geared for the analysis and processing of the XBRL format of the SEC EDGAR reports (and other filings using the XBRL format). It is an open source platform available for all the common operating systems. It does not focus on content analysis of the XBRL format filings, it rather focuses on well-formedness checks and semantic validation given numerous XBRL-related standards. While *Arelle* is a powerful environment to process and search XBRL documents, to our knowledge it has no capabilities to include advanced NLP technologies for content analysis and graph databases for knowledge graph representations of content.

A more detailed description of the available NLP components, pipelines and technologies is presented in the following section, where we discuss the common technologies and their application in our specific scenario.

The business literature has demonstrated interest both in director networks and in the risks identified by firms, often relying on proprietary, curated databases focused on large firms. However, to the best of our knowledge, we are not aware of any publication that employs techniques such as our to accomplish these analytical goals at scale, i.e. a time mapping of relations between individuals and institutions, and the topic modeling analysis of identified risks and risk management strategies mapped onto the time axis over SEC filings. We welcome further feedback on additional research questions that this project's goals and methods may be well suited to address.

3. Implementation

Each of the SEC filings is composed of multiple sections, each of which may provide information of different scholarly interest. Thus, we employ Machine Learning (ML) (Scikit Learn (Buitinck et al., 2013; Pedregosa et al., 2011) and our own set of ML classifiers, e.g. using Bayesian or Support Vector Machine approaches). The models are extracted and the algorithms are trained on the manually annotated corpus that represents the distinct portions of SEC filings, in order to split them into appropriate segments for further analysis. For instance, a typical table of contents for the 10-K would include the following disclosures:

Item 1. Business (Description of the filing company's background, products, strategy and competition), Item 1A. Risk Factors (the factors that could most substantially affect the company's profitability and therefore the risk that shareholders may lose their investment), Item 3 Legal Proceedings, and so on. Additional key items of note include Management's Discussion and Analysis of Financial Condition and Results of Operations (Item 7) and Financial Statements (Item 8). Some items required to be part of a particular form may be incorporated by reference to another form. For instance, certain items in the 10-K are not provided therein, but are instead met by referencing the company's proxy statement.

Given the various sections and the unique content within each, we first pursued steps to delineate the content into sub-corpus that represented these component sections. We discuss these steps now: Creating the Training Corpus, Detection of Content in Paragraphs, and Natural Language Processing.

3.1. Creating the Training Corpus

We used the XML Path Language (XPath) (DeRose and Clark, 1999) to markup an initial training corpus from existing SEC reports manually using only a web-browser and a database table for the selected paragraphs and document reference. Our assumption is that the SEC filing URLs do not change, neither the final document structure and content for every single URL. Thus a corpus definition using XPath and URL tuples seems appropriate.

A team of corpus annotators browsed a selection of the SEC EDGAR reports using a web-browser that provided XPath information for selected paragraphs (e.g. Google Chrome). We collected those XPath descriptions with the document URLs in a data table for the target chapter types. The annotators marked the paragraphs of interest, in one particular case for example the sections of the 20-F which provided biographical information on one or more senior executives or board members of the company. This manual markup provided an XPath specification and a document URL referencing the start and end points of each sentence or paragraph within the document that was in-scope for the specified purpose.

The selected reference pointers to the remote documents were stored in database tables and validated by the corpus annotators.

We developed a crawler to fetch the text portions from the specific documents given the XPath description. The crawler aggregates the text paragraphs and generates a raw text corpus that serves as the training corpus for our machine learning (ML) content detection algorithms. This training corpus was used at the same time as one part of the textual data for extraction of information for the network study of individuals and firms, as well as the risk analysis using a topic modeling framework.

With this configuration we are able to process large volumes of corporate filings and extract the paragraphs of interest.

3.2. Detection of Content in Paragraphs

To train an algorithm to detect the desired content automatically, we used supervised machine-learning approaches. We utilized a Bayesian text similarity algorithm using frequency profiles of unigrams over the target text portions, as well as Support Vector Machines to detect the optimal hyperplane between binary models, e.g. CV vs. non-CV sections or risk-management vs. non-risk-management descriptions.

We assessed the performance of the algorithms on the training texts against the human-classification of the same texts. By automatic and manual modification of model weights assigned to the classes given the N-gram models, we improved the accuracy of the classifiers to predict either “relevant” or “not-relevant.”

In basic evaluations we concluded that the ML-based algorithm can achieve a detection rate for sections with the curriculum vitae of the management or the description of board members at higher than 95%. As for the detection of sections discussing risk factors we achieve an accuracy above 90%.

Automatically detected relevant paragraphs that are not part of the manually generated and validated training corpus, are extracted and saved as a separate corpus for further analysis, including the relevant meta-information as for example the URL and XPath information.

3.3. Natural Language Processing

Next, our project requires the ability to parse the relevant text to extract focal information. We found that available parsers were inconsistent in their treatment of domain, time, and company specific information. Further, our use-case has considerable nuance in regard to pronoun-resolution and each individual may have time-variant characteristics or positions.

To evaluate the quality of the parsing and to ensure we obtained the highest-quality extraction of information, we built an interface to allow comparison of multiple parsing tools, including Stanford CoreNLP (Manning et al., 2014), spaCy,[2] NLTK (Bird et al., 2009) components, and various other NLP technologies.

The different NLP components provide linguistic analytical output for various domains in different quality. The majority of those tools are able to generate Part-of-Speech (PoS) tags for the text tokens (i.e. the words and punctuation marks). The problem is that the PoS-taggers make use of different tagsets, and the tagsets are utilizing a limited number of tags that ignore detailed morpho-syntactic features of high importance for a deep linguistic analysis. CoreNLP and spaCy segment text into sentences and generate for example dependency parse trees[3] for each sentence. These Dependency Trees describe rather shallow relations at the functional level of sentences (the subject and the object of the main predicate, etc.), leaving out essential information about scope and semantic hierarchies of sentential or clausal elements. A segmentation of clauses is not provided, but we were able to translate the Dependency Parse Trees and Constituent Trees that were generated by phrase structure parsers into clause segmentation information. Only CoreNLP provides an essential component for the resolution of anaphora and coreference (Clark and Manning, 2015; Clark and Manning, 2016).

To improve the named entity recognition and to add to the NLP components a sub-classification, we created an extended English morphological analyzer. We collected freely available resources with names of people known in the business world with their titles and roles. Additionally, we extended the analyzer with all variations of the name used in the business documentation. For example, Timothy Donald Cook, the CEO of Apple Inc., is also identified as the same entity referenced by strings like *Timothy* or *Tim Cook*. The morphological analyzer not only generates a named entity label for a **person name**, but also a sub-class ID for the **business domain**, and a unique identifier for the entity. The morphological analyzer includes also a list of

[2] The parser used in spaCy is an implementation of Honnibal and Johnson (2015).

[3] See for example (2014) for the CoreNLP Dependency Parser output.

to us known institutions and firm names with variants, e.g. *Apple Inc.* and simply *Apple*. The domains of institutions are emitted as tags as well, i.e. subtypes like **financial**, **IT**, or business types like **manufacturing** or **service** business. The morphological analyzer is implementation as a bidirectional Finite-state Transducer using the *Foma* (Hulden, 2009) compiler, with the named-entity databases containing extensive lists of people and institutions and detailed type information translated to *Lexc*[4] grammars.

Our goal was to utilize freely available NLP pipelines and components as much as possible, developing parallel processing chains that then validate the output via comparison. We developed uniform wrappers for CoreNLP, spaCy, OpenNLP, some NLTK components, and our own modules, that translate the NLP analyses into a uniform data structure (a Python class instance). The *LingData* class serves as this kind of wrapper. It translates PoS-tags into feature structures represented as Directed Acyclic Graphs. Constituent trees are translated into clause hierarchies and all scope relations between tokens and phrases are mapped to an API, allowing for requests like "does the main clause contain a sentential negation," or "is the embedded clause in the scope of future tense (from the matrix clause)." Such requests are wrapped into method or function calls within the *LingData* class. They are essential for advanced mapping of content at deep linguistic levels to Knowledge Graphs or Representations (using free graph representations or OWL-based ontologies). Dependency Graphs are similarly translated into data structures that can be queried for clause level dependencies, e.g. checking for core relations like *subject* and *object* of a *predicate*, mapping semantic triples from dependency representations. They are also mapped on phrase level dependencies, usually not obvious relations in a dependency parse tree, such that the entire *subject phrase* can be asked for. As long as our NLP components are able to resolve anaphoric expressions, the references for every pronoun and coreferent concepts are made available via methods in *LingData*. Entities and basic concepts in the NLP output are annotated for their synonyms, hypo-, and hypernyms using WordNet (Miller, 1995; Fellbaum, 1998) in NLTK.

Integrating the outputs of NLP components into one uniform data structure allows us to unify the outputs and identify mismatches, generate selection models, and weight those outputs to identify the best representation from often failing or insufficiently specific NLP results.

The mentioned NLP components exhibit systematic errors with specific constructions and sentence types. Common errors occur with constructions containing coordination, complex embeddings, constructions with gapping or ellipsis, and simply with very long sentences.[5]

To overcome issues and limitations with these purely data driven and probabilistic methods, we utilize hybrid approaches as in the Free Linguistic Environment (FLE) project (Cavar et al., 2016). Such approaches allow us to combine grammar engineering with data driven machine learning approaches to improve the NLP performance and generate deep linguistic annotations that go beyond the limited dependency parse or simple constituent structure trees. Our Probabilistic Lexical-functional Grammar approach in FLE combines and links hierarchical structural relations with functional annotations, elementary semantic and morpho-syntactic relations and properties. A detailed description of the deep linguistic analysis/annotation that can be achieved with a parser like FLE (or simpler forms implemented using NLTK components) would go beyond the scope of this abstract. We would be happy to discuss these properties in the full paper version and in the presentation of this work.

Our NLP pipelines are implemented as parallel Remote Procedure Call (RPC) (Microsystems, 1988) services that can be distributed over multiple instances and thus scaled for big data annotation. We provide wrappers in Python (and Go) for the mentioned pipelines. The code is mostly independent of the underlying operating system.

3.3.1. Topic Modeling of Risk Discussions

For topic modeling and analysis of the risk management content we used initially Mallet (McCallum, 2002). The large number of different implementations of topic modeling libraries in different systems allows us to extend our evaluations in various ways. Topic models generated by Mallet provide essentially a set of n groups of tokens from the text that represents n underlying or latent topics. Since the particular topic modeling technologies at the time of our experiments did not provide any systematic strategy to display variation of topics over time, we decided to generate the models for individual reports and generate a mapping of the topic related token list over the time axis by iterating the analyses.

Setting the number of topics to an initial 100, we used Mallet incrementing the optimization intervals that lead to an increasing number of empty topics, which we were inclined to take to be a good indicator for the optimal number of topics in the content of a particular report.

Our goal was to map these results on the different business domains and sectors, to identify time-dependent changes in the perception of risks and emerging mitigation strategies. This is ongoing research that we will report on independently.

3.4. Mapping Concepts and Relations on KRs

Our uniform *LingData* data structures contain detailed information about the structure of every single clause in the analyzed text. We are able to extract sentence internal clause structure, identify clause features and scope relations between clauses, core semantic relations within a clause (e.g. subject – predicate – object), as well as detailed properties of concepts, their semantic and morpho-syntactic features. This information is essential for the mapping of mentioned concepts on KRs.

Our graph-mapping approach includes the extraction of *subject – predicate – object* tuples from raw text. We validate the semantic relations using linguistic analyses, e.g. whether the utterance is embedded in a subjunctive or fu-

[4] See for details on the *Lexc*-formalism the documentation of *Foma* (Hulden, 2009).

[5] We are preparing an independent documentation of the fallacies of the common NLP pipelines and tools, and systematic issues with the inherent probabilistic technologies used in those.

ture tense context, or whether it is in the scope of such a context, or negative operators. The relations that are identified as valid facts and assertive statements are added to the graph, representing the subject and object as concepts, and the predicate as a relation with a time-stamp and meta-information about the source.

As an example, we can differentiate future tense claims like *Facebook will buy Oculus* from past tense assertions like *Facebook bought Oculus*. The latter allows us to extend the KR with a factual assertion of concepts and relations *Facebook* $\rightarrow$ *buy* $\rightarrow$ *Oculus*, while the future tense clause does not justify this mapping. Analyzing the same factual relation in an embedded clause as in *I would not claim that Facebook bought Oculus* in the scope of a negated matrix clause cannot be rendered as a factual assertion.

There are multiple interesting aspects to emphasize here, when it comes to the correct mapping of semantic relations extracted from NLP outputs. Due to space limitations we confine ourselves to this particular example.

As a graph database we utilize commercial products, Neo4J (`neo4j.com`) and Stardog (`www.stardog.com`). While both graph DBs offer similar capabilities with respect to graph storage and retrieval functionalities, Stardog provides a standardized SPARQL (The W3C SPARQL Working Group, 2013) interface and it offers the possibility to integrate Ontologies in form of OWL (W3C OWL Working Group, 2012; W3C OWL Working Group, 2009) and utilize the integrated Pellet reasoner (Sirin et al., 2007). Using ontologies in the graph DB allows us to infer a wider array of information about a concept by inheriting properties from the general class definition. As an example, if we assert that *Tim Cook isA CEO*, and if our ontology encodes the fact that *CEO isA Human*, we can infer that *Tim Cook* must be a human with all the attributes and relations that follow from that, without this being explicitly mentioned in any text. At the same time, an assertion like *Apple isA Institution* cannot be followed by an obviously wrong assertion like *Apple isA CEO (of Google)*, if the concept *Institution* is not subsumed under the concept *Human* in the ontology.

4. Discussion

Corporate reporting with the Securities and Exchange Commission (SEC) represents a significant, recurring domain-specific corpus with tremendous academic and practical value. To date, most research on corporate filings has employed limited parsing techniques or only sentiment or dictionary-based approaches. Our work applies the latest advances in linguistic approaches to one of the largest, publicly available corpus of business documents. In doing so, we make significant contributions to advancing the techniques of deep natural language processing. In particular, linguists currently seek to determine how to make their tools more specifiable to particular domains as well as to account for rapid-changes in terminology use. Further, we bridge linguistic processing techniques, including mapping of semantic properties, with network analysis of individuals and entities.

In addition, we provide a valuable platform for future work in business domains. Accounting, finance and management scholars all rely heavily on SEC filings, including 10-Ks, 10Qs, and 20-Fs, and these scholars are increasingly seeking to employ various forms of computer aided text analysis. These include for instance the readability of the disclosures (Loughran and McDonald, 2014), dictionary-based approaches (Andriy Bodnaruk and McDonald, 2015) including sentiment analysis and Naive Bayesian approaches (Loughran and McDonald, 2016). Scholars in the business domain are seeking to move beyond single-word approaches into phrase (n-gram) approaches that also draw on a greater wealth of linguistic components (Pandey and Pandey, 2017). Our pipeline extends even further, showing how deep semantic processing allows for the identification of previously unobservable relationships, by providing a mechanisms for surfacing deep interrelationships between concepts, entities and related parties. Our pipeline also allows for more traditional analysis of the word content in SEC disclosures.

Since we are lacking a gold standard corpus or data set for most of our tasks, we are not yet able to measure the effectiveness of extracting concepts and relations in a formal and quantitative way. The only evaluation criterion that we can apply is whether by human judgment the results are true and useful. The same can be applied to topic modeling results from the risk description sections. We hope to be able to provide much better evaluation criteria in the near future, given that we are able to generate corpora and initial data sets using our architecture.

5. Acknowledgments

We are grateful for the programming expertise of Lwin Moe and other valuable contributions from Katherine Spoon, Jonathan Anderson, Prateek Srivastava, Anthony Meyer, Ji Li, Celine Marquet, Atreyee Mukherjee and students in Dr. Cavar's Fall 2017 AI and Machine Learning seminar at Indiana University. We received useful comments and suggestions related to this research from colleagues at Indiana University, particularly participants in our workshop at the IU Network Institute. Cavar and Josefy are Fellows of the National Center for the Middle Market at the Fisher College of Business, The Ohio State University, and acknowledge the Center's support for this research, which enabled us to gather and validate data on middle market firms for inclusion in the larger dataset. This research is also funded by the Office of the Provost of Research at Indiana University in form of a collaborative interdisciplinary research grant.

6. Bibliographical References

Andriy Bodnaruk, T. L. and McDonald, B. (2015). Using 10-k text to gauge financial constraints. *Journal of Financial and Quantitative Analysis*, 50(4):623–646.

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media.

Blei, D. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, April.

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., VanderPlas, J., Joly, A., Holt, B., and Varoquaux, G. (2013). API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122.

Cavar, D., Moe, L., Hu, H., and Steimel, K. (2016). Preliminary results from the free linguistic environment project. In Doug Arnold, et al., editors, *Proceedings of the Joint 2016 Conference on Head-driven Phrase Structure Grammar and Lexical Functional Grammar*, pages 161–181. CSLI Publications.

Chen, D. and Manning, C. D. (2014). A fast and accurate dependency parser using neural networks. In *Proceedings of EMNLP 2014*.

Clark, K. and Manning, C. D. (2015). Entity-centric coreference resolution with model stacking. In *Association for Computational Linguistics (ACL)*.

Clark, K. and Manning, C. D. (2016). Deep reinforcement learning for mention-ranking coreference models. In *Empirical Methods on Natural Language Processing*.

DeRose, S. and Clark, J. (1999). XML path language (XPath) version 1.0. W3C recommendation, W3C, November. http://www.w3.org/TR/1999/REC-xpath-19991116/.

Engel, P., Hamscher, W., Shuetrim, G., vun Kannon, D., and Wallis, H. (2013). Extensible business reporting language (xbrl) version 2.1. XBLR recommendation, XBRL International, February. http://www.xbrl.org/Specification/XBRL-2.1/REC-2003-12-31/XBRL-2.1-REC-2003-12-31+corrected-errata-2013-02-20.html.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Fischer, H. and Mueller, D. (2011a). Enabling comparability and data mining with the Arelle open source unified model, September. Paper presented at First Conference on Financial Reporting in the 21st Century: Standards, Technology, and Tools.

Fischer, H. and Mueller, D. (2011b). Open source & XBRL: the Arelle project, April. Paper presented at 5th University of Kansas Conference on XBRL.

Fischer, H. (2013). Evolution and future trends for XBRL development, April. Paper presented at 6th University of Kansas Conference on XBRL.

Honnibal, M. and Johnson, M. (2015). An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal, September. Association for Computational Linguistics.

Hulden, M. (2009). Foma: A finite-state compiler and library. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 29–32. Association for Computational Linguistics.

Jonathan L. Rogers, D. J. S. and Zechman, S. L. (2017). Run edgar run: Sec dissemination in a high-frequency world. *Journal of Accounting Research*, 55(2):459–505.

Josefy, M., Harrison, J., Sirmon, D., and Carnes, C. (2017). Living and dying: Synthesizing the literature on firm survival and failure across stages of development. *Academy of Management Annals*, 11(2):770–799.

Loughran, T. and McDonald, B. (2014). Measuring readability in financial disclosures. *Journal of Finance*, 69(4):1643–1671.

Loughran, T. and McDonald, B. (2016). Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*, 54(4):1187–1230.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

McCallum, A. K. (2002). MALLET: A machine learning for language toolkit.

Microsystems, S. (1988). RPC: Remote procedure call protocol specification. Request For Comment (RFC) 1057.

Miller, G. A. (1995). WordNet: A lexical database for english. *Communications of the ACM*, 38(11):39–41.

Newman, M., Barabási, A.-L., and Watts, D. J. (2006). *The Structure and Dynamics of Networks*. Princeton Studies in Complexity. Princeton University Press.

Pandey, S. and Pandey, S. K. (2017). Applying natural language processing capabilities in computerized textual analysis to measure organizational culture. *Organizational Research Methods*, In-press:1–33.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Sirin, E., Parsia, B., Grau, B. C., Kalyanpur, A., and Katz, Y. (2007). Pellet: A practical owl-dl reasoner. *Web Semantics: Science, Services and Agents on the World Wide Web*, 5(2):51 – 53. Software Engineering and the Semantic Web.

The W3C SPARQL Working Group. (2013). SPARQL 1.1 overview. W3C recommendation, W3C, March. http://www.w3.org/TR/2013/REC-sparql11-overview-20130321/.

W3C OWL Working Group. (2009). OWL 2 web ontology language document overview. W3C recommendation, W3C, October. http://www.w3.org/TR/2009/REC-owl2-overview-20091027/.

W3C OWL Working Group. (2012). OWL 2 web ontology language document overview (second edition). W3C recommendation, W3C, December. http://www.w3.org/TR/2012/REC-owl2-overview-20121211/.

NTUSD-Fin: A Market Sentiment Dictionary for Financial Social Media Data Applications

Chung-Chi Chen, Hen-Hsen Huang, Hsin-Hsi Chen
Department of Computer Science and Information Engineering
National Taiwan University, Taipei, Taiwan
{cjchen, hhhuang}@nlg.csie.ntu.edu.tw; hhchen@ntu.edu.tw

Abstract

Extracting information in textual data for further applications is one of the popular topics in financial domain in last decade. Although there exist some dictionaries for news and financial reports, few dictionaries focus on financial social media data. This paper constructs a market sentiment dictionary based on more than 330K labeled posts crawled from financial social media. There are 8,331 words, 112 hashtags and 115 emojis in our dictionary. The statistic results shows the difference between the sentiment and the market sentiment of the investors. Furthermore, the comparison of (1) general sentiment analysis and market sentiment analysis, and (2) the market sentiment of social media data and formal reports are discussed with the constructed dictionary. We find that some neutral words in general sentiment dictionary should be considered as the bullish/bearish words. The experimental results of our dictionary and that of the dictionary for financial formal documents show the usefulness of our dictionary in financial social media application.

Keywords: Financial dictionary, social media, sentiment analysis

1. Introduction

Textual data has been regarded as an important source when analyzing economic and financial phenomena. There are three major kinds of text resources, including official documents, news, and social media. They stand for different viewpoints of the same event. Official documents, published by government or company, provide the insiders' opinions. News are expected to propose objective opinons of the incident. Social media data, which are the most popular text data recently, contain plenty of crowd views. The informal vocabulary and syntax make social media data quite different from the other documents and make the analysis tasks more challenging.

Sentiment analysis is one of the hot topics in financial domain in last decade. Many empirical results show that textual sentiment is highly correlated to different aspects of financial phenomena. Bollen et al. (2011) show the correlation between tweet mood and Dow Jones Index. Sul et al. (2014) find that the emotion of tweets for certain company significantly affects its stock price. Furthermore, El-Haj et al. (2016) extend the sentiment analysis for sentence in PEAs into internal or external attribute, and compare the performance of human and machine.

Sentiment dictionary plays a crucial role in both dictionary-based and machine learning approaches. There are dictionaries for official documents (Loughran and McDonald, 2011) and news (Huang et al., 2013), but few dictionaries are available for social media data. Chen et al. (2014) present a dictionary for social media data applications. However, it is constructed based on the word list of Loughran and McDonald (2011), which is collected for formal documents like 10-K (annual report of company). Furthermore, the posts in the Seeking Alpha platform used by Chen et al. (2014) are dissimilar to nowaday Twitter-like social media posts. Li and Shah (2017) construct a domain specific sentiment lexicon with StockTwits data, and propose a state-of-the-art model for sentiment analysis. In this paper, we will construct a market sentiment dictionary based on over 330K labeled posts craweled from financial social media. Besides words, hashtags and emojis are also included.

In some topics, writers' sentiments are strongly related to their opinions. For example, in the review of a hotel, positive/negative sentiment of a customer is associated with good /bad opinion of the hotel. In constrast, the market sentiment (bullish/bearish) of an investor may not be derived from the positive/negative sentiments of the investor directly. More details will be discussed in Sections 4 and 5.

The sturcture of this paper is organized as follows. Details of the dataset are specified in Section 2. The methods to construct a sentiment dictionary are introduced in Section 3. Overview of the dictionary is shown in Section 4. We discuss some findings in Section 5. Finally, Section 6 concludes the remarks.

2. Fianacial Social Media Data

2.1 Data Source

StockTwits is a Twitter-like social media for investors to share their information and opinions of the market or a certain company. Figure 1 shows the graphical user interface (GUI) provided by StockTwits. The same as Twitter, it limits the length of each post to 140 characters. Under this limitation, users have to focus on a few main points they want to share in the posts. Users usually use cashtag ($ before ticker) to mark the instruments they mention. For instance, $MSFT stands for the security of Microsoft Corporation. In particular, the bullish and bearish bottoms allow users to label their market sentiment in the post.

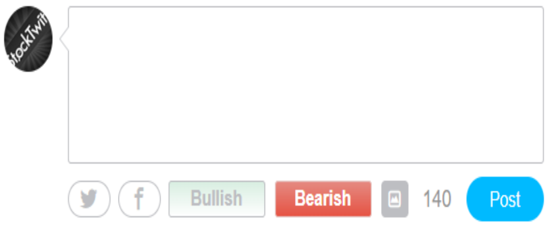


Figure 1: GUI of Stocktwits

2.2 Dataset

From StockTwits, we crawled 334,798 labeled posts from 13,059 users. (The detail about StockTwits API please refer to the related documents [1].) In total there are 75,376 unique words, 3, 041 unique hashtags, and 451 unique emojis in the collection. The distribution of these elements for both sentiments is shown in Table 1.

Since the bull market is much longer and more profitable than the bear market in history, it is reasonable that people tend to find the bullish targets. The other reason for the unbalance distribution may be that short stock is costly than long. Therefore, 95.35% of users had published the bullish posts, while only 44.67% of users had published the bearish posts.

	Bullish	Bearish
Post	289,416	45,382
User	12,452	5,834
Word	69,114	25,956
Hashtag	2,507	715
Emoji	427	174

Table 1 : Distribution of dataset.

2.3 Quality of the Dataset

The collected dataset in this paper is the large financial social media dataset labeled by the original writers. Compared with the datasets that are labeled by additional annotators, ours is advantageous in the consistency between the text meaning and the label since the writers would not misunderstand the meaning in the posts written by themselves. Therefore, it is reasonable to assume this dataset is a high quality dataset.

3. Methods

To mine the bullish/bearish sentiment tokens, we applied four methods including chi-squared test, collection frequency, pointwise mutual information, and a convolutional neural network classifier. Before that, we perform the data preprocessing as follows. First, stopwords, punctuations, digits, URLs, user ids, and tickers are removed from the posts. Second, all characters are converted to lowercase. Third, we remove the posts less than 2 words. For example, users may just post one cashtag and give a sentiment label. Finally, the tokens appearing less than *n* times are not taken into consideration, where *n* is set to 100 for words and 10 for hashtags and emojis. We do not perform word stemmings, beacuse we would like to maintain the original results and keep the largest flexibility for the uses of the proposed sentiment dictionary.

3.1 Chi-Squared Test

Chi-squared test is used to determine if there exist the difference between expected and observed frequency. It is adopted to decide whether the token should be remained in our dictionary with the confidence level set to 95%.

3.2 Collection Frequency-Inverse Document Frequency

Collection frequency(CF) is calculated as

$$cf_s(t, D_s) = log(1 + f_{t,D_s}) \quad (1)$$

,where t is one of the tokens in the list of words, hashtags or emojis, and s stands for a sentiment (i.e., bearish or bullish). D_s is a set of posts labeled as s, and f_{t,D_s} is the frequency of the token t appearing in D_s. Inverse document frequency (IDF) is the most common weighting scheme used to extract the keywords of documents.

$$idf_s(t, D_s) = log\frac{N_s}{|\{d \in D_s : t \in d\}|} \quad (2)$$

where N_s is the number of posts in D_s. Collection Frequency- Inverse Document Frequency (CFIDF) can be computed as follows.

$$cfidf(t, s) = tf_s(t, D_s) \times idf_s(t, D_s) \quad (3)$$

We can obtain the degree of importance of a token according to its CFIDF score in sentiment *s*.

3.3 Pointwise Mutual Information

Pointwise mutual information (PMI) is used to measure the dependence of the events. With PMI, we can observe how much the token *t* is correlated to the sentiment *s*.

$$pmi(t, s) = log\frac{p(t,s)}{p(t)p(s)} \quad (4)$$

In order to maintain the information of frequency, we use the probability of the tokens in the dataset to weight pmi as (5), where f_t is the frequency of the token t and T is the total number of tokens in the dataset.

$$wpmi(t, s) = \frac{f_t}{T} \times pmi(t, s) \quad (5)$$

3.4 Convolutional neural networks

Convolutional neural networks (CNN) is adopted to train the word embedding of each token in our dataset. The input is the text in a StockTwits post. The output of CNN model is the classification result, i.e., bullish or bearish, of a input post. The sturcture of the model is shown in Figure 2. The loss function is binary cross entropy, and we use Adam algorithm (Kingma and Ba, 2015) to optimize the parameters.

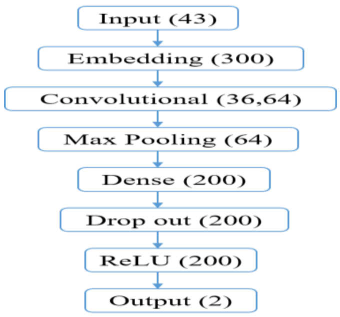


Figure 2: Structure and output size of CNN model

The word embedding scheme with neural network is widely used in natural language processing. We use the CNN model to train a classifier for market sentiment with our dataset, and use the output vector of embedding layer as the representation of each token. We calculate the cosine

[1] https://api.stocktwits.com/developers/docs

Bullish						Bearish					
Word		Hashtag		Emoji		Word		Hashtag		Emoji	
buy	14,489	stocks	202	😂	927	short	3,653	stocks	68	😂	184
today	13,191	sharkalerts	110	👍	518	going	2,094	noshamenate	40	💩	55
like	11,624	bitcoin	104	🚀	517	stock	2,075	sanofiwasright	34	🐻	35
go	10,959	cesium	102	😎	332	like	1,989	mnkdsecinvestigation	29	😎	26
get	10,829	trading	92	😉	290	sell	1,897	forex	27	😭	24
going	10,203	bullish	71	😁	282	get	1,678	trading	26	😊	21
back	9,768	stock	71	👀	242	lol	1,663	elliottwave	16	😉	21
day	9,400	brachytherapy	63	🤔	233	today	1,626	earnings	15	😁	20
stock	8,590	tradesmart	60	💰	210	back	1,595	scambags	15	🤔	20
next	8,451	btfd	59	🤑	186	buy	1,592	markets	14	😜	17

Table 2 : Top 10 highly frequent tokens for sentiments *Bullish* and *Bearish*.

Bullish						Bearish					
Word		Hashtag		Emoji		Word		Hashtag		Emoji	
bully	86.74	stocks	80.19	🚀	82.75	bye	64.25	sanofiwasright	66.01	😂	60.78
dilly	86.74	sharkalerts	77.04	👍	82.23	ah	61.87	stocks	57.22	💩	56.64
bye	85.32	bitcoin	76.77	👀	82.14	junk	61.29	noshamenate	56.25	🐻	53.38
blah	83.87	cesium	76.54	😎	81.95	million	60.89	mnkdsecinvestigation	52.77	😎	51.40
energy	83.78	trading	75.83	😂	81.71	death	60.60	forex	51.43	😭	50.47
vs	83.77	pebbleproject	74.24	😉	81.36	bubble	60.53	trading	51.13	😊	49.33
sma	83.36	stock	73.85	😁	81.29	cents	60.39	elliottwave	46.82	😉	49.33
billion	83.32	bullish	73.85	🤔	80.76	dip	60.33	scambags	46.19	🤔	48.90
candle	83.18	brachytherapy	72.85	💰	80.72	debt	60.33	earnings	46.19	😁	48.90
phase	82.99	tradesmart	72.43	💥	80.65	weekly	60.33	pennymikey	45.50	😜	47.40

Table 3 : Top 10 tokens ranking with CFIDF for sentiments *Bullish* and *Bearish*

similarity of each token with the "bullish" and "bearish", and subtact the cosine similarity with "bearish" from the cosine similarity with "bullish" to measure the tendency of each token. The token with positive score is considered to be the token with bullish tendency, and the token with negative score is considered to be the token with bearish tendency.

4. Analysis of Dataset

4.1 Analysis in Single Sentiment

The top 10 highly frequent tokens of both sentiments are shown in Table 2, respectively. First, we can find that "buy" is the most frequent word used in bullish posts. It is reasonable because the bullish label means a writer expects the price of the metioned instrument will rise, and may write down her/his action or ask the others to buy the mentioned instrument. "buy" is also in the top-10 tokens in bearlish posts. The reason is that it is unlimited for the rising price, but the minimun of the falling price is 0. In some bearish posts like (P1) "buy" is used to mention the buy back price of the instrument. Compared to bullish posts, the writers of bearish posts tend to use "short" but not "sell" to narrate the action they take.

(P1) *$SKX if you have profits, sell now and buy back below 30. You will thank me later*

Second, the uses of hashtags in financial social media can be formulated in the following four ways: (1) To tag the instruments they focus on (stock, bitcoin, forex), (2) To use an unique tag to store their posts (e.g., mnkdsecinvestigation), (3) To label their sentiments (e.g., bullish, sharkalerts, btfd), and (4) To tag the method they used (e.g., elliottwave). Compared to common social media, hashtags are not frequently used in financial social media. Only 1.37% of bullish posts and 1.75% of bearish posts contain at least one hashtag.

Third, emoji is more popular than hashtag in financial social media. Total 2.81% of bullish posts and 1.86% of bearish posts contain at least one emoji. In other domains, e.g., hotel review, the smile emoji stands for positive comment. In contrast, the Face With Tears of Joy (😂) emoji gets the first place in both bullish and bearish posts. It is a special phenomenon in sentiment analysis. Since the sentiment of an investor may depend on her/his return from trading, the investor who longs the intrument will feel happy if the price rise, but the investor who shorts the insturment will feel sad in this situation. Therefore, the market sentiment of investors should be discriminated from the positive/negative sentiment of investors in financial social media data. We will show some evidence for this issue in Section 5. Some emojis may not imply sentimentss, for example, Thinking Face (🤔) appears in both top-10 lists.

Furthermore, we use CFIDF to calculate the weights of tokens. The results are shown in Table 3. In bullish post, some positive words such as "bully" and "dilly", get the first place, and some negative words like "junk", "death" and "dip" are shown in the top-10 list of bearish.

4.2 Analysis between Two Sentiments

To analyze the correlation of tokens and both sentiments, we adopt PMI to sort out the critical tokens. The top-10 results are shown in Table 4. In bullish posts, "dominant", "bully", "blast" and "undervalued" get the front place. If the instrument is described with "undervalued", it is similar to the writer expects that this instrument is bullish. The same case is also shown in bearish posts. The word "overvalued" gets the third place in the list. In bearish posts, "junk", "garbage", "trash" and "turd" appear in the top-10 list, and these words are considered as negative words, when they are used to described things. Furthermore, "puts", standing for one kind of options that buyers have the right to sell the underlying asset at certain price, is in the bearish list. The emoji results could also show some clues for the writer's sentiments. The Ox (🐂) and Airplane Departure (🛫) emojis get the front place in bullish posts, and the Down-Pointing Triangle (🔻), Thumbs Down Sign (👎), Pile of Poo (💩), and Bear Face (🐻) emojis are in the top-10 list of bearish posts.

