Same Domain Different Discourse Style
A Case Study on Language Resources for Data-driven Machine Translation

Monica Gavrila, Walther v. Hahn, Cristina Vertan

University of Hamburg
Hamburg, Germany
{gavrila, vhahn}@informatik.uni-hamburg.de, cristina.vertan@uni-hamburg.de

Abstract

Data-driven machine translation (MT) approaches became very popular during last years, especially for language pairs for which it is difficult to find specialists to develop transfer rules. Statistical (SMT) or example-based (EBMT) systems can provide reasonable translation quality for assimilation purposes, as long as a large amount of training data is available. Especially SMT systems rely on parallel aligned corpora which have to be statistical relevant for the given language pair. The construction of large domain specific parallel corpora is time- and cost-consuming; the current practice relies on one or two big such corpora per language pair. Recent developed strategies ensure certain portability to other domains through specialized lexicons or small domain specific corpora. In this paper we discuss the influence of different discourse styles on statistical machine translation systems. We investigate how a pure SMT performs when training and test data belong to same domain but the discourse style varies.

Keywords: Statistical machine translation, Discourse style, evaluation of Machine Translation, Linguistic analysis of training data, Moses.

1. Introduction

Data-driven machine translation (MT) approaches became very popular during last years, especially for language pairs for which it is difficult to find specialists to develop transfer rules. Statistical (SMT) or example-based (EBMT) systems can provide reasonable translation quality for assimilation purposes, as long as a large amount of training data is available. Especially SMT systems rely on parallel aligned corpora which have to be statistical relevant for the given language pair. Given the intrinsic features of natural language as ambiguity, vagueness and polysemy parallel corpora have to be domain dependent. Within one domain, words tend to have with higher probability a certain meaning and therefore the disambiguation process in the automatic alignment step is more precise. The construction of large domain specific parallel corpora is time- and cost-consuming; the current practice relies on one or two big such corpora per language pair. Recent developed strategies ensure certain portability to other domains through specialized lexicons or small domain specific corpora.

For European languages two parallel corpora are largely used: JRC-Acquis[1] (parallel corpora for all combinations of 23 languages) and Europarl[2], which focuses in its last version on 21 EU[3]-Languages. Recently corpora with moderate size were added in Europarl, involving some of the languages from the countries that joined the community after EU-Enlargements in 2004 and 2007.

While the portability to other domains received a lot of attention in the recent years e.g. in (Niehues and Waibel, 2010), less research was performed to analyze how discourse style within same domain may affect the translation quality. In (Calude., 2002) the behavior of a rule-based system is tested across several text genres. However the chosen text genres are quite different (news, scientific, novel). Experiments across different genres have also been presented in (Monz, 2011). This paper showed that a phrase-based baseline system can benefit from using POS information by building lexically anchored local models. Test data of different genres have been also used in (Habash and Sadat, 2006): one is a mix of news, editorials and speeches, whereas the second, like the training data, is purely news. However. the focus of the paper is on Arabic preprocessing schemes for statistical MT.

In this paper we discuss the influence of different discourse styles on statistical machine translation systems. We investigate how a pure SMT performs when training and test data belong to same domain but the discourse style varies.

The paper is organized as follows: in section two we describe the experimental set-up, the data we used and the parameter setting for the SMT system. We introduce the broader framework within we tested the system as rationale for broader discourse style variability. We include also a linguistic analysis of the used corpora. Sections three is dedicated to the presentation of the evaluation results and discuss the variation of automatic metrics. Finally we present conclusions and further work in section four.

2. Experimental Set-up

2.1. The ATLAS -System

Current content management systems (CMSs) do not embed advanced techniques from language technology and information retrieval. The ICT PSP EU project ATLAS - Applied Technology for Language-Aided CMS[4] - aims to fill this gap by providing three innovative Web services within a web content management system (WCMS). The Web services: i-Librarian, EUDocLib and i-Publisher are not only

[1] http://optima.jrc.it/Acquis/.
[2] http://www.statmt.org/europarl/.
[3] EU = European Union
[4] http://www.atlasproject.eu

thematically different but offer also different levels of intelligent information processing.
The ATLAS WCMS makes use of state-of-the art text technology methods in order to extract information and cluster documents according to a given hierarchy. A text summarization module and a machine translation engine as well as a cross-lingual semantic search engine are embedded. The system is addressing for the moment seven languages (Bulgarian, Croatian, English, German, Greek, Polish and Romanian) from four language families. However, the chosen framework allows additions of other languages at a later point.
Machine translation (MT) is a key component of the ATLAS WCMS, and it will be embedded in all three services of the system. The development of the engine is particular challenging as the translation should be used in different domains and on different text-genres. Additionally the considered language-pairs belong most of them to the less resourced group, for which bilingual training and test material is available in limited amount.
The machine translation engine is integrated in two distinct ways into the ATLAS platform:

1. for i-Publisher (meta service for generating web sites) the MT is serving as a translation aid tool for publishing multilingual content. Text is submitted to the translation engine and the result is subject to the human post processing

2. for i-Librarian and EuDocLib (on-line content management systems generated with i-Publisher) the MT-engine provides a translation for assimilation, which means that the user retrieving documents in different languages will use the engine in order to get a clue about the documents, and decide if he will store them. If the translation is considered as acceptable it will be stored into a database.

The integration of an MT engine into a web based content management system in general and the ATLAS system in particular, presents from the user point of view two main challenges:

1. the user may retrieve documents from different domains. Domain adaptability is a major issue in machine translation, and in particular in corpus-based methods. Poor lexical coverage and false disambiguation are the main issues when translating documents out of the training domain

2. the user may retrieve documents from various time periods. As language changes over time, language technology tools developed for the modern languages do not work, or perform with higher error rate on diachronic documents.

With the current available technology it is not possible to provide a translation system which is domain and language variation independent and works for a couple of heterogeneous language pairs. Therefore our approach envisage a system of user guidance, so that the availability and the foreseen system-performance is transparent at any time.

Given the fact that the ATLAS platform deals with languages from different language families, and that the engine should support at least several domains an interlingua approach is not suitable. Building transfer systems for all language pairs is also time consuming and does not make the platform easily portable to other languages. Given the user and system requirements corpus based MT-paradigms are the only ones to be considered.
For the MT-Engine of the ATLAS -System we decided on a hybrid architecture combining EBMT (Gavrila, 2011) and SMT at word-based level (no syntactic trees will be used) (Koehn et al., 2007)). For the SMT-component part-of-speech (PoS) and domain factored models as in (Niehues and Waibel, 2010) are used, in order to ensure domain adaptability. An original approach of our system is the interaction of the MT-engine with other modules of the system:
The document categorization module assigns to each document one or more domains. For each domain the system administrator has the possibility to store information regarding the availability of a correspondent specific training corpus. If no specific trained model for the respective domain exists, the user is provided with a warning, telling that the translation may be inadequate with respect to the lexical coverage.
The output of the summarization module is processed in such way that ellipses and anaphora are omitted, and lexical material is adapted to the training corpus.
The information extraction module is providing information about meta-data of the document including publication age. For documents previous to 1900 we will not provide translation, explaining the user that in absence of a training corpus the translation may be misleading.
The domain and dating restrictions can be changed at any time by the system administrator when an adequate training model is provides. The described architecture is presented in Figure 1.

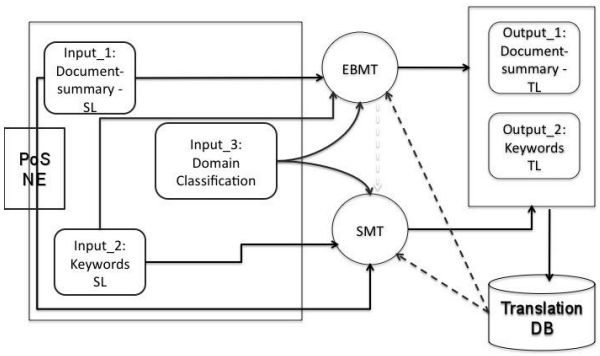

Figure 1: System architecture for the ATLAS-engine.

The design of the system was preceded by a study of portability of results among domains and discourse genres. Especially the latter aspect plays a major role within EuDocLib where documents relevant to the European Union have

to be processed. This involves: parliamentary speeches, laws or news and normal regulation. As for most part of the involved languages JRC-Acquis is the only available large parallel corpora within the law domain, we investigated to which extent documents within same domain but with different discourse structure can be processed by the translation engine

2.2. The Machine Translation Engine

The SMT system follows the description of the baseline architecture given for the EMNLP 2011 Sixth Workshop on SMT[5]. The system uses Moses[6], an SMT system that allows the user to automatically train translation models for the language pair needed, considering that the user has the necessary parallel-aligned corpus. More details about Moses can be found in (Koehn et al., 2007).
While running Moses, we used SRILM (Stolcke, 2002) for building the language model (LM) and GIZA++ (Och and Ney, 2003) for obtaining word alignment information. The training data has been 'cleaned'[7], in the sense of removing all sentences longer than 40 tokens[8]
The tuning step was realized using two different data-types:

1. 1500 sentences from JRC-Acquis (same domain and discourse type as the training data);
2. 1500 sentences from the EU-Constitution corpus[9] (same domain as the training data, but discourse style identical with the test data). The Eu-Constitution corpus (EUconst) is a parallel corpus collected from the European Constitution, which includes 21 languages (Tiedemann, 2009).

2.3. Test and Training Data

The training data is part of the JRC-Acquis corpus for German - English. JRC-Acquis is a freely available parallel corpus in 22 languages, which consists of European Union documents of legal nature. From the two types of sentence alignments available (Vanilla and HunAlign[10]), we used the Vanilla[11] alignments. The same alignments have been also used in (Ignat, 2009). In order to reduce possible errors, only one-to-one alignments have been considered for the experiments presented in this paper. More details on the JRC-Acquis corpus can be found in (Steinberger et al., 2006). The corpus is not (manually) corrected. Therefore, translation, alignment or spelling errors can have an influence on the output quality. The corpus contains 1190025 sentences containing 236965 unique tokens.
We used test data from three different corpora:

- JRC-Acquis itself (**Case A**) - see Section 2.4.1..
- EU-Constitution (**Case B**) - see Section 2.4.2..
- Europarl (**Case C**) - see Section 2.4.3.. More details on the corpus can be found in (Koehn, 2005).

The test and tuning data has been extracted from the middle of each of the corpora used. More information about the training, tuning and test data is presented in Table 1[12].

Data	No. sent.	Voc. size	Average sent. length
Training JRC-Acquis	1,190,025	236,965	9.03
Test JRC-Acquis	2,000	4,044	23.16
Test EUconst	2,000	4,329	19.19
Test Europarl	2,000	8,034	27.41
Tuning JRC-Acquis	1,500	4,308	29.10
Tuning EUconst	1,500	3,699	17.17

Table 1: Data description (No.=number, Voc.=vocabulary, sent.=sentence)

2.4. Linguistic Analysis of the Input Data

In our experiments we considered German as source language (SL) and English as the target one (TL).
In the following paragraphs we describe in detail morphological and syntactic features of the three corpora. The focus is set on the source language.

2.4.1. JRC-Acquis

From the morphological point of view, we remark a strong nominal style, interspersed capitalizing and the presence of rare words (e.g. "jeglichen") There is a high density of compounds and long compounds (up to 4 (derivated) nouns) are quite frequent, such as "Erstattungsbeträge", "Durchführungsvorschriften", etc. Quite difficult for MT systems is the presence of spontaneous composition (such as "Weltzuckermarkt") and of compounds with similar elements which are difficult to decompose ("Ausfuhrerstattungshöchstbetrag"). Semantic decomposition ("x of y by z") often fails because of unclear semantics: "(((Ausfuhr)erstattung)s((höchst)(betrag)))". Some unusual plurals with botanic terms are also present.
From the point of view of syntax following features are relevant for the texts being part of the corpus:

- Many bracketed sections with references to laws, named entities, and numbers;
- Calendar date indications;
- A strong dependency from the case of the prepositions, caused by the high number of prepositional structures (prepositional objects, adjuncts and complements);
- Asyntactic headlines ("Artikel 2");

[5] www.statmt.org/wmt11/baseline.html - last accessed on July 14th, 2011.
[6] www.statmt.org/moses/ - last accessed on July 14th, 2011.
[7] For more information about the 'cleaning' step please see the MT system description on www.statmt.org/wmt11/baseline.html.
[8] A token can be a lexical item, a punctuation sign, a number, etc.
[9] http://opus.lingfil.uu.se/EUconst.php.
[10] http://mokk.bme.hu/resources/hunalign/.
[11] See http://nl.ijs.si/telri/Vanilla/ - last accessed January 12th, 2011.
[12] The third column "Vocabulary size" counts the number of different tokens in the data.

- Long distance dependencies, partly with unclear structure;
  Example: ”DIE KOMMISSION DER EUROPÄISCHEN GEMEINSCHAFTEN gestützt auf den Vertrag ..., gestützt auf die Verordnung ...in Erwägung nachstehender Gründe: (1) - (10) HAT FOLGENDE VERORDNUNG ERLASSEN:”[13]
  The whole phrase/paragraph has a length of 50 lines with 30 intermediate full stops, 531 words (431 words without the last part after the colon) and 15 subphrases in a frame of 4 syntactic phrases. In the first long phrase the subj-NP consists of several complete sentences. A formal analysis of the whole phrase from full stop to full stop is impossible, because the VP follows only in the end of the paragraph after 14 interposed moderately syntactic structures. The recognition of the internal structure is rather difficult.

- Asyntactic phrases (partly ellipses);

- Foreign language segments, partly with foreign characters;

- Officialese syntax (e.g. ”Sind diese Bedingungen nicht erfüllt,” instead of: ”wenn diese ...” or ”Um zu überprüfen”);

- Ambiguity even in officialese syntax;

- Typing errors.

2.4.2. EUconst

Although quite close to the topic of JRC-Acquis the morphological features of this corpus are quite different. There are few technical terms except legal terms. Compared to JRC-Acquis there are much less and shorter compounds, few spontaneous compounds and more internationalisms (e.g. ”Kommission”, ”Parlament”, ”Union” etc., but with national inflectional morphology).
Also the syntactical analysis reveals new phenomena:

- Largely syntactic phrases, partly excessively structured;

- Passive style;

- Numerous modal verbs: ”sollen”, ”sollten”, ”können”, dürfen”;

- Significantly more enumerations;

- Highly structured text by headings and numbered paragraphs;

- Excessive hierarchical nested structures of conjunctions or disjunctions;

- Predominance of relative clauses against conjunctional clauses.

2.4.3. Europarl

The morphological features in Europarl are quite different form JRC-Acquis and EUconst, as the corpus records spontaneous speech.
Following features are worth to be mentioned:

- Meta signs in brackets;

- Persons are addressed by their names;

- Idioms with deviant inflexion;

- Multi-word expressions;

- Foreign language elements or mixed language expressions (i.e. ”Naming-and-Shaming-Verfahren”);

- Moderate number of officialese compounds;

- Frequently demonstrative pronouns instead of articles;

- Enumerations in the text are always written verbally (”erstens”), no numbers as in the previous corpora;

- Metaphoric expressions and idioms prevent literal translations (segment by segment): ”Finanzvehikel”;

- Rare compounds from abbreviations: ”PPE-DE-Fraktion”.

While syntactically comparing Europarl with JRC-Acquