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 bold-faced 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)

sentence

Worldwide sales for **Fisher-Price Friends** were down **13%** for **the quarter**

representation

QUANTITY	'sales'
UNIT	%
VALUE	'13'
SIGN	−
MANNER	`relative`
THEME	'Fischer-Price Friends'
LOCATION	'worldwide'
TIME	'the quarter'

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.

'$'), VALUE (e.g. '2 million'), MANNER (`absolute` or `relative`), and SIGN ($\pm$). 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 Online 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.