Bullish						Bearish					
Word		Hashtag		Emoji		Word		Hashtag		Emoji	
dominant	1.22	buy	1.27	🌞	1.14	bagholders	3.46	mnkdsecinvestigation	3.53	🔻	4.42
bully	1.21	early	1.27	🎄	1.14	junk	3.45	pennymikey	3.53	👎	3.68
updates	1.21	gainers	1.27	🐂	1.14	overvalued	3.42	noshamenate	3.53	💩	3.53
runner	1.21	mattel	1.27	💎	1.14	pumpers	3.37	sanofiwasright	3.53	🐻	2.59
binance	1.21	oprah	1.27	💲	1.14	scam	3.34	scambags	3.53	🎯	2.01
blast	1.21	shortsqueeze	1.27	🔑	1.14	garbage	3.32	short	3.39	😂	1.82
floater	1.21	analyis	1.27	🛫	1.14	pig	3.25	scam	3.29	🙂	1.82
undervalued	1.21	biotech	1.27	🤘	1.14	trash	3.25	forex	3.04	😳	1.67
accumulating	1.20	blocks	1.27	😳	1.14	turd	3.19	market	2.75	😀	1.57
blackberry	1.20	boolish	1.27	🥃	1.14	puts	3.11	elliottwave	2.40	🤣	1.52

Table 4 : Top 10 tokens ranking with PMI for sentiments *Bullish* and *Bearish*

Bullish						Bearish					
Word		Hashtag		Emoji		Word		Hashtag		Emoji	
bully	1.22	sharkalerts	1.34	🚀	1.15	bagholders	3.42	short	3.01	👎	3.56
updates	1.22	club	1.34	💥	1.15	junk	3.41	scam	2.91	💩	3.40
runner	1.22	bullboard	1.33	🤞	1.13	overvalued	3.38	forex	2.66	🐻	2.47
binance	1.22	dontbeasheep	1.33	👀	1.13	pumpers	3.33	markets	2.37	🎯	1.89
blast	1.22	crypto	1.32	🚂	1.12	scam	3.30	elliottwave	2.02	😂	1.70
floater	1.22	pennystocks	1.29	💪	1.12	garbage	3.28	futures	1.48	🙂	1.69
undervalued	1.21	blockchain	1.28	💰	1.12	pig	3.21	earnings	1.41	😳	1.55
accumulating	1.21	moviepass	1.28	🏃	1.12	trash	3.21	stocks	1.16	😀	1.45
blackberry	1.21	stockmarket	1.28	👍	1.12	turd	3.15	trading	0.97	🤣	1.40
partnerships	1.21	timestamp	1.27	🌙	1.11	puts	3.07	study	0.86	😊	1.39

Table 5 : Top 10 tokens ranking with PMI appearing in both sentiments *Bullish* and *Bearish*

Bullish						Bearish					
Word		Hashtag		Emoji		Word		Hashtag		Emoji	
buy	0.01	stocks	0.04	😂	0.10	short	0.01	stocks	0.04	😂	0.20
today	0.01	sharkalerts	0.02	🚀	0.06	stock	0.01	trading	0.02	💩	0.04
like	0.01	bitcoin	0.02	👍	0.06	going	0.01	forex	0.01	🐻	0.03
go	0.01	trading	0.02	😎	0.04	like	0.01	elliottwave	0.01	🤣	0.03
get	0.01	stock	0.01	😉	0.03	sell	0.01	earnings	0.01	😊	0.02
going	0.01	btfd	0.01	😀	0.03	lol	0.01	markets	0.01	😀	0.02
back	0.01	stockmarket	0.01	👀	0.03	get	0.01	futures	0.01	😎	0.02
day	0.01	club	0.01	🤔	0.03	back	0.01	study	0.01	😜	0.02
shares	0.01	study	0.01	💰	0.03	money	0.01	scam	0.01	💥	0.02
stock	0.01	optionpros	0.01	🤑	0.02	today	0.01	short	0.01	🤔	0.02

Table 6 : Top 10 tokens ranking with WPMI appearing in both sentiments *Bullish* and *Bearish*

Bullish						Bearish					
Word		Hashtag		Emoji		Word		Hashtag		Emoji	
streamline	1.55	brent	1.25	💥	1.16	fuh	-1.55	getyourshinebox	-1.26	⚰	-1.10
dinghy	1.55	crossover	1.23	🙌	1.11	pumptards	-1.35	timberrrrr	-1.23	🔻	-1.07
bitc	1.27	qnx	1.18	🗺	1.11	foolishness	-1.28	overpriced	-1.22	💭	-0.99
awakes	1.26	buffett	1.17	🏃	1.11	bleeds	-1.28	pennymikey	-1.22	🥔	-0.94
rap	1.25	overwatch	1.15	🎞	1.11	grasshoppa	-1.26	scam	-1.18	👎	-0.93
brent	1.25	powerhour	1.15	🚀	1.11	leeches	-1.24	overbought	-1.18	😷	-0.92
crossover	1.23	paytheask	1.14	👊	1.11	downgraded	-1.24	overvalued	-1.16	🌏	-0.89
attend	1.23	letsgo	1.14	⛽	1.11	timber	-1.24	bankruptcy	-1.16	💩	-0.87
shaken	1.23	epyc	1.13	🐐	1.09	barev	-1.24	bitcoinfork	-1.15	⚫	-0.86
varta	1.22	wallstreetgames	1.13	🌪	1.09	myant	-1.23	forex	-1.15	🚽	-0.78

Table 7 : Top 10 tokens ranking with word embedding and cosine similarity for sentiments *Bullish* and *Bearish*

	NTUSD-Fin						SentiWordNet	
Word	Market sentiment	Chi squared	Bull freq.	Bull cfidf	Bear freq.	Bear cfidf	Sentiment	Word ID
buy	0.59	14711.71	14489	61.54	1592	52.32	0.00	buy#1
sell	-0.98	3581.53	5800	71.02	1897	51.60	0.00	sell#4
call	0.44	2211.63	2259	78.16	277	59.76	0.00	call#13
put	-0.49	973.82	1326	80.52	310	59.59	0.00	put#1
overvalued	-3.42	1625.99	54	71.49	172	59.75	0.25	overvalue#1
undervalued	1.21	1095.06	844	81.71	9	40.81	-0.38	undervalue#1

Table 8 : Comparision of NTUSD-Fin with SentiWordNet

To do the in-depth analysis, the tokens that appear in the posts of both sentiments are shown in Table 5. The results of hashtag and emoji of bullish posts are different from the results in Table 4. The hashtag "sharkalerts" gets the first place in the hashtag result. Moreover, none of emoji is the same as the emojis in Table 4. Rocket (🚀), Steam Locomotive (🚂) and Flexed Biceps (💪) are highly related to bullish. Moreover, we add the frequency information with WPMI, and the results are shown in Table 6. With frequency information, WPMI tends to pick out the general tokens. Most of tokens in Table 6 are the same as those scored with frequency in Table 2. Therefore, comparing with frequency and WPMI, the results scored by PMI contain most of specific tokens for market sentiment.

The top 10 tokens ranking by cosine similarity based on on word embedding for both "bullish" and "bearish" are listed in Table 7. Some tokens are different from those proposed by the other methods. The word, "downgraded", is picked out by this scoring method in bearish words, and the "overpriced", "overbought" and "overvalued" hashtags show high tendency toward bearish.

4.3 Dictionary Format

Because different information may have dissimilar usage, our dictionary provides various scoring methods including frequency, CFIDF, chi-squared value, market sentiment score and word vector for the tokens. Only the tokens appeared at least ten times and shown significantly difference between expected and observed frequency with chi-squared test are remained in our dictionary. The predetermined significance level is 0.05. The market sentiment score is calculated by substracting the bearish PMI from the bullish PMI. There are 8,331 words, 112 hashtags and 115 emojis in the constructed dictionary, NTUSD-Fin. Some examples are shown in Table 8. The distribution of these elements for both sentiments is shown in Table 9.

	Bullish	Bearish
Word	6,670	1,661
Hashtag	97	15
Emoji	103	12

Table 9 : Distribution of NTUSD-Fin.

5. Discussion

First, we discuss the gengeral sentiment and the market sentiment in financial social media. As we mentioned in Section 4, for the sentiment analysis in hotel reviews, the positive/negative sentiment of a writer is associated with a good/bad opinion of the hotel. This case is the same in movie reviews and product reviews. However, the sentiment of the investors may depend on the positions they hold. Therefore, the positive sentiment of the investor does not imply the bullish market sentiment for the mentioned target of this investor. It is worth distinguishing the market sentiments of the investors from the sentiments of the investors. To in-depth analysis, we compare the market sentiment scores with the sentiment scores in SentiWordNet 3.0 (Baccianella et al., 2010), a dictionary annotated with sentiments of all synsets in WordNet (Fellbaum, 2005) in Table 8. The results show the differences between general sentiments and market sentiments. For instance, "buy" and "sell" are neutral words in SentiWordNet, but they get the positive and the negative market s scores in our dictionary, respectively. (P2) shows a bullish instance containing "buy", and (P3) shows a bearish instance containing "sell". With dictionary-based approach with SentiWordNet, the sentiment scores for both posts are zero. In contrast, the market sentiments provided by our dictionary are 0.64 and -0.93, suggesting correct information for market sentiments of investors. The words "call" and "put" are examples to show a similar phenomenon.

(P2) *$CHGG buy while you can...*
(P3) *$CADC sell*

In addition, the scores of "overvalued" and "undervalued" in our dictionary are opposite to the scores in SentiWordNet. This result shows the different between market sentiment and common sentiment. Because the price of overvalued instruments is expected to fall down, and the price of the undervalued instrument is expected to rise up, it is reasonable to be considered as bearish word and bullish word in financial data. However, "overvalue" and "undervalue" are the same as overestimate and underestimate in SentiWordNet, and get the positive and negative sentiment scores. The evidence illustrates the difference between general sentiment and market sentiment. (P4) and (P5) are the posts that will be misled by using SentiWordNet.

(P4) *$WTW very overvalued like $HLF*
(P5) *$MTBC so undervalued at these prices*

Second, we compare the words in our dictionary with those in the dictionary for formal documents in finance (Loughran and McDonald, 2011). There are 354 positive words and 2,355 negative words in their dictionary, and only 152 positive words and 329 negative words appear in our dictionary. This circumstance shows that the words used by investors in social media are different from those used in formal documents such as 10-K annual reports. Besides, "easy" is a positive word in the their dictionary, but gets a negative market sentiment score in our dictionary. It implies that not only the words are different between social media data and annual reports, but also the tendency of market sentiment of the same words may also different.

	Micro	Macro
Loughran and McDonald	21.67	23.02
NTUSD-Fin	61.23	40.22

Table 10 : Dictionary-based experimental results. (%).

Furthermore, we use dictionary-based model to test the performance of the dictionary of Loughran and McDonald (2011) and our dictionary. We use the number of positive words minus the number of negative words in the dictionary to score the sentiment of each tweet. The tweets get positive (negative) score will be considered as bullish (bearish) instances, and the tweets with zero score will be considered as neutral instances. SemEval-2017 Task 5 dataset (Cortis et al., 2017), which was collected from Twitter and StockTwits, is adopted for this experiment. In order to confirm all instances are discussing financial instruments, each instance contains at least one cashtag. There are total 2,030 instances in this dataset, including 1,318 bullish instances, 676 negative instances, and 36 neutral instances. Table 10 shows the micro- and macro-averaged F-score of both dictionaries. Our dictionary outperforms Loughran and McDonald's dictionary, which is constructed for formal documents. The experimental results shows the usefulness of the media-oriented dictionary.

In summary, the usage of the NTUSD-Fin dictionary is different from that of general sentiment dictionaries from several aspects. The purpose of this dictionary is to capture the market sentiment of the investors for the mentioned instruments in the social media platform, but not to predict the sentiments of the investors.

6. Conclusion and Future Work

In this paper, we distinguish the market sentiment of investors from the sentiments of investors. Not only the emoji shows the evidence of this phenomenon, the comparison with general sentiment dictionary shows too. The constructed market sentiment dictionary is based on a large-scale labeled data from financial social media. Words, hashtags and emojis are included in the dictionary. The POS tagging and the meaning of the words will be added in the future version. Our dictionary[2] is publicly available for research purpose.

7. Acknowledgements

This research was partially supported by Ministry of Science and Technology, Taiwan, under grants MOST 107-2634-F-002-011-, MOST 106-3114-E-009-008- and MOST 105-2221-E-002-154-MY3.

8. References

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*, pages 2200–2204.
Bollen, J., Mao, H., & Zeng, X.J. (2011). Twitter Mood Predicts the Stock Market, *Journal of Computational Science*, 2(1): 1–8.
Chen, H., De, P., Hu, Y. J., & Hwang, B. H. (2014) Wisdom of Crowds: The Value of Stock Opinions Transmitted Through Social Media. *Review of Financial Studies* 27: 1367–1403.
Cortis, K., Freitas, A., Daudert, T., Huerlimann, M., Zarrouk, M., Handschuh, S., & Davis, B. (2017). Semeval-2017 task 5: Fine-Grained Sentiment Analysis on Financial Microblogs and News. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 519-535.
El-Haj, M., Rayson, P. E., Young, S. E., Walker, M., Moore, A., Athanasakou, V., & Schleicher, T. (2016). Learning tone and attribution for financial text mining. In *Proceedings of the tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1820-1825.
Fellbaum. C. (2005) Wordnet and wordnets. In K. Brown, editor, *Encyclopedia of Language and Linguistics*, pages

[2] http://nlg.csie.ntu.edu.tw/nlpresource/NTUSD-Fin/

665–670, Oxford, Elsevier.

Huang, A., Zang, A., & Zheng, R. (2013). Large sample evidence on the informativeness of text in analyst reports. *The Accounting Review*, 89(6): 2151-2180.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv, arXiv:1412.6980.

Li, Q., & Shah, S. (2017) Learning Stock Market Sentiment Lexicon and Sentiment-Oriented Word Vector from StockTwits. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 301–310.

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, 2011, 66.1: 35-65.

Sul, H. K., Dennis, A. R., & Yuan, L. I. (2014). Trading on Twitter: The Financial Information Content of Emotion in Social Media. In *Proceedings of 47th Hawaii International Conference on System Sciences (HICSS)*, pages 806–815.

QSRL: A Semantic Role-Labeling Schema for Quantitative Facts

Matthew Lamm[1,3], Arun Chaganty[2,3],
Dan Jurafsky[1,2,3], Christopher D. Manning[1,2,3], Percy Liang[2,3]
[1]Department of Linguistics, Stanford University, Stanford, CA, USA
[2]Stanford Computer Science, Stanford University, Stanford, CA, USA
[3]Stanford NLP Group
{mlamm, jurafsky}@stanford.edu
{chaganty, manning, pliang}@cs.stanford.edu

Abstract

Financial text is replete with quantitative information about company, industry, and economy-level performance. Until now however, work on financial narrative processing has overlooked this information in favor of softer forms of meaning like textual sentiment. In this paper, we examine such language from two sources—newswire and publicly available quarterly reports—to define an annotation schema for quantitative facts in text to be used in future information extraction (IE) work. The Quantitative Semantic Role Labels (QSRL) representation takes a situationist perspective on quantitative facts, describing quantities not only in terms of hard numerical values, but also the context in which they take on those values. Unlike other semantic role-labeling frameworks however, it is specifically designed with quantitative language in mind, and hence is a much simpler representation. We conclude with a description of some of the challenges we face in quantitative information extraction, as highlighted by the data we consider throughout the paper.

Keywords: semantic role labeling, information extraction, quantities, numbers, financial text

1. Intro

Research on financial narrative processing distinguishes the "hard," quantitative information reflected in financial tables, such as balance sheets and income statements, from the "soft" information reflected in the financial language of earnings calls, such as textual sentiment (Engelberg, 2008; Demers et al., 2008; Lee, 2014). So posed however, the soft/hard dichotomy is misleading. Financial text contains quantitative information that contributes to our understanding of companies' fundamentals that is not reflected in the standard tables of the financial reporting repertoire. In this paper, we present a particular methodology for representing such facts such that they can be extracted using tools for shallow semantic parsing.

A quick survey of the most recent quarterly reports from Boeing reveals a variety of examples. Included in these are quantities that need not be reflected in standard financial reporting tables, and elaborations thereupon (the first boldfaced figure and the second, respectively):

(1) Backlog at Defense, Space & Security was **$46 billion**, of which **35 percent** represents orders from international customers.(Boeing Company, The, 2017a)

forward-looking quantitative assessments that do not fall under the auspices of guidance:

(2) Our 20-year commercial market outlook forecasts demand for approximately **41,000 new airplanes** over the next 20 years.(Boeing Company, The, 2017b)

as well as quantitative facts that provide industry-wide perspectives and hence are not about a particular company per se, as in:

(3) In the Commercial Airplanes market, airlines continue to report solid profits. And passenger traffic growth continues to outpace GDP, with traffic growth of **8%** through August.(Boeing Company, The, 2017b)

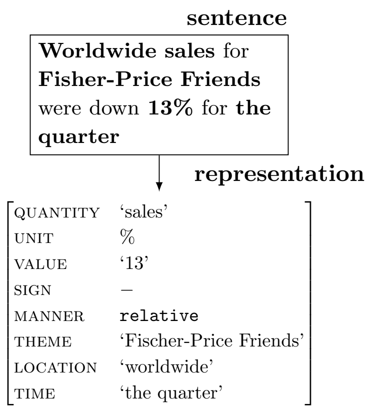


Figure 1: An analysis of a quantitative financial sentence using QSRL.

This list is non-exhaustive.

The abundance of such quantitative information in financial text begs the question of how to extract it automatically, which in turn forces us to define what "it" even is. We thus examine data from two sources—quarterly reports and newswire—to define a general semantic representation of quantitative language. We call the representation Quantitative Semantic Role Labels (QSRL). See, for example, Figure 1.

QSRL expresses quantitative facts in terms of a fixed set of information slots, or roles, which can be divided into two groups. The first of these comprises the explicitly quantitative roles, such as QUANTITY (e.g. 'profit'), UNIT (e.g.

'$'), VALUE (e.g. '2 million'), MANNER (`absolute` or `relative`), and SIGN (±). The rest of the roles summarize the broader context in which a quantity manifests, such as the TIME at which a particular value is taken on by a quantity, or a quantity's THEME—an entity in the world that the quantity parametrizes. We define all of QSRL's roles, with examples, in Section 3; In Section 4, we show how to apply it to a handful of examples.

A fundamental concern in defining any semantic representation is the expressivity-sparsity tradeoff: On the one hand, a semantic representation should faithfully replicate the meaning of language as we understand it. On the other hand, representations that are too complex induce sparsity over even very large datasets, and thus cannot be recognized by statistical algorithms for information extraction. QSRL is no exception to this rule.

In order to justify the introduction of any new role in QSRL, we thus look for instances in the data where the *absence* of that role would lead to a vague, and hence uninformative analysis. The following Wall Street Journal sentence, for example, makes a strong case for a TIME role, among others:

(4) Genentech Inc. said **third-quarter** profit more than doubled to $ 11.4 million from a depressed **1988 third-quarter** performance of $ 5.3 million.

For, without a way of representing the time at which a quantity took on a particular value, our meaning representation would be unable to distinguish between the fact that Genentech had a profit of $ 11.4 million in the 1989 third quarter and the fact that it had a profit of $ 5.3 million in the 1988 third quarter. To identify such sentences, we simply analyze instances in the data that have multiple numerical mentions, as indicated by their part of speech tags.

The need to contextualize quantities is amplified when we consider downstream inferences we expect to be able to perform with a numerical information extraction system at hand. Suppose, hypothetically, that we extracted from language a fact about the number of shares outstanding for a some company, and another fact about that company's earnings. In order to compute earnings-per-share for that company, we would first have to confirm that those facts refer to the company at the same point in time.

In summary, our contributions are as follows: (a) a survey of quantitative financial language from multiple data sources, (b) QSRL: a general-purpose representation of quantitative language that emphasizes context, and (c) a survey of challenges for future efforts in quantitative information extraction using QSRL.

2. Related Work

Quantitative information extraction is an underrepresented area in NLP. Recently however, Madaan et al. (2016) showed that distant supervision techniques used in relation extraction can be applied to extract relation triples of the form `inflation_rate`(*India*, *11%*) from web text. QSRL can be thought of as an n-ary representation of quantitative facts that better capture the situational context (Devlin, 2006; Forbus, 1984) in which a quantity manifests.

In essence, QSRL augments the Quantity-Value Representation (QVR) of numerical facts in (Roy et al., 2015; Roy, 2017) using variants on standard roles from semantic role labeling (SRL) (Levin, 1993). The four components of QVR find correlates in QSRL roles like QUANTITY, VALUE, UNIT, and MANNER.

In order to generate representations using the considerably more general semantic sing frameworks like VerbNet (Kipper Schuler, 2005), PropBank (Palmer et al., 2005), and FrameNet (Baker et al., 1998), one must identify predicates in text, retrieve from a hard-coded lexicon a set of semantic roles or frame elements specific to that predicate, and then identify the elements in the sentence which correspond to those roles. A great deal of this effort is ancillary to the more specific goal of representing quantitative language in particular.

On the other hand, being designed to extract hard quantitative facts in particular, a QSRL annotator has recourse to a much simpler regime. All it must do is identify a numerical mention in text, e.g. '$ 2 million', which can be done in a rule-based manner, and assign roles to contextualizing phrases in the surrounding language. Recent work suggests this focused procedure can be performed in an end-to-end manner (He et al., 2017).

The general idea of representing the context surrounding numbers in financial text is not our own. The Extensible Business Reporting Language (XBRL) (XBRL International Inc., 2013), for example, is a widely-adopted data reporting standard that similarly represents quantitative facts in terms of a set of roles, called elements. QSRL is importantly distinct in two ways. Firstly, its role set is more expressive than the set of XBRL elements. For example, whereas XBRL has a single `entity` attribute describing the organization or business entity described by a fact, QSRL has a more nuanced entity roles THEME, AGENT, and SOURCE that express distinct semantic relations in a fact. Secondly, whereas XBRL annotations wrap numerical mentions in a document with normalized metadata, annotated QSRL roles are aligned to text, and hence can provide strong supervision to information extraction systems.

3. QSRL

In this section, we define the roles of QSRL by examining data from newswire and earnings call transcripts. Where mentioned, roles are written in small-caps, e.g. MANNER and THEME. While QUANTITY is a specific role in QSRL, we only write it in small caps when we are explicitly referring to it as a role, instead of a general concept. In the rest of the paper, any sentence that does not have an associated citation is from the Wall Street Journal (Mitchell P. Marcus, 1999).

3.1. QUANTITIES and VALUES

Quantities are the crux of QSRL, in relation to which all of its roles are defined. Thus, a robust semantics of quantities is of the utmost importance. At the same time, following recent work in SRL (He et al., 2015) we maintain that identifying quantities in a text should be as easy as asking the question "what does this number measure?" Asking

this question of data from newswire and earnings call transcripts, we identify a variety of quantitative modes.
Note that following the aforementioned Quantity-Value Representation (Roy et al., 2015), we distinguish QUANTITIES, e.g. 'profit', from the VALUES they take on, e.g. '2 million' and the UNITS in which they are measured e.g. '$'. Sometimes, quantities can be seen as measuring the extent of some explicit predicate. Consider the following

(5) In 2013, Mattel **returned** almost **$1 billion** to shareholders.(Mattel Inc., 2014)

Here, the predicate in question is 'return', which comes associated with a dollar-valued argument measuring the extent to which Mattel returned money to shareholders.
Relatedly, the set of quantities includes conventionalized measures like *profit* and *loss* exemplified in the following sentence

(6) Priam had a **loss** of **$ 25.4 million** for the fiscal year ended July 7, compared with year-earlier **profit** of **$ 543,000**, or two cents a share.

Profit and loss are quantities which parametrize some financial entity, and serve as conventional measures of the entity's performance.
We distinguish these from quantities measuring the extent to which some predicate obtains over a *set* of entities. This is commonly the case where the grammatical subject is a cardinal- or percent-quantified subject, as in

(7) The poll showed that company size had a bearing on a manager's view of the problem, with **65%** of those in companies of more than 15,000 employees **saying stress-related problems were "fairly pervasive"** and **55%** of those in companies with fewer than 4,000 employees **agreeing**.

In the above, we annotate the bolded predicates as QUANTITIES. This analysis amounts to saying that what is being measured is the extent to which stress-related problems obtain in specific settings.
Quantities also come in an existential flavor, measuring the existence of some entity or class of entities in the world. Consider the following

(8) The two-part issue consists of $ 200 million of **senior subordinated reset notes** maturing in 1997 and $ 150 million of **subordinated floating rate notes** also maturing in 1997.

Here, amounts of two financial entities (bold-faced) are quantified in terms of their monetary value. Labeling these entities QUANTITIES and their associated numerical mentions VALUES amounts to saying that what is being measured is the extent of the existence of those quantities.

3.2. CO-QUANTITIES

In some cases, answers to our diagnostic question "what does this number measure?" are less clear. Consider

(9) The company was to **repay** $ 58 million in **debt** on Dec. 31 and $ 15 million on March 31.

In one sense, the numerical values associate with the predicative quantity 'repay', and measure the amount the company in question was to repay on specific days of the year; in another sense, they measure amounts of 'debt' held by the company. We take signal from context, maintaining that at a higher level the sentence is about repayment, and thus 'repay' is the quantity. We call 'debt' a CO-QUANTITY describing the nature of what is being repaid.

3.3. MANNERS of measure

Following previous work (Roy et al., 2015), QSRL distinguishes between **absolute** and **relative** values taken on by a given quantity. This distinction is exemplified in the following sentence:

(10) Echo Bay Mines **rose** $\frac{5}{8}$ to $15\frac{7}{8}$.

We say that the two numerical values above refer to the same quantity—implicitly, the price on shares of Echo Bay Mines—but differ in the MANNER in which they relate to that quantity. The first of these is a **relative** measure denoting the degree of change undergone by the quantity; the second is an **absolute** measure denoting a value taken on by the quantity within a particular interval of time.
Since, in our framework, there are only two general classes of measure, it suffices to annotate for one and leave the other implied. We nominate for annotation predicates of change, e.g. 'rose', which signify the relative manners of measure.
This is based on the observation that predicates of change not only encode a manner, but also sign. Consider

(11) AMR, which owns American Airlines, **rose** $3\frac{3}{8}$ to $72\frac{1}{4}$; USAir Group **fell** $1\frac{1}{2}$ to $38\frac{5}{8}$, and Delta Air Lines **rose** $\frac{1}{2}$ to $66\frac{1}{2}$ after posting higher earnings for the September quarter.

In addition to the common verbs of change, English also has recourse to case modifiers like 'up' and 'down', as in

(12) Other winners include real estate issues Mitsubishi Estate, which closed at 2,500, **up** 130

Despite the lexical diversity in predicates of change, these ultimately constitute a closed class. This is a useful fact, because sign ($\pm$) can thus be extracted from change predicates in a deterministic way, for example by using a hard-coded dictionary.

3.4. SIGN modifiers

The sign of a number is sometimes explicitly indicated in text, for example as an adjectival modifier:

(13) Together, the six government-controlled or essentially insolvent Arizona thrifts have tangible capital of a **negative** $ 1.5 billion

More frequently however, sign is built-in to the semantics of another word in a sentence. For example, the aforementioned predicates of change not only encode manner, but also sign. Consider

(14) AMR, which owns American Airlines, **rose** $3\frac{3}{8}$ to $72\frac{1}{4}$; USAir Group **fell** $1\frac{1}{2}$ to $38\frac{5}{8}$, and Delta Air Lines **rose** $\frac{1}{2}$ to $66\frac{1}{2}$ after posting higher earnings for the September quarter.

Here, the bolded verbs 'fell' and 'rose' are negatively and positively signed predicates of change, respectively.

3.5. PRECISION modifiers

There is a diverse set of ways to indicate that a number is an approximate figure, which we call PRECISION modifiers. For example:

(15) He said the third-quarter **estimate** indicates profit for the nine months of $ 4.65 a share

In this context, it is clear that the (third-quarter) profit of some company is in question, but that the value ascribed to it, '$ 4.65 a share' is an estimated figure. Apart from adding brevity, approximation frequently occurs in contexts where forward-looking, and hence approximated, statements are made before the actual numbers become details of historical fact.

3.6. AGENTS and THEMES

Just as in other semantic role frameworks, we define the roles THEME and AGENT to represent the "key players" in a quantitative fact. The central distinction between the two is one of obliqueness: Whereas AGENTS play a direct role in influencing a quantity, THEMES relate to a quantity obliquely. This is exemplified in in the following sentence, which features three AGENTS (bolded) and a THEME (italicized).

(16) **New England Electric**, based in Westborough, Mass., had offered $ 2 billion to acquire *PS of New Hampshire*, well below the $ 2.29 billion value **United Illuminating** places on its bid, and the $ 2.25 billion **Northeast** says its bid is worth.

Here, the AGENTS each have control over the amount they bid, whereas the THEME of that bid, 'PS of New Hampshire', at most indirectly influences the bid values. Seen another way, the AGENTS determine a bid value which is imputed to the THEME.

QSRL also classifies as THEMES entities which are parametrized by the quantity in question, as in, a bond (THEME) associated with a yield (a QUANTITY):

(17) **The bonds**, rated double-A by Moody's and S&P, were priced to yield from 6.20 in 1992 to 7.10 in 2008 and 2009.

It is often the case that the THEME associated with a quantity is mentioned explicitly, but the quantity in question is left implicit. Sentences (10)-(12), for example, exhibit a form of synecdoche in which a company name stands in for the price of its stock. As we discuss in the next section, implicit information is a major challenge for numerical information extraction.

3.7. Parts and WHOLES

We define a WHOLE argument that is entity-like, but semantically distinct from THEMES and AGENTS. Note we are not the first to do so; part/whole is considered to be a major semantic relationship (Miller et al., 1990; Girju et al., 2003). In the data we have examined, WHOLES manifest in two ways. The first is the common sense of the term as it associates with percents, as in a pie-chart. Consider:

(18) Of **the 1,224 companies surveyed**, 31 % expect to cut spending on plant equipment and machinery, while only 28 % plan to spend more.

In the sense used here, the bold-faced WHOLE argument describes a set of companies, of which a portion will cut spending and another portion will increase it.

In syntax, WHOLE phrases are often linked with a %-value by way of the preposition *of*, as in '% of people', or '% of all surface waters.'

Another role we identify with WHOLE is the description of some entity that is comprised of, or bundles, some set of existential quantities, as in:

(19) In addition, the **$ 3 billion bid** includes $ 1 billion debt that will be assumed by IMA, $ 600 million of high-yield junk bonds that will be sold by First Boston Corp. and $ 285 million of equity.

This sense of WHOLE is of course closely related to the one previously described.

3.8. TIME

Quantities vary over time. In language, they are described with temporal modifiers referring to *when* a quantity takes on particular value, as in the sentence (1), and in the following comparative sentence from a recent Wall Street Journal article:

(20) The yield on the benchmark 10-year Treasury note settled at 2.542% **Tuesday**, compared with 2.480% **Monday**.(Goldfarb and Kruger, 2018)

Here, a QUANT, the yield, of some THEME, the benchmark 10-year Treasury note, takes on two different VALUES, 2.480% and 2.542%, at distinct points in TIME—Monday and Tuesday, respectively.

We define TIME to be the time at which a quantity takes on a particular value, or a quantitative event. Common time arguments in financial discourse include days of the week, financial quarters, or years. More formally, following (Angeli et al., 2012), values taken on by TIME are ranges on a temporal continuum.

We identify another temporal argument, REFERENCE_TIME (like VerbNet's INIT_TIME) which is often necessary for contextualizing and disambiguating change events. Consider the following:

(21) The Belgian consumer price index rose a provisional 0.1% in October **from the previous month** and was up 3.65% **from October 1988**, the Ministry of Economic Affairs said.

Here, the extent of the described changes undergone by the Belgian consumer price index are measured at the same time, 'October' (1989, that is). The values 0.1% and 3.65% stem from comparing the price index against values taken on at the respective REFERENCE_TIMEs 'the previous month' and 'October 1988'.

Of course, application-specific exigencies may call for refinements on this admittedly simple temporal representation, though in a survey of WSJ sentences we find TIME and REFERENCE_TIME to give ample coverage.

One candidate for refinement may be forward-looking temporal arguments associated with financial entities such as bonds. Consider:

(22) Capital appreciation bonds are priced to yield to maturity from 7.10% **in 2003** to 7.25% **in 2007 and 2008**.

The subtlety in question is that, technically speaking, bonds are priced at a specific point in time, in this case in the late 1980s, but the yield at maturity does not itself realize until some later date.

Thus, by one reading, the bold-faced phrases above refer to TIMEs, associated with yields on capital appreciation bonds. By another reading, they refer to some future, promissory time, distinct from the time at which those yields are assessed.

3.9. PLACE

Geographic location is another important conditioning variable for quantities. For example, companies with global operations will report the performance of efforts in a specific place, as in the following statement from a 2013 Mattel earnings call:

(23) And in **Latin America**, we achieved about $1 billion in sales for the third year in a row despite some economic headwinds.(Mattel Inc., 2014)

and the following Wall Street Journal sentence (ellipsis ours):

(24) **World-wide** sales of Warner-Lambert's non-prescription health-care products ... increased 3% to $ 362 million in the third quarter; **U.S.** sales rose 5% last year

In the latter, the change in two quantities (sales of non-prescription healthcare products) associated with a theme (Warner-Lambert) vary according to the geographic region over which those sales are considered.

Of course, the above is but one example of the way in quantities can be indexed to location. Another salient example of location modifiers in the financial domain include statements about economy level-trends, as in

(25) Bourbon makes up just 1% of **world-wide** spirits consumption but it represented 57% of **U.S.** liquor exports last year

Here, a discrepancy is observed between two distinct, but semantically related quantities—liquor exports and spirits-consumption—when indexed to location.

3.10. CONDITION and CAUSE

Sometimes, quantities that manifest in text can only be said to obtain subject to the satisfaction of certain future CONDITIONS. Consider the following statement from a recent WSJ article about the implications of a recently passed tax bill:

(26) Farmers would get a smaller deduction—about 20% of income—**if they sell grain or other farm products to privately held or investor-owned companies like Mr. Tronson's**.(Bunge and Rubin, 2018)

Without sensitivity to these conditional statements, one might infer that a guaranteed implication of the tax bill is that farmers receive a deduction of 20% of their income. However, what the above sentence actually says is that such deductions would obtain *given* the satisfaction of the condition in bold-face.

Similarly, in the following sentence

(27) The Short Term Bond Fund...would deliver a total return for one year of about 10.6% **if rates drop one percentage point** and a one-year return of about 6.6% **if rates rise by the same amount**.

one might deduce the seemingly contradictory facts that one-year returns on the Short Term Bond Fund are, simultaneously, 10.6% and 6.6%. Of course, the above sentence differentiates between these scenarios with the use of the bold-faced, opposing conditional antecedents.

Relatedly, quantities are described as the result of conditions that have already been satisfied. These frequently manifest in because-phrases, as in the following sentence:

(28) Fireman's Fund Corp. said third-quarter net income plunged 85% to $7 million from last year's $49.1 million ... because of **ravages of Hurricane Hugo and increased reserves for legal expenses**.

in which a decline in third-quarter net income from one year to the next was caused by a natural disaster.

More generally, a CAUSE argument is some phrase describing a state of affairs in the world that *has already occurred*, and upon which the possibility or plausibility of a quantity is conditioned.

3.11. The SOURCE of information

Another entity-like role in quantitative facts is the SOURCE of the quantity in question. An obvious function of the SOURCE is signaling the credibility of a fact. More subtly, it allows the reader to interpret a quantitative analysis on the basis of the source's motives and status in a discourse. Consider, for example, the following WSJ sentence:

(29) **Mr. Einhorn of Goldman Sachs** estimates the stock market will deliver a 12% to 15% total return from appreciation and dividends over the next 12 months–vs. a "cash rate of return" of perhaps 7% to 8% if dividend growth is weak.

Aside from analysts' perspectives, financial discourse cite industry-specific metrics reported by third-party analysis firms, as in the following citation of ShopperTrak from a Mattel earnings call:

(30) Consumers came out much later and less frequently to brick-and-mortar stores with **ShopperTrak** showing retail foot traffic in-stores to be down as much as 15%.(Mattel Inc., 2014)

Recall our earlier discussion in which we used deliberate and direct action to be the standard distinguishing AGENTS from THEMES. In sentences like (22), it seems that deliberateness applies of the SOURCE as well. This occasional commonality between AGENTS and SOURCES may serve a source of confusion. However, SOURCE arguments are distinct by virtue of their third-party, outsider status with respect to a quantitative fact.

4. Worked Examples

QSRL is designed to capture a diverse range of quantities across multiple syntactic categories, and has a significant set of roles for doing so. To show how it all works together, we apply QSRL in full to some of the data we have previously considered. Each sentence considered is excerpted herein for the reader's convenience.

Let us begin with a statement about a company's sales, focusing on the particular numerical mention '3%':

(31) World-wide sales of Warner-Lambert's non-prescription health-care products... increased **3%** in the third quarter.

Here, the QUANTITY associated with the bold-faced number is 'sales'. More particularly, 3% denotes a positive change (hence, a `relative` MANNER of measure), in 'sales'.

The sentence provides several contextualizing details about this quantity. Namely, the sales in question were for 'Warner-Lambert's non-prescription health-care products', the sales in question are global, and were achieved in the third quarter Putting these together gives the following analysis of the quantitative fact:

QUANTITY	'sales'
THEME	'Warner-Lambert's non-prescription...'
VALUE	'3'
UNIT	%
SIGN	+
MANNER	`relative`
TIME	'the third quarter'
LOCATION	'world-wide'

Importantly, QSRL takes a syntactically invariant perspective on quantities. In the following, the word in the sentence describing what the number measures is the verbform 'offered'

(32) New England Electric, based in Westborough, Mass., had offered **$ 2 billion** to acquire PS of New Hampshire

Here, the number in question measures the magnitude of the offer made *by* New England Electric *for* PS of New Hampshire. Thus the former is labeled an AGENT and the latter a THEME. Here we do not include 'Westborough, Mass.' as PLACE as it does not serve as a geographic modifier to the offer, in the way that 'world-wide' modifies 'sales' in (31), for example. At best 'word-wide' is a modifier on the THEME, which might be included in some augmented role set.

QUANTITY	'offered'
AGENT	'New England Electric'
THEME	'PS of New Hampshire'
VALUE	'2 billion'
UNIT	$
SIGN	+
MANNER	`absolute`

As a final example, we apply QSRL to a percent-quantified subject of a predicate:

(33) The poll ... with **65%** of those in companies of more than 15,000 employees saying stress-related problems were "fairly pervasive"

We interpret such constructions as `absolute` measures of the extent to which the predicate in question, here 'saying stress-related problems were 'fairly pervasive", obtains in the world. Hence we label the predicate as the QUANTITY:

QUANTITY	'stress related problems...pervasive'
VALUE	'65'
UNIT	%
SIGN	+
MANNER	`absolute`
WHOLE	'those in companies...15,000 employees'
SOURCE	'the poll'

The subject in such cases constitutes a WHOLE of which the percent-extent of the predicate is measured.

Note that we include phrases such as 'the poll' in our analyses of SOURCE, despite that when abstracted away from the text the precise referent of such a phrase is lost. We take this to be outside of the scope of QSRL itself, leaving it instead to be a matter of post-processing of the sort discussed in the next section.

QSRL and other role-labeling frameworks do not only tell you what to look for, but also what you are missing if you haven't found it. It is often the case that only some of the contextual information is mentioned in language. For example, the TIME at which a company had a specific profit may be suggested only from context rather than explicitly manifest in language, and the SOURCE of a given numerical figure may be the reporting document itself, such as an earnings call or press release. As described, this contextualizing information is important and should be preserved when possible.

5. Challenges New and Old in QSRL-based Information Extraction

In this section we review some of the major challenges of a QSRL-based information extraction. These are not insurmountable barriers. They are, rather, issues that are highlighted in the course of employing QSRL for close reading of actual data. Any robust quantitative information extraction system will have to address them.

5.1. Anchoring

Traditionally speaking, semantic role-labeling schemata are anchor-specific. For example, VerbNet is not just a list of roles, but a lexicon of verbs and the roles those particular verbs select for. FrameNet, on the other hand, defines a set of scenarios, each of which can be triggered by some list of so-called "target" words.

On the other hand, noting that (i) QSRL is far less granular than these other schemata and (ii) assuming that QSRL annotation is anchored to the appearance of numerical mentions, like '$ 2 million', which are easy to identify using simple rules, the anchoring task in our case is considerably simpler.

5.2. Normalization

All of the standard problems of normalization in information extraction apply to QSRL. These include entity-linking (Rao et al., 2013), or the task of mapping entity mentions in text, as in 'Barack Obama' or 'President Obama' to a common entity reference; time normalization (Chang and Manning, 2012; Angeli et al., 2012), or the task of mapping a temporal phrase, as in 'the 1988 fiscal third quarter' to a domain-independent representation of time; and quantity normalization (Roy, 2017), which is like time normalization but maps numerical mentions, as in '$ 1 million' and '$ 1,000,000' to domain-independent, standard representations with their associated units.

In addition to these, QSRL introduces a new challenge for normalization: Its syntactic invariance requires a way to map essentially synonymous words like 'earned' (a verb) and 'earnings' (a noun) to the same underlying form.

5.3. Implied arguments

A common issue in information extraction that also applies to QSRL is that of implied arguments, in which some meaningful piece of information is only implied by context, rather than being explicitly manifest in syntax.

We have already seen one example of implied arguments so far in sentence (14), where a company name stands in for its stock price (and hence the quantity in question is not explicitly obvious). Another commonly problematic form of implied arguments for information extraction is *gapping* (Schuster et al., 2017). As in

(34) Mary drinks coffee, and John tea.

in which there is an implicit 'drinks' predicate linking 'John' and 'tea' that is resolved constructionally with reference to the preceding clause, 'Mary drinks coffee'. We find several instances in the data of gapping constructions in quantitative facts, e.g.

(35) Merck's **profit climbed** 25 %, Warner-Lambert's 22 % and Eli Lilly's 24 % .

Note that here, the QUANTITY 'profit' and the MANNER 'climbed' appear only in the first clausal conjunct, and are elided in the next two. In these cases, common rule-based approaches to IE that use sets of predefined patterns will fail due to their context-insensitivity.

One quantity-specific implied argument phenomenon is that of unit ellipsis, which is similar to gapping constructions in that a unit (or some part of it) is left out, and thus only implied by context. This is exemplified in the following sentence.

(36) Third-quarter net income slid to $ 5.1 million, or **six cents a share**, from $ 56 million, or **65 cents**, a year earlier.

Here, the second bold-faced value '65 cents' exists in parallel with the first 'six cents a share'. From the context, it is clear that, implicitly, the unit is 'cents a share' despite that part of this unit is left out.

As another more extreme example, consider the following

(37) Sumitomo Metal Mining fell **five yen** to **692** and Nippon Mining added **15** to **960**.

Here, the unit associated with the first numerical mention in the text is subsequently elided in the following three numerical mentions. We can assume that because the change in stock price of Sumimoto Metal Mining is cited in yen, the final value '692' is as well. Inferring units on the two subsequent numerical mentions requires more complex real-world knowledge: Namely, that at the time of utterance Nippon Mining was also a Japanese company whose price was thus stated in yen.

5.4. Intersententiality

Despite the way they are commonly presented and employed, SRL annotation schemata are not limited to operate within the confines of the sentence. Indeed, contextualizing details of quantitative facts often appear outside of the sentence in which a particular value is mentioned. Consider the following:

(38) For **the quarter**, BCA generated revenue of $15 billion on a record 202 deliveries. **Operating margins** of **9.9%** reflect higher 787 margins and strong operating performance on production programs(Boeing Company, The, 2017b)

Here, the QUANTITY 'operating margin' and its VALUE/UNIT '9.9%' are mentioned in the second sentence, but the TIME argument associated with this quantity carries over from the previous sentence.

Such intersentential information extraction remains overwhelmingly under-explored, partly because it is at least as hard as its intrasentential counterpart, which is itself an unsolved problem. Nevertheless, recent work (Peng et al., 2017) has shown promise for end-to-end algorithms for such IE. Thus, any annotation effort using QSRL should take care to acknowledge its intersentential scope.

6. Concluding Remarks

We began this paper by observing that text from a variety of sources contains a great deal of quantitative financial information. Importantly, many of these facts go beyond the information in the standard tables of financial reporting. They provide critical perspectives at the levels of company, industry, and economy.

In order to extract these facts automatically, we must first represent them. We define a new annotation schema for doing so, called QSRL, that homes in on quantities in text and the context in which they manifest.

Applying QSRL to linguistic data reveals several interesting challenges for quantitative information extraction in general. Quantitative language is replete with traditionally challenging phenomena for semantic processing such as implied arguments and intersententiality.

We take QSRL as defined here to be a stepping stone for annotation efforts in quantitative information extraction. Upon further development of QSRL, future work will employ it for the purposes of supervised quantitative information extraction.

7. Bibliographical References

Angeli, G., Manning, C. D., and Jurafsky, D. (2012). Parsing Time: Learning to Interpret Time Expressions. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 446–455. Association for Computational Linguistics.

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The Berkeley FrameNet Project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*, pages 86–90. Association for Computational Linguistics.

Boeing Company, The. (2017a). News Release, October.

Boeing Company, The. (2017b). Q3 2017 Earnings Call, October.

Bunge, J. and Rubin, R. (2018). Tax Law's Effect Fuels Farm Outcry. *The Wall Street Journal*, CCLXXI(9).

Chang, A. X. and Manning, C. D. (2012). SUTime: A Library for Recognizing and Normalizing Time Expressions. In *LREC*, volume 2012, pages 3735–3740.

Demers, E., Vega, C., et al. (2008). Soft Information in Earnings Announcements: News or Noise?

Devlin, K. (2006). Situation Theory and Situation Semantics. *Handbook of the History of Logic*, 7:601–664.

Engelberg, J. (2008). Costly Information Processing: Evidence from Earnings Announcements.

Forbus, K. D. (1984). Qualitative Process Theory. *Artificial Intelligence*, 24(1):85–168.

Girju, R., Badulescu, A., and Moldovan, D. (2003). Learning Semantic Constraints for the Automatic Discovery of Part-Whole Relations. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 1–8. Association for Computational Linguistics.

Goldfarb, S. and Kruger, D. (2018). Investors Prepare for Inflation. *The Wall Street Journal*, CCLXXI(9).

He, L., Lewis, M., and Zettlemoyer, L. (2015). Question-Answer Driven Semantic Role Labeling: Using Natural Language to Annotate Natural Language. In *EMNLP*, pages 643–653.

He, L., Lee, K., Lewis, M., and Zettlemoyer, L. (2017). Deep Semantic Role Labeling: What Works and What's Next. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Kipper Schuler, K. (2005). *VerbNet: A broad-coverage, comprehensive verb lexicon*. Ph.D. thesis, University of Pennsylvania.

Lee, J. (2014). Scripted Earnings Conference Calls as a Signal of Future Firm Performance. *SSRN scholarly paper ID*, 2426504.

Levin, B. (1993). *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press.

Madaan, A., Mittal, A., Mausam, G. R., Ramakrishnan, G., and Sarawagi, S. (2016). Numerical Relation Extraction with Minimal Supervision. In *AAAI*, pages 2764–2771.

Mattel Inc. (2014). Q4 2013 Earnings Call, January.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. J. (1990). Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography*, 3(4):235–244.

Palmer, M., Gildea, D., and Kingsbury, P. (2005). The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71–106.

Peng, N., Poon, H., Quirk, C., Toutanova, K., and Yih, W.-t. (2017). Cross-Sentence N-ary Relation Extraction with Graph LSTMs. *Transactions of the Association for Computational Linguistics*, 5:101–115.

Rao, D., McNamee, P., and Dredze, M. (2013). Entity Linking: Finding Extracted Entities in a Knowledge Base. In *Multi-source, Multilingual Information Extraction and Summarization*, pages 93–115. Springer.

Roy, S., Vieira, T., and Roth, D. (2015). Reasoning about quantities in natural language. *Transactions of the Association for Computational Linguistics*, 3:1–13.

Roy, S. (2017). *Reasoning about quantities in natural language*. Ph.D. thesis, University of Illinois at Urbana-Champaign.

Schuster, S., Lamm, M., and Manning, C. D. (2017). Gapping constructions in Universal Dependencies v2. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*.

XBRL International Inc. (2013). Extensible Business Reporting Language (XBRL), December.

8. Language Resource References

Mitchell P. Marcus, Beatrice Santorini, Mary Ann Marcinkiewicz, Ann Taylor. (1999). *Treebank-3 LDC99T42*. Linguistic Data Consortium, None, 1.0, ISLRN 141-282-691-413-2.

Towards a Multilingual Financial Narrative Processing System

Mahmoud El-Haj[1], Paul Rayson[1], Paulo Alves[2], and Steven Young[3]
[1]School of Computing and Communications, Lancaster University, UK
[2]Management School, Lancaster University, UK
[3]Universidade Católica Portuguesa, Portugal
{[1]m.el-haj, p.rayson}@lancaster.ac.uk, [2]palves@porto.ucp.pt, [3]s.young@lancaster.ac.uk

Abstract

Large scale financial narrative processing for UK annual reports has only become possible in the last few years with our prior work on automatically understanding and extracting the structure of unstructured PDF glossy reports. This has levelled the playing field somewhat relative to US research where annual reports (10-K Forms) have a rigid structure imposed on them by legislation and are submitted in plain text format. The structure extraction is just the first step in a pipeline of analyses to examine disclosure quality and change over time relative to financial results. In this paper, we describe and evaluate the use of similar Information Extraction and Natural Language Processing methods for extraction and analysis of annual financial reports in a second language (Portuguese) in order to evaluate the applicability of our techniques in another national context (Portugal). Extraction accuracy varies between languages with English exceeding 95%. To further examine the robustness of our techniques, we apply the extraction methods on a comprehensive sample of annual reports published by UK and Portuguese non-financial firms between 2003 and 2015.

Keywords: Financial Narrative Processing, NLP, annual reports, Information Extraction, Multilingual

1. Introduction

There are a number of different financial reporting requirements and legislative frameworks for national and international companies in terms of how they must report to their shareholders, potential investors and the financial markets. Companies produce a variety of reports containing both textual and numerical information at various times during their financial year, including annual financial reports, quarterly reports, preliminary earnings announcements and press releases. Additionally, conference calls with analysts are transcribed and made available publicly, and other sources of information such as media articles and online social media are employed by companies, analysts and the general public. This creates a vast financial information environment which can be impossible to keep track of. Previous academic research in accounting and finance areas has tended to focus on numerical information, or small scale manual studies of textual information. Over the last few years, we have been able to contribute to the scaling up of the textual analysis component by applying Information Extraction (IE), Natural Language Processing (NLP) and Corpus Linguistics (CL) methods to the data. [1] have focussed on the UK context where annual financial reports are released in glossy PDF format with a variety of different looser structures, and these have made it harder to apply normal research methods on a large scale. In contrast, much of the previous research has been targetted at the US context where annual 10-K forms are required to follow a rigid structure with a standard set of headings, and are written in plain text. A standard format enables more straightforward selection of relevant sections for further analysis. In this paper, we describe not only the structure detection and extraction process that we have designed and implemented for English annual reports, but also our initial work to extend this research to another national context, in this case to Portugal. We report on our experiments to port the system from English annual reports to those published in Portuguese, and describe the adaptations made to the system to enable this. Our methods extract information on document structure which is needed to enable a clear distinction between narrative and financial statement components of annual reports and between individual sections within the narratives component. The resulting software is made freely available for academic research.

[1]For more details, see the CFIE projects described at http://ucrel.lancs.ac.uk/cfie/

2. Related Work

Previous related work on financial narrative analysis has taken place in a number of areas including accounting and finance research, natural language processing and corpus linguistics. Some early approaches in the accounting and finance literature employed painstaking manual approaches and were therefore limited in scale due to time constraints. Further studies have become larger scale but are still using manually constructed word lists for detecting features without considering local context for disambiguation purposes or more advanced machine learning methods. Well known studies include one by Li (2010) which considered forward-looking statements in 10-K (annual) and 10-Q (quarterly) filings in the US and found a link between positive statements and better current performance and other indicators. Li also found that general content analysis dictionaries (such as Diction, General Inquirer and LIWC) are not helpful in predicting future performance. Loughran and McDonald (2011) also found that negative words in the general purpose Harvard Dictionary were not typically considered as negative in financial contexts, and so were less appropriate than domain specific versions. They also considered US 10-K reports for their study. Schleicher and Walker (2010) found that companies with impending performance downturns bias the tone in outlook sections of the financial narrative. A good survey of text analysis methods in accounting and finance research was recently published by Loughran and McDonald (2016).

In the natural language processing research area, previous research has been carried out to extract document structure mainly from scientific articles and books (Doucet et al., 2009; Teufel, 2010; McConnaughey et al., 2017). Other than this, there has been much recent work in using text mining and sentiment analysis, in particular to Twitter, with the goal of predicting stock market performance (Devitt and Ahmad, 2007; Schumaker, 2010; Im et al., 2013; Ferreira et al., 2014; Neuenschwander et al., 2014) although presumably any really successful methods would not be published. From the other end of the language analysis spectrum, in linguistics, there has been a large amount of research on the language of business communication. Merkl-Davies and Koller (2012) introduced the Critical Discourse Analysis (CDA) approach to the accounting community and showed how it can be used to systematically analyse corporate narrative documents to explore how grammatical devices can be used to obfuscate and guide interpretations. Brennan and Merkl-Davies (2013) considered communication choices and devices which contribute to the phenomena of impression management, where individuals or companies use language to present themselves favourably to others.

3. Dataset

In our work we focus on UK and Portuguese annual reports for large firms listed on the stock exchange market in each country. The number of UK annual reports exceeds 10,000 of mostly UK non-financial firms listed on the London Stock Exchange. The annual reports cover a period of years in the range 2003 and 2014. The extraction methods have been tested and evaluated on English annual reports and were later adapted to work with other languages. We collected 627 Portuguese annual for 77 firms for the period for the period 2006-2015. All firms are listed on the Portuguese Stock Exchange. The annual reports were collected automatically from Perfect Information[2].

3.1. Description of Dataset

We first start with explaining an annual report is. An annual report is an analysis and assessment of the financial trend of the business over the past year. An annual report consists of a description of the accounting activities seen within the report. For example, a description of the principles used for determining the accounting items in both the income statement and the balance sheet. An annual report could also include information on the events that have influenced the company's accounting throughout the year, a statement from management showing an accurate picture of the company's economic standing and development, and an auditor's report.

It was not until legislation was enacted after the stock market crash in 1929 that the annual report became a regular component of corporate financial reporting. Typically, an annual report will contain the following sections:

- Financial Highlights
- Letter to the Shareholders
- Narrative Text, Graphics and Photos
- Management's Discussion and Analysis
- Financial Statements
- Notes to Financial Statements
- Auditor's Report
- Summary Financial Data
- Corporate Information

Most of the published annual reports are in PDF file format. The different variation of annual reports' formatting makes it difficult to automatically extract relevant information or even detect the report's structure. The annual reports vary in respect to their style and number of pages. In the US firms are required to disclose their annual reports by following and filling a preset template by the US Securities and Exchange Commission (SEC). This allows a standard structure to be followed by each company making it easy to extract information and easily detect structure. In contrast to the US, stock exchange-listed firms in UK and Portugal do not present their financial information and accompanying narratives in a standardised format when creating annual reports. Firms in the aforementioned countries have much more discretion regarding the structure and content of the annual report. Added to this is the problem of nomenclature: no standardised naming convention exists for different sections in UK annual reports so that even firms adopting the same underlying structure and content may use different terminology to describe the same section(s).

Table 3.1. shows the dataset size in words in addition to the number of reports for each language.

Language	Reports	Words
English (UK)	11,009	300M
Portuguese	396	7.50M

Table 1: Dataset Size

4. Extraction Methods

To extract information from our dataset of PDF annual reports we used Information Extraction (IE) and Natural Language Processing (NLP) methods and techniques. This helps in extracting sections and their narratives. The methods automatically detect the annual report's table of contents, synchronise page numbers in the native report with page numbers in the corresponding PDF file, and then use the synchronised page numbers to retrieve the textual content (narratives) for each header (hereinafter section) listed in the table of contents. The extraction methods rely on the table of contents by using section heading presented in the table of content to partition into the audited financial statements component of the report and the "front-end" narratives component, with the latter sub-classified further into a set of generic report elements including the letter to shareholders, management commentary, the governance statement, the remuneration report, and residual content.

[2]http://www.perfectinfo.com

4.1. Structure Extraction Process

This section explains in details the steps and process needed to be able to detect the structure of the PDF annual reports of both UK and Portuguese datasets. The process was first applied to the 10,000 UK annual reports (Section 4.1.), we then applied the same process to the smaller Portuguese dataset.

As mentioned in Section none of the UK or Portuguese annual reports follow a standard reporting template as in the US Stock Exchange. Firms and management in the UK have more discretion when it comes to the the format, structure and the contents of the annual reports. On the other hand the US Securities and Exchange Commission forces firms to follow a standard format and a pre–labeled annual reports template which they publish in HTML file format. This has helped in creating a reporting standard making it easy for investors, firms and analysts to access and acquire information automatically from a bulk of annual reports. This is different in the UK where firms tend to publish their annual reports in PDF file format. Despite being cross-platform and a portable file format it is deemed a difficult task to automatically extract information from PDF annual reports since companies' reports vary significantly especially when it comes to the contents and the section headers. In order to automatically analyse a large dataset of UK annual reports we first needed to automatically detect the structure of the PDF annual reports so we can extract the information needed.

To detect and extract the structure of the annual reports each PDF file goes through the following five steps: 1) detecting the contents-page, 2) parsing the detected contents-page and extracting the sections, 3) detecting page numbering, 4) adding the extracted sections to the annual report PDFs as bookmarks, and 5) using the added bookmarks to extract the narrative sections under each heading.

4.1.1. Detecting the Contents Page

An annual report contents page includes information about the main sections of the report and its associated page numbers. Information in the contents page helped us detect the structure of the annual reports. However, detecting the contents page was not a straightforward task. We created a list of gold–standard section names extracted manually from the contents page of a random sample of 50 annual reports. We filtered the gold–standard keywords by removing duplicates and preserving the structure of how they appeared in the annual reports. We matched each page in the annual report against the list of section names in gold–standard, then we selected the page with the highest matching score as the *potential* contents page. The score was calculated by an increment of 1 for each match. To improve the matching process and avoid false positives, we match the gold–standard keywords against lines of text that follow a contents–page–like style (e.g. a section name followed by a page number, such as "Governance Report 22").

4.1.2. Parsing the Contents Page

In order to get the structure of the annual report we automatically parse the selected contents page by extracting the name of each section and its associated page number. To do this we matched each line of text in the selected contents page against a regular expression commands that will extract any line starting or ending with a number between 1 and the number of pages of the annual report.

We built a simple filtering tool that filters out any block of text that matches our regular expression commands. This is done by removing text containing addresses, dates, and postal codes. The filtering tool can also detect email addresses, websites, references to branches and locations using regular expression commands and a gazetteer.

We differentiate between dates and actual page numbers to avoid extracting incorrect section headers. However, lines containing text such as an address (e.g., 23 Robert Avenue) might still be confusing for the tool. We tackled this problem by matching the list of extracted sections against a list of gold–standard section synonyms which we explain in more details in Section 4.1.5..

The structure of the PDF files makes it difficult to extract text in its actual format. Extracting plain text from PDFs results in many line breaks being added in between the text. This makes extracting a section name that is split into two lines a difficult task. To tackle the problem of broken sections (i.e., appearing on two lines or more), we implemented an algorithm to detect broken section headers and fix them by concatenating lines that end or begin with prepositions such as 'of', 'in' ...etc. The algorithm also concatenates sentences ending with singular or plural possessives, symbolic and textual connectors (e.g. 'and', 'or', '&'...etc), and sentences ending with hyphenations. This method was also adapted to Portuguese prepositions and other stop-words needed to concatenate lines of text by forming a list of most common stop-words for each language.

4.1.3. Detecting Page Numbering

The page numbers appearing on the contents page do not usually match with the actual page numbers in the PDF files. For example, page 4 in the annual report could refer to page 6 in the PDF file, which may lead to incorrect extraction[3]. We address this problem by creating a page detection tool that crawls through annual report pages taking three consecutive pages in each iteration. The tool aims to extract a pattern of sequential numbers with an increment of 1 (e.g. 16, 17, 18) but with the complex structure of the PDF files this has been proven to be a difficult task. The tool starts by reading the contents three pages at a time starting from the report's number of pages minus one. For example, assume we are trying to detect the page numbering pattern for a report of 51 pages. The tool starts by extracting text from pages 48, 49 and 50. A regular expression command is then used to extract all the numbers in each page contents that is made up of maximum three digits creating a vector of numbers for each page. Figure 1 shows a sample of 3 vectors for the pages 48, 49 and 50. As shown in Figure 1 the algorithm will only keep numbers that are within a range of 10 pages those linked with small double arrows. The algorithm will then try to form a pattern of sequential

[3]The algorithm responsible for extraction of sections uses start and end page numbers to locate the text and therefore accurate page numbers are required.

numbers with an increment of 1. Figure 1 shows that the pattern 49, 50 and 51 (dark circles) has been found which is equivalent to a one page difference (*page–increment*) between the reports page numbering and those found in the PDF file. The tool will repeat the same process for all the pages in the annual report until it reaches pages 1, 2 and 3 where it stops.

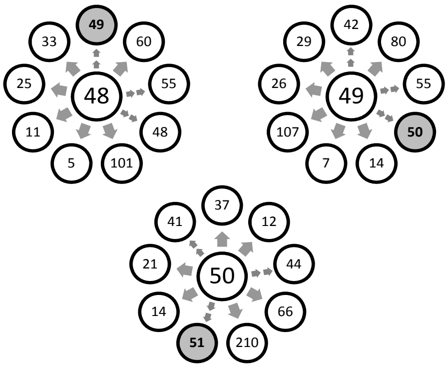


Figure 1: Detecting Page Numbering

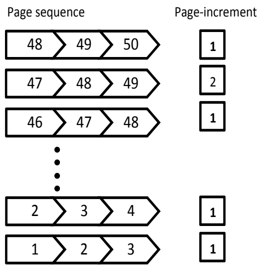


Figure 2: Popular Page Increment

As shown in Figure 2 for each 3 vectors the tool will store the page–increment in an array of numbers and at the end of the process the most popular (most frequent) page–increment will be selected as the difference between the annual report and the PDF numbering.
This process on the sample yielded an accuracy rate of more than 95%. Manual examination of the remaining less than 5% revealed the following reasons for non-detection:

- Encoding Error, unrecognised text
- Images or empty pages interrupting the sequence of pages
- Page numbers appeared on even or odd pages only
- Unusual numbering format (e.g. "001001001029" refers to page 29).
- Page numbers appeared in a written format (e.g. Twenty One)
- Page numbers restarted on each section
- Some pages had no page numbers available
- Every other page has two numbers (e.g. 26/27) with no numbers available on the next page
- Two pages on each PDF page
- Some other errors due to formatting

4.1.4. Adding Section Headers as Bookmarks

Using the sections and their correct page numbers from Sections 4.1.1. and 4.1.3. we implemented a tool to insert the extracted contents page sections as bookmarks (hyperlinks) to sample PDFs. This process helped in extracting narratives associated with each section for further processing (see Section 4.1.5. below).

4.1.5. Extracting Sections' Narratives

We implemented an automatic extraction algorithm to crawl through the data collection and, for each PDF file, extract all inserted bookmarks and their associated pages. Since UK firms do not follow a standard format when creating annual reports, a long list of synonyms are possible for a single section header. For example the section header "Chairman's Statement" may also appear as "Chairman's Introduction", "Chairman's Report" or "Letter to Shareholders". The same case applies to Portuguese as well. To solve this problem, we semi-automatically and by the help of experts in accounting and finance, created a list of synonyms for each of the generic annual report sections (see the list below). This was done by extracting all sections containing "Chairman", "Introduction", "Statement", "Letter to"...etc from a sample of 250 annual reports of 50 UK firms (the quoted unigrams were selected by the same experts). We refined the list by removing redundancies. The accounting experts then manually examined the list and deleted irrelevant or incorrect sections. We used the refined list as gold–standard synonyms to extract all the sections related to each of our generic sections (e.g. all sections about the "Chairman's Statement"). To overcome the problem of different word–order or additional words included in the headline (e.g. "The Statement of the Chairman"), we used *Levenshtein Distance* string metric algorithm (Levenshtein, 1966) to measure the difference between two sections. The Levenshtein distance between two words is the minimum number of single-character edits (insertion, deletion, substitution) required to change one word into the other. To work on a sentence level we modified the algorithm to deal with words instead of characters. All the sections with a Levenshtein distance of up to five were presented to the accounting expert.
We used the above process to create gold–standard synonym lists for the following 8 generic section headers that we wished to extract for further analysis:

1. Chairman Statement
2. CEO Review
3. Corporate Government Report
4. Directors Remuneration Report

5. Business Review
6. Financial Review
7. Operating Review
8. Highlights

Having detected and extracted section headers (or their gold–standard synonyms) and their sections, we then extract the sections' narratives using iText[4], an open source library to manipulate and create PDF documents (Lowagie, 2010), to apply our text analysis metrics, which include readability measurement and counting word frequencies using financial domain hand–crafted word lists.

5. Extraction Tools

We used the extraction methods described in Section 4. to create publicly available web and desktop tools for users to automatically and freely analyse annual reports in different languages. The tools deal with multilingual annual reports of firms within the UK and Portugal written in either English or Portuguese and distributed in PDF file format[5].

The tool is called CFIE-FRSE standing for Corporate Financial Information Environment (CFIE) -Final Report Structure Extractor (FRSE). The tool is available as a web application[6] or as desktop application, which is freely available on GitHub[7]. The tools detect the structure of annual reports by detecting the key sections, their start and end pages in addition to the narrative contents. This works for both languages. The tools provide further analysis for reports written in English such as readability metrics, section classification and tone scores. This is because the tool was built to analyse UK annual reports where we have a large dataset to train the system to provide an extra level of analysis.

The extra level of analysis will be made available for Portuguese at a later stage. For now we do not have enough reports for Portuguese to be able to train the system. As explained earlier the aim of this paper is to show that our extraction methods can be applied to a second language, a vital step towards fully analysing reports in other languages in the future.

6. Multilingual Extraction

In this section we explain the process we followed to extracting sections from annual reports in both English and Portuguese.

6.1. English

As mentioned earlier the work was first designed to analyse UK English annual reports (El-Haj et al., 2014). We automatically harvested more than 10,000 annual reports for firms listed on the London Stock Exchange (LSE). Prior to analysing the annual reports we first worked on sorting them by firm and we created our own unique report identifier which we called "LANCS_ID". Sorting annual reports was done semi-automatically were we used a Java tool to match firm names and extract the reports' years. This was followed by manual post editing to make sure the matching was correct. Firms without a match could be firms that do not exist anymore or firms with a new name due to merging with another firm, those had to be manually matched. PDF filenames do not contain a unique firm identifier. For example, reports collected from Perfect Information use a standard naming convention comprising firm name and publication year. We use filenames as the basis for a fuzzy matching algorithm that pairs firm names extracted from the PDF filename with firm names provided by Thomson Datastream. Matching on name is problematic because firms can change their name over the sample period. The matching procedure must therefore track name changes. To address this problem, we combine firm registration numbers and archived names from the London Share Price Database with Datastream's firm name archive in our fuzzy matching algorithm. For those cases where our algorithm fails to find a sufficiently reliable match, we perform a second round of matching by hand. Further details of the matching procedure, including a copy of the algorithm and a step-by-guide to implementing the matching procedure in SAS are available at `http//cfie.lancaster.ac.uk.8443/`. Licensing restrictions prevent direct publication of proprietary identifiers.

Annual report structures vary significantly across reporting regimes and therefore to make the initial development task feasible we focus on reports for a single reporting regime. We select the UK due to the LSE's position as largest equity market by capitalisation outside the US. The extraction process is nevertheless generic insofar as reports published in other reporting regimes and languages can be analysed by modifying the language- and regime-dependent aspects of our tool without editing the underlying Java source code. Further guidance will be provided in an online appendix, together with full technical details of our method, in due course.

Table 6.1. shows the structure detection and extraction accuracy for UK annual reports.

Number of downloaded annual reports	11,009
Number of reports analysed	10,820
% of correctly retrieved table of contents	98.28
% of correctly retrieved pages	95.00
% of correctly retrieved text from sections	95.00

Table 2: UK Annual Reports Analysis

As shown in the table the tool analysed more than 98% of the downloaded annual reports. Firms management in the UK have more discretion over what, where, and how much information on topics such as risk, strategy, performance, etc. is reported, this lead the reports to vary significantly in terms of structure and design. Despite the dissimilarity between the structure of the downloaded annual report, our methods were able to accurately analyse the majority of the reports. Those failing the analysis process were due to one

[4] `http://itextpdf.com/api`

[5] For now only the Desktop version of the tool can work with multilingual annual reports

[6] `https://cfie.lancaster.ac.uk:8443`

[7] `https://github.com/drelhaj/CFIE-FRSE`

of the following reasons:

1. The file does not allow the text to be extracted (image-based documents). This problem is more common in the early years of our sample (i.e. 2000-2005), as some of the annual reports were poor quality scanned files. Reports from the more recent years tend not to be of this type.

2. Reports with a table of contents that could not be read due to the limitation imposed by how the table was designed. For example where a table of contents is designed with numbers and text in two different columns, or where the table of contents is split into two pages which causes problems for the PDF library.

3. Absence of page numbers.

6.2. Portuguese

The adaptation of our software to other languages must deal with problems that are both specific to the financial reporting environment and to the language itself. As in most countries, Portuguese market regulations allow a certain degree of flexibility in relation to the content and structure of the annual report[8] and concerning a firm's governance structure. For instance, the board of directors (or its equivalent) and the Fiscal Committee can adopt different structures and names. As an example, we detected 7 alternative titles for the CEO's message, 12 different titles for the chairman's letter and 35 alternatives for the auditor and related governance mechanisms. We believe that this will be a common problem across the different language implementations. The approach we adopted was to list all the alternatives, create a list of synonymous and assign a unique classification to each alternative. On the other hand, the language related issues are specific to each language. During the implementation of the Portuguese version, we identified several different problems. Firstly, the English language is one of the few western languages that does not use phonetic modifications of common characters, such as "À", "Á", "Ã", "Â" and "Ç". These phonetic modifications are common in other languages and can also vary across countries. Secondly, Portuguese is a gender-based language, which increases the complexity in developing a list of stop-words to deal with the line breaking. One such example is the proposition "of", which can be translated as "de", "do", "da", "dos" and "das", depending on the gender of the following word. Thirdly, Portugal signed the Portuguese Language Orthographic Agreement of 1990[9]. This agreement changed the spelling of some words (e.g. the word "Accionistas", is now spelt "Acionistas",). It also allowed an alternative spelling for some words (e.g. the word "Sector" can also be spelt as "Setor",). During the transition period, which ended in 2015, the adoption of the new spelling was voluntary and different firms used different spelling variations for some words. As a result, the algorithm must recognise all spelling variations. To test the adaption of the software to Portuguese Annual Reports, we retrieved from Perfect Information all annual reports published in Portuguese by firms listed on Euronext Lisbon for the period 2006 to 2015, totaling 627 reports for 77 firms (Table 6.2.).

Year	2005	2006	2007
# Downloaded Reports	51	52	60
# Processed	23	26	38
%	%45	%5	%63
Year	2008	2009	2010
# Downloaded Reports	61	61	62
# Processed	37	38	40
%	%61	%62	%65
Year	2011	2012	2013
# Downloaded Reports	64	62	58
# Processed	44	43	42
%	%69	%69	%72
Year	2014	2015	Total
# Downloaded Reports	59	37	627
# Processed	36	29	396
%	%61	%78	%63

Table 3: Number of Reports Per Year

The software was able to process 396 reports (63%) of the annual reports. We then focused on understanding the reason for the non-processed reports and the accuracy of the processed reports. The software failed to process 231 reports:

- Table of contents does not exist: 62 reports
- Table of contents could not be detected: 52 reports
- Table of contents presented in an unconventional format: 45 reports
- Table of contents without page numbers: 39 reports
- Table of contents with more than one page: 12 reports
- Image based file: 21 reports

The adaptation to Portuguese was based on 2 steps. We started by listing all table of contents entries for 67 randomly selected annual reports. This procedure produced a list of 2,053 entries that, after cleaning for errors and minor differences, included 694 different table of contents entries.

[8] The Companies Act (Código das Sociedades Comerciais) and Portuguese market regulations require a firm's Annual Report to include, amongst other items a review of the firm's activities, performance and financial position, a description of the main risks and uncertainties, financial risk management goals and policies, including details of hedging operations and risk exposures, a description of subsequent events, the expected evolution of the firm and a proposed net income allocation and dividends. In addition, firms are required to submit a Corporate Governance Report. Firms can opt to include this report in the Annual Report or to submit a separate document. Disclosure requirements are summarised in Circular sobre Contas Anuais – 9th February 2017.

[9] An international agreement aiming at the creation of a unified orthography for the Portuguese language across all the countries with Portuguese as their official language.

This variety reflects the lack of standardisation of the structure of the annual report that is common to most countries. To deal with this problem, for the second step, we assigned each entry to a pre-defined section (Chairman, CEO, Performance, Auditor, Financial Statements and Other), which reflects the common structure of a Portuguese annual report at a very basic level. We also tested the accuracy of the adaptation to Portuguese by manually checking 100 annual reports processed and we concluded that the software performs with an accuracy comparable to the English implementation (Tables 6.1. and 6.2.).

7. Conclusion

The methods reported in this paper demonstrate the adaptability of our extraction and classification procedures to non-English annual reports published in regulatory settings other than the UK, and we examine Portuguese reports in this paper. This adaptation was achieved by developing methods that are language independent where the extraction process relies on the structure of the annual reports rather than the deep language characteristics. The methods will still require dictionaries and word-lists to be in the same language as the annual reports but the extraction process remains the same across languages. The reported work paves the way for investors, firms and analysts to access and acquire information automatically from a large volume of annual reports in languages other than English.

8. Acknowledgements

We acknowledge support for this research in three projects. The first Corporate Financial Information Environment (CFIE) project was funded (2012-14) by the Economic and Social Research Council (ESRC) (reference ES/J012394/1) and The Institute of Chartered Accountants in England and Wales (ICAEW). This work also continued in the Understanding Corporate Communications sub-project funded as part of the ESRC Centre for Corpus Approaches to Social Science (CASS) (reference ES/K002155/1). Most recently, the research is funded under the new project which started in January 2018 Analysing Narrative Aspects of UK Preliminary Earnings Announcements and Annual Reports (reference ES/R003904/1).

9. Bibliographical References

Brennan, N. M. and Merkl-Davies, D. M. (2013). Accounting narratives and impression management. In *The Routledge Companion to Communication in Accounting*.

Devitt, A. and Ahmad, K. (2007). Sentiment polarity identification in financial news: A cohesion-based approach. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 984–991.

Doucet, A., Kazai, G., Dresevic, B., Uzelac, A., Radakovic, B., and Todic, N. (2009). Icdar 2009 book structure extraction competition. In *Proceedings of the Tenth International Conference on Document Analysis and Recognition (ICDAR'2009)*, pages 1408–1412, Barcelona, Spain, July.

El-Haj, M., Rayson, P., Young, S., and Walker, M. (2014). Detecting document structure in a very large corpus of UK financial reports. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014.*, pages 1335–1338.

Ferreira, J. Z., Rodrigues, J., Cristo, M., and de Oliveira, D. F. (2014). Multi-entity polarity analysis in financial documents. In *Proceedings of the 20th Brazilian Symposium on Multimedia and the Web*, WebMedia '14, pages 115–122, New York, NY, USA. ACM.

Im, T. L., San, P. W., On, C. K., Alfred, R., and Anthony, P. (2013). Analysing market sentiment in financial news using lexical approach. In *Open Systems (ICOS), 2013 IEEE Conference on*, pages 145–149, Dec.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710.

Li, F. (2010). The information content of forward-looking statements in corporate filings—a naïve bayesian machine learning approach. *Journal of Accounting Research*, 48(5):1049–1102.

Loughran, T. and McDonald, B. (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance*, 66(1):35–65.

Loughran, T. and McDonald, B. (2016). Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*, 54(4):1187–1230.

Lowagie, B. (2010). *iText in Action*. Covers iText 5. Manning Publications Company.

McConnaughey, L., Dai, J., and Bamman, D. (2017). The labeled segmentation of printed books. In *Proceedings of the EMNLP 2017 conference*.

Merkl-Davies, D. and Koller, V. (2012). 'Metaphoring' people out of this world: a critical discourse analysis of a chairman's statement of a UK defence firm. *Accounting Forum*, 36(3):178–193, 9.

Neuenschwander, B., Pereira, A. C., Meira, W., and Barbosa, D. (2014). Sentiment analysis for streams of web data: A case study of Brazilian financial markets. In *Proceedings of the 20th Brazilian Symposium on Multimedia and the Web*, WebMedia '14, pages 167–170, New York, NY, USA. ACM.

Schleicher, T. and Walker, M. (2010). Bias in the tone of forward-looking narratives. *Accounting and Business Research*, 40(4):371–390.

Schumaker, R. P. (2010). An analysis of verbs in financial news articles and their impact on stock price. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media*, WSA '10, pages 3–4, Stroudsburg, PA, USA. ACL.

Teufel, S. (2010). *The structure of scientific articles: Applications to Citation Indexing and Summarization*. CSLI Studies in Computational Linguistics. Center for the Study of Language and Information, Stanford, California.

Trust and Doubt Terms in Financial Tweets and Periodic Reports

Martin Žnidaršič, Jasmina Smailović, Jan Gorše, Miha Grčar, Igor Mozetič, Senja Pollak
Jožef Stefan Institute
Jamova cesta 39, 1000 Ljubljana, Slovenia
martin.znidarsic@ijs.si

Abstract

In this paper we present a study on expressions of trust and doubt in financial tweets and official periodic reports of companies. We use the trust and doubt wordlists that we created and analyze the presence of trust and doubt terms in both textual collections after some domain-specific text processing. In tweets, we have found that doubt is more frequently expressed than trust and forms higher peaks. Next, we have analyzed the relation between the filing dates of reports and the peaks in financial tweets with regard to their overall volume, trust tweets volume and doubt tweets volume. The analysis indicates that the Twitter community reacts more often to the quarterly than yearly reports and that the peaks are usually at the day of report, not before or after. As a result of corresponding analysis of textual content in annual reports, we present the frequencies of different trust/doubt terms in these reports and indicate some notable differences among their use by different companies.

Keywords: trust, doubt, tweets, periodic reports, wordlists

1. Introduction

Given the popularity of on-line social networking platforms, such as Twitter or Facebook, there has been a growing body of literature focused on analyzing social media content and its relation to various economic, political and social issues. For example, studies have analyzed the relationship between the Twitter data and financial indicators (Bollen et al., 2011; Smailović et al., 2014; Sprenger et al., 2014; Ranco et al., 2015; Gabrovšek et al., 2017), voting results (Tumasjan et al., 2010; Borondo et al., 2012; Eom et al., 2015; Grčar et al., 2017), crime (Gerber, 2014) or public health (Paul and Dredze, 2011). Other studies have focused on more formal documents (companies' reports) that have been analyzed in relation to various phenomena, such as company's financial performance (Qiu et al., 2006; Hajek et al., 2014), the cost of capital (Kothari et al., 2009), or fraud detection (Goel and Uzuner, 2016). Interestingly, compared to numerous applications of sentiment analysis, the aspect of trust is not a very common phenomenon to study in financial texts, although it is an important component of business.

In social networks, trust can be assessed from different aspects: from analyzing which users trust each other (Adali et al., 2010), to estimating trustworthiness of posted information (Zhao et al., 2016), and measuring expressed trust regarding an entity mentioned in a post. In addition, the methods for assessing trust are diverse: one can analyze the network structure, interactions between users, or examine content of posts. Sherchan et al. (2013) discuss definitions, aspects, properties and models of trust, and provide a survey of trust in social networks. They categorize sources of information regarding trust into attitudes, behaviours and experiences, while methods for calculating trust are grouped into network-based, interaction-based and hybrid ones.

Also periodic reporting has become an appealing topic of research. The main goal of financial reporting is to ensure high quality, useful information about the financial position of firms, their performance and changes (IASB, 2015) to a wide range of users (e.g. investors, financial institutions, employees, the government). Firms publish annual (and other periodical) reports, in which they—as summarised by Fuoli (2017) —construct and promote positive corporate image and gain trust. Related research focuses on various aspects, including the devices used to create an ethical image in corporate social responsibility (CSR) reports (Aiezza, 2015), impression management in chairman's statements (Merkl Davies et al., 2011), differences in stance expressions (strongly related to trust building) between annual reports and CSR reports (Fuoli, 2017). El-Haj et al. (2016) focus only on performance sentences (in the UK Preliminary Earning Announcements) on which they also test machine-learning methods for their identification, as well as for identifying the expressed attribution (internal or external factor related to the expressed performance) and tone.

In our paper, we focus on explicit mentions of trust and doubt terms in financial communications. In a preliminary study of correlations between linguistic characteristics and financial performance of companies, limited to only four firms (Smailović et al., 2017), we have used the trust and doubt wordlist for the first time, and the analysis showed that doubt terms are correlated to the financial indicators of failure (interestingly, more than the words from more frequently used lexicons of positive and negative words). For this reason, in our paper we further explore the expression of trust and doubt in financial communication. We are particularly interested in trust and doubt terms in Twitter posts (tweets) and periodic reports, and in observing the reaction to the periodic reports on Twitter. We focus on companies from the Dow Jones Industrial Average 30 (DJIA) index in a two-year period (2014-2015).

After introducing the manually created trust and doubt wordlists (Section 2.), we describe a lexicon-based approach for assigning the tweets into the two categories (trust and doubt) (Section 3.1.) and analyze the reactions to periodic reports in Twitter (through the peaks in the volume of all tweets, trust tweets or doubt tweets). In Section 3.2., we focus on trust and doubt terms in annual reports and, based-on a frequency analysis, report on some differences in the usage of trust and doubt terms between different companies. Finally, we conclude the paper and present the future research steps.

2. Trust/Doubt Wordlists

The wordlists used in this paper contain manually collected (near) synonyms of words *trust* and *doubt* from WordNet (Fellbaum, 1998; Miller, 1995)[1] and online dictionaries[2]. In the current version (v1.1), we included 25 terms for trust and 77 terms for doubt. The wordlists are publicly available at `http://kt.ijs.si/data/trust_doubt_wl.zip`.

A selection of trust/doubt terms from the wordlists is shown in Figure 1. The part of speech (POS) of each term is specified in parentheses (*n*-noun, *v*-verb, *adj*-adjective), while the # sign denotes that a term is in its derived form.

Trust	Doubt
assurance(n)	disbelief(n)
#assurances(n)	#disbeliefs(n)
confidence(n)	disenchantment(n)
confident(adj)	disillusion(n)
credence(n)	disillusionment(n)
faith(n)	distrust(n)
faithful(adj)	distrust(v)
reliable(adj)	#distrusted(v)
reliance(n)	distrustful(adj)
reliant(adj)	distrustfulness(n)
...	...

Figure 1: A selection of terms from the trust and doubt wordlists.

The wordlists have been used for the first time in our preliminary study of reports of four companies (Smailović et al., 2017), where we have shown the correlation between the doubt terms and financial performance. In the current version, some new terms were added (and some mistakes removed)[3]. We also describe here the resource in more detail and use it on other datasets and for other purposes.

The resource the most similar to ours is a list of 30 trust and distrust related words, presented in the study of (Jian et al., 2000). The authors study terms used for expressing three types of trust (trust towards machines, towards humans and trust in general). In future, we will investigate if the terms are useful for extending our wordlists.

If based on our wordlists, one wanted to create the trust and doubt wordlists for another language, lexical resources with linked senses over different languages could be considered (e.g. BabelNet[4] or Multilingual WordNet[5]).

3. Wordlist-based study of Trust and Doubt in Financial Communication

We use the trust/doubt wordlists that are described in Section 2. for conducting a wordlist-based study on two types of text: (i) tweets, which discuss the DJIA 30 companies in the years 2014 and 2015, and (ii) the corresponding annual reports of the companies.

3.1. Tweets

We analysed the presence of trust and doubt terms in tweets about the 30 DJIA companies over a two-year period. The data was acquired by the Twitter Search API, where a query is specified by the stock cashtag. A cashtag is a word with the dollar sign as the first character. Cashtags with stock ticker symbols (short codes that represent specific stocks) are used in Twitter messages to refer to particular stocks or companies (e.g., "$MSFT" for Microsoft). We collected 5,570,817 tweets in the period from January 1, 2014 until December 31, 2015.

Lexicons have been frequently used as a resource for sentiment analysis, where documents (e.g. tweets) are assigned to positive or negative class based on the frequencies of words from the positive or negative wordlists (Loughran and McDonald, 2016). In our case, we use the lexicon-based approach to classify the tweets in the categories of trust and doubt. The initial approach for categorizing tweets as expressing trust and/or doubt is straight-forward: for each tweet, the trust value of a company-day combination is increased by 1, if at least one trust term from the trust wordlist is matched.[6] The same is repeated for each tweet for the doubt terms. This results (for each tweet) in the increase of 1 for trust, doubt, none or both values for the company-day. The aggregated results for all DJIA 30 companies are presented in Figure 2, where we display tweet volume, trust and doubt over time. As it can be seen from the figure, there are several peaks in doubt and a substantial increase in trust over several months in the year 2015.

Content analysis of the tweets revealed that there is a need for handling two important aspects, which were not taken into account by our basic approach for assessing trust and doubt. First, there is a need for handling negated trust/doubt terms, and second, we noticed that the word "trust" is used not only related to, e.g., *reliability* and *confidence*, but often also in the context of mentioning investment funds. We did not find the ambiguity of any of the other considered terms so prominent.

In order to handle negations, we checked if there is a negation word (e.g., *no*, *not*, *isn't*, *aren't*, *wasn't*, etc., but also grammatically incorrect forms such as *dont*, *didnt*, *wont*,

[1] `http://wordnet.princeton.edu`

[2] E.g., `http://thesaurus.yourdictionary.com/doubt`

[3] We added more than 10 new terms and corrected some minor mistakes (some forms were marked as base forms, but were in fact derived forms, we removed some duplicates etc.).

[4] `http://babelnet.org/`

[5] `http://globalwordnet.org/wordnets-in-the-world/`

[6] The approach for retrieving tweets based on trust/doubt terms uses unique terms from the trust/doubt wordlists, without taking into account their POS tags or information regarding their derived forms.

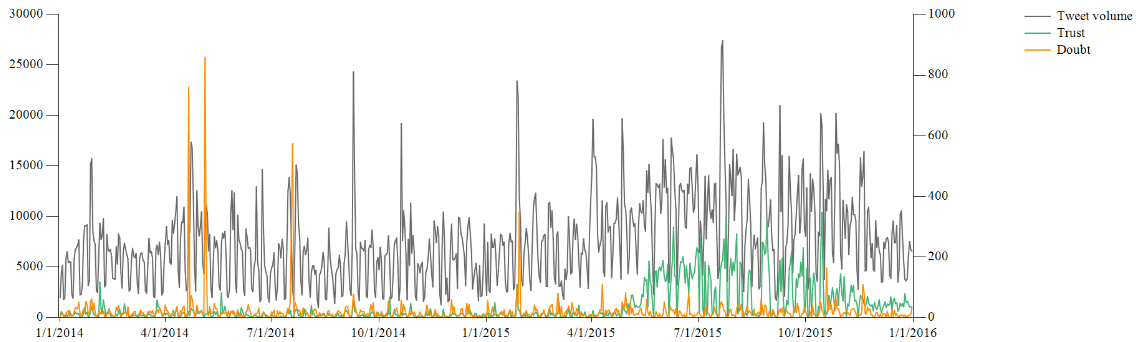


Figure 2: Tweet volume (gray, left y axis), trust (green, right y axis) and doubt (orange, right y axis) over time aggregated for all DJIA 30 companies (time labels have MM/DD/YYYY format).

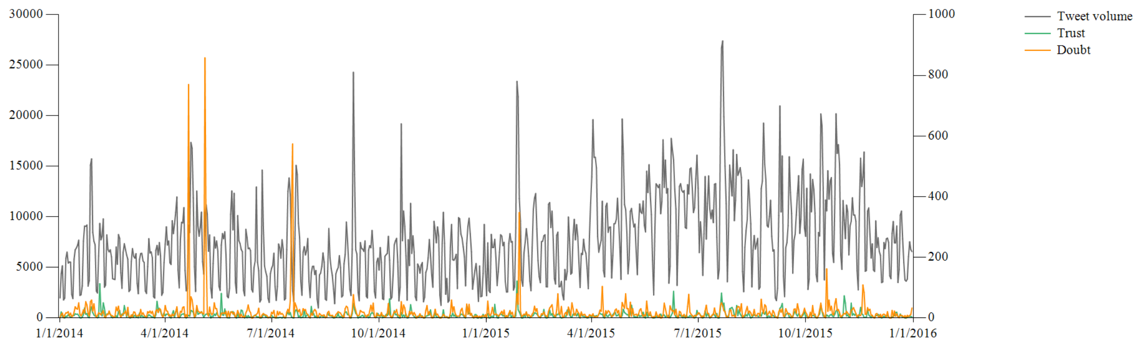


Figure 3: Tweet volume (gray, left y axis), trust (green, right y axis) and doubt (orange, right y axis) over time aggregated for all DJIA 30 companies (time labels have MM/DD/YYYY format). Both modifications of the methodology (handling negations and removing trust terms related to the investment funds) are taken into account.

etc.) immediately before the trust/doubt term under consideration. We treated such negated trust terms as doubt terms, and vice versa. This adjustment had only a small influence on the overall results: only 1.49% of the matched trust terms and 3.15% of doubt terms were found to be negated, which caused also only small changes in the visualization of the aggregated DJIA 30 results.

The second modification has, however, changed the results considerably. In this scenario, in order to avoid trust terms related to the context of investment funds, we did not take into account the word "Trust" if it appeared in the capitalized form, unless it was at the beginning of a tweet or positioned after a selection of punctuation marks ("."/"!"/"?"). By applying such an approach, we excluded 20,024 capitalized words "Trust" out of the 21,841 words *trust* (regardless of the capitalization) detected in the DJIA 30 tweets.

The aggregated results for DJIA 30 companies, after applying both modifications (handling negations and removing trust terms related to the investment funds) are shown in Figure 3. From the comparison with Figure 2, it can be seen that the increase in trust in the year 2015 vanished, which indicates that in that time period the Twitter community discussed investment funds. Several peaks of the doubt score remained and we made a more detailed analysis of the tweets which contributed to these peaks. The analysis revealed that high increases in doubt might be to some extent explained by retweets of certain Twitter posts, written by specific analysts or journals. Additionally, in Figure 5, we show an example of results for an individual company, i.e. the Apple company. The vertical red and blue lines mark the 10-K and 10-Q filing dates enabling one to observe if there exist tweet volume, trust or doubt changes around such dates.

3.1.1. Peaks and Trends in Tweet Volume, Trust and Doubt

After the analysis of Twitter posts from the perspective of trust and doubt expressions, we analysed the presence of such expressions near specific business-reporting events. We examined if there is a connection between changes in tweet volume, trust or doubt tweets, and the filing dates of the periodic reports. Specifically, we examined if there exist local peaks or trends in tweet volume/trust/doubt around the filing dates of 60 10-K and 180 10-Q reports of DJIA 30 companies in years 2014 and 2015. In this experiment, both adjustments of the methodology for assessing trust and doubt (handling negations and trust terms related to the investment funds) were applied.

The results are shown in Table 1, where we display percentages of reports that coincide with peaks in tweet vol-

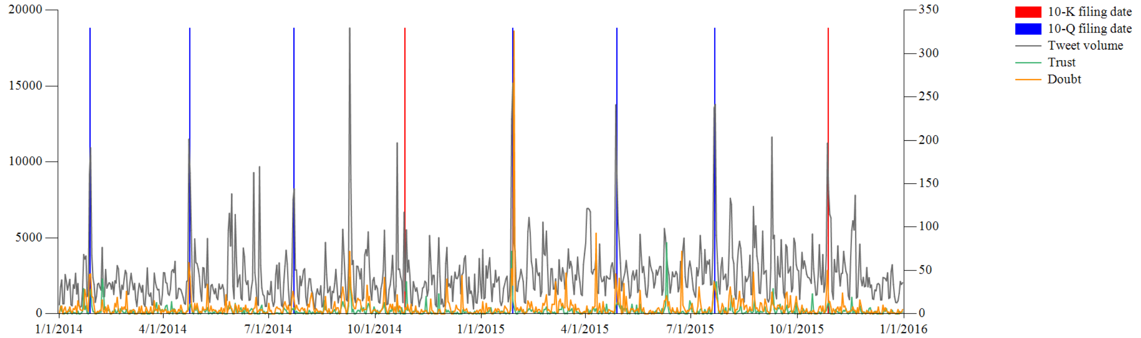


Figure 5: Tweet volume (gray, left y axis), trust (green, right y axis) and doubt (orange, right y axis) over time for the Apple company (time labels have MM/DD/YYYY format). The vertical red and blue lines mark the 10-K and 10-Q filing dates, respectively. Both modifications of the methodology (handling negations and trust terms related to the investment funds) are applied.

Random-All	Peak(-1)	Peak(0)	Peak(+1)	Inc. trend	Dec. trend
Volume	25.42%	27.92%	23.33%	17.92%	25.83%
Trust	7.50%	6.67%	7.08%	0.83%	1.25%
Doubt	10.00%	10.83%	7.50%	2.08%	1.25%
Report-All	Peak(-1)	Peak(0)	Peak(+1)	Inc. trend	Dec. trend
Volume	18.75%	44.17%	24.58%	15.00%	12.90%
Trust	9.58%	14.17%	10.83%	2.50%	2.50%
Doubt	12.08%	12.91%	11.67%	2.92%	2.50%
10-K	Peak(-1)	Peak(0)	Peak(+1)	Inc. trend	Dec. trend
Volume	23.33%	38.33%	18.33%	20.00%	23.30%
Trust	6.67%	8.33%	6.67%	0.00%	0.00%
Doubt	11.67%	16.67%	10.00%	0.00%	0.00%
10-Q	Peak(-1)	Peak(0)	Peak(+1)	Inc. trend	Dec. trend
Volume	17.78%	47.78%	27.22%	19.44%	13.89%
Trust	11.67%	16.67%	12.22%	3.33%	3.33%
Doubt	12.22%	12.78%	12.78%	5.00%	3.89%

Table 1: Percentages of random dates and filing dates of all reports, 10-K reports and 10-Q reports related to peaks and trends of tweet volume/trust/doubt around their filing dates. See Figure 4 for illustration of different types of peaks and trends.

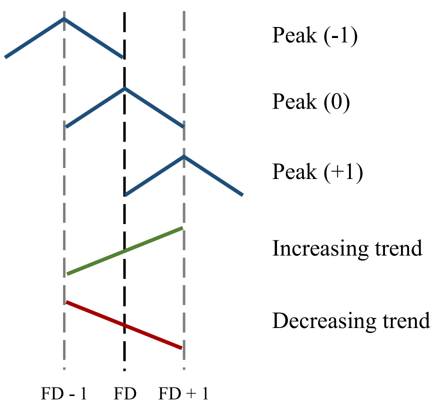


Figure 4: Illustration of changes and their labels in tweet volume, trust or doubt around a filing date (FD), 1 day preceding (FD-1) and following (FD+1) the filing date.

ume, trust and doubt on the exact day (Peak (0)), one day before (Peak (-1)), and one day after (Peak (+1)) filing the reports. Additionally, the table displays results of detecting increasing and decreasing trends in tweet volume/trust/doubt in a time period of 3 days around the filing dates (the filing date, and 1 day preceding and following the filing date). See Figure 4 for illustration of changes in tweet volume/trust/doubt and corresponding labels. Note that we did not apply thresholding or quantification, so for example a peak is equally detected given tweet frequencies (2,10,2) or (10,11,10) for three consecutive days.

The results in Table 1 are presented in terms of percentages of all reports, 10-K and 10-Q reports related to the described peaks and trends, while the first triplet of rows shows percentages for the same number of days as for all reports, but chosen at random dates. As it can be seen from the results, the largest percentage is observed in tweet volume on the exact day when the periodic reports are filed (and less on the preceding or the following day). This is ev-

Random-All	Peak(-1)	Peak(0)	Peak(+1)
Volume	-0.0046 (0.5043)	0.0047 (0.4954)	-0.0069 (0.3255)
Trust	-0.0048 (0.4967)	-0.0065 (0.3506)	-0.0013 (0.8514)
Doubt	-0.0023 (0.7412)	0.0040 (0.5652)	-0.0071 (0.3078)
Reports-All	Peak(-1)	Peak(0)	Peak(+1)
Volume	**-0.0153 (0.0285)**	**0.0495 (1.5e-12)**	-0.0005 (0.9438)
Trust	0.0092 (0.1899)	**0.0283 (5.3e-05)**	**0.0143 (0.0403)**
Doubt	0.0121 (0.0846)	**0.0152 (0.0298)**	0.0104 (0.1361)

Table 2: Correlation among the days with peaks of volume, trust and doubt and the reporting days. The values presented are Pearson correlation coefficients (and p-values) for tweet peaks and report dates (data was concatenated for all of the reports and studied companies). Significant results ($p<0.05$) are marked in bold.

ident also for both types of reports (10-K and 10-Q), however it seems that the Twitter community reacts more often to 10-Q than 10-K reports. Furthermore, it seems that the increasing and decreasing trends around the filing dates of periodic reports may be observed in tweet volume, but very rarely in trust or doubt.

For a comparison, we have also calculated the results for random dates (see top rows in Table 1). We took the same number of random dates as for the joint report dates (so the values of *Random-All* are directly comparable with the the *Report-All* values). We can notice that the reporting dates with peaks are higher in all the categories, with the exception of the peak volume on the preceding day (Peak(-1)). The largest difference (16.25%) is in the peak of volume on the reporting day. The results for trends are less consistent as they occur very rarely.

Next, to verify whether the peaks in volume, trust and doubt appear more often near report filing dates, we compared the actual dates and random dates with regard to their correlation with the peaks in the three phenomena of interest. Results of this analysis are presented in Table 2 and they suggest that there is a non-coincidental correlation among the report filing dates (Peak(0)) and the volume, trust and doubt peaks in tweets.

3.2. Annual Reports

Our analysis of trust and doubt terms appearance in tweets was mostly focused on behaviour during publishing dates of periodic reports, so we briefly analysed also the use of these terms in the reports. For this purpose we have collected the 10-K annual reports for the firms in DJIA 30. We selected the reports corresponding to the years of Twitter collection, i.e. the reports with filing dates in years 2014 and 2015. For the reports we have selected only the Part I, and Items 7 and 7A from Part II. These are less regulated textual parts of reports, allowing for more flexibility and expression of opinion.

We were interested in finding out, which terms from trust and doubt wordlists are more frequently used, and which companies use more of trust/doubt terms. For the list of terms from the trust and doubt wordlists, we have calculated the number of occurrences per firm and in total, including the relative and absolute frequencies.[7] The results of the absolute frequencies of trust and doubt terms joint for all companies are presented in Table 3 showing that the most used term from the Trust wordlist is *assurance*, followed by terms *confidence, trust* and *reliable*. For the terms from the Doubt wordlist, *uncertainty* takes the lead, while term *doubt* has only two occurences in the corpus.

To find out which firms use the largest amount of the trust and doubt terms, we calculated the relative frequencies of the trust and doubt terms (per 1000 words). The trust terms are the most frequently used in the reports of JPM (1.486 permille), followed by CAT, IBM and AAPL. Less than one permille of trust terms characterizes the reports of CVX and XOM, which are both from Oil and Gas industry.

The doubt terms are more frequent in the reports of TRV (0.832 permille), UNH, CSCO and AAPL, while interestingly, IBM and JPM together with WMT are the ones that express doubt (by using doubt terms) the least.

[7]For the lemmatization, we used the LemmaGen lemmatizer (Jursic et al., 2010).

Trust terms	freq.	Doubt terms	freq.
assurance	331	uncertainty	603
confidence	134	question	25
trust	111	doubtful	19
reliable	68	suspicious	16
reliance	46	tentative	3
reliant	8	doubt	2
faith	2		
TOTAL	700	TOTAL	668

Table 3: Trust and doubt terms (lemmas) and their absolute frequencies in the corpus of cleaned annual reports.

4. Conclusion

The work presented in this paper discusses the expression of trust and doubt in financial tweets and periodic reports corresponding to companies from the DJIA 30 Index in years 2014 and 2015. We use a wordlist-based approach to categorize the tweets into trust and doubt categories and analyze them from the perspective of interactions between social media posts and selected business reporting events. We analyzed the relationship between changes in tweet volume, trust or doubt, and the filing dates of the annual and quarterly reports. Results show that the Twitter users react mostly on the same day a report is filed, which was further confirmed by a comparison with random dates. Finally, our results indicate that the increasing and decreasing trends around the filing dates may be observed in tweet volume, but almost never in expressions of trust or doubt.

For the annual reports we selected the parts, where management has more freedom to express their opinion, and calculated the frequencies of the words from the trust/doubt wordlists. We showed that the most used term from the trust list is *assurance*, possibly an ambiguous term, and the most used term from doubt terms is *uncertainty*. We have also analyzed which firms use the trust and doubt terms more than others, and showed that JPM and IBM are between the companies that use the trust terms the most and the doubt terms the least. In future work we will correlate the usage of trust and doubt terms with financial performance information and see if trust and doubt expressions are correlated with financial indicators on the dataset of companies of DJIA 30, which would confirm our findings from the preliminary study of the correlation between content of annual reports and firms' financial performance (Smailović et al., 2017), where doubt terms showed significant correlations. Moreover, we plan to focus on an in depth corpus analysis of the terms from the wordlists, and extract and analyze their collocations, change over time and their original textual context.

5. Acknowledgements

The authors acknowledge the financial support from the Slovenian Research Agency for research core funding no. P2-0103, and project no. J5-7387 (Influence of formal and informal corporate communications on capital markets). We would like to thank the collaborators of the project A. Valentinčič, M. Pahor and I. Lončarski. This work was also supported in part by the H2020 FET project DOLFINS (grant no. 640772). We thank Sowa Labs (`http://www.sowalabs.com/`) and S. Rutar for providing data on financial tweets, as well as the anonymous reviewers for their valuable suggestions.

6. Bibliographical References

Adali, S., Escriva, R., Goldberg, M. K., Hayvanovych, M., Magdon-Ismail, M., Szymanski, B. K., Wallace, W. A., and Williams, G. (2010). Measuring behavioral trust in social networks. In *IEEE International Conference on Intelligence and Security Informatics*, pages 150–152.

Aiezza, M. C. (2015). "We may face the risks"..."risks that could adversely affect our face." A corpus-assisted discourse analysis of modality markers in csr reports. *Studies in Communication Sciences*, 15(1):68–76.

Bollen, J., Mao, H., and Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8.

Borondo, J., Morales, A., Losada, J., and Benito, R. (2012). Characterizing and modeling an electoral campaign in the context of Twitter: 2011 Spanish presidential election as a case study. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 22(2).

El-Haj, M., Rayson, P., Young, S., Moore, A., Walker, M., Schleicher, T., and Athanasakou, V. (2016). Learning tone and attribution for financial text mining. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016.*

Eom, Y.-H., Puliga, M., Smailović, J., Mozetič, I., and Caldarelli, G. (2015). Twitter-based analysis of the dynamics of collective attention to political parties. *PLoS ONE*, 10(7):e0131184.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Fuoli, M. (2017). Building a trustworthy corporate identity: A corpus-based analysis of stance in annual and corporate social responsibility reports. *Applied Linguistics*, page amw058.

Gabrovšek, P., Aleksovski, D., Mozetič, I., and Grčar, M. (2017). Twitter sentiment around the earnings announcement events. *PLoS ONE*, 12(2):e0173151.

Gerber, M. S. (2014). Predicting crime using Twitter and kernel density estimation. *Decision Support Systems*, 61:115–125.

Goel, S. and Uzuner, O. (2016). Do sentiments matter in fraud detection? Estimating semantic orientation of annual reports. *Intelligent Systems in Accounting, Finance and Management*, 23(3):215–239.

Grčar, M., Cherepnalkoski, D., Mozetič, I., and Novak, P. K. (2017). Stance and influence of Twitter users regarding the Brexit referendum. *Computational Social Networks*, 4(1):6.

Hajek, P., Olej, V., and Myskova, R. (2014). Forecasting corporate financial performance using sentiment in annual reports for stakeholders' decision-making. *Technological and Economic Development of Economy*, 20(4):721–738.

IASB. (2015). International Accounting Standards Board: Conceptual Framework. `http://www.ifrs.org/`. Accessed: 2015-07-10.

Jian, J.-Y., Bisantz, A., and Drury, C. (2000). Foundations for an empirically determined scale of trust in automated systems. 4:53–71, 03.

Jursic, M., Mozetic, I., Erjavec, T., and Lavrac, N. (2010). Lemmagen: Multilingual lemmatisation with induced ripple-down rules. *J. UCS*, 16(9):1190–1214.

Kothari, S., Li, X., and Short, J. E. (2009). The effect of disclosures by management, analysts, and business press on cost of capital, return volatility, and analyst forecasts: A study using content analysis. *The Accounting Review*, 84(5):1639–1670.

Loughran, T. and McDonald, B. (2016). Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*, 54(4):1187–1230.

Merkl Davies, D. M., Brennan, N. M., and McLeay, S. J. (2011). Impression management and retrospective sense making in corporate narratives: A social psychology perspective. *Auditing & Accountability Journal*, 24(3):315–344.

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41.

Paul, M. J. and Dredze, M. (2011). You are what you tweet: analyzing Twitter for public health. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, volume 20, pages 265–272.

Qiu, X. Y., Srinivasan, P., and Street, N. (2006). Exploring the forecasting potential of company annual reports. *Proceedings of the American Society for Information Science and Technology*, 43(1):1–15.

Ranco, G., Aleksovski, D., Caldarelli, G., Grčar, M., and Mozetič, I. (2015). The effects of Twitter sentiment on stock price returns. *PLoS ONE*, 10(9):e0138441.

Sherchan, W., Nepal, S., and Paris, C. (2013). A survey of trust in social networks. *ACM Computing Surveys (CSUR)*, 45(4).

Smailović, J., Grčar, M., Lavrač, N., and Žnidaršič, M. (2014). Stream-based active learning for sentiment analysis in the financial domain. *Information Sciences*, 285:181–203.

Smailović, J., , Žnidaršič, M., Valentinčič, A., Lončarski, I., Pahor, M., Martins, P. T., and Pollak, S. (2017). Automatic analysis of financial annual reports: A case study. *Computación y Sistemas (Special Issue on Advances in Human Language Technologies)*, 21(4):809–818.

Sprenger, T. O., Tumasjan, A., Sandner, P. G., and Welpe, I. M. (2014). Tweets and trades: the information content of stock microblogs. *European Financial Management*, 20(5):926–957.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., and Welpe, I. M. (2010). Predicting elections with Twitter: what 140 characters reveal about political sentiment. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*, volume 10, pages 178–185.

Zhao, L., Hua, T., Lu, C.-T., and Chen, I.-R. (2016). A topic-focused trust model for Twitter. *Computer Communications*, 76:1–11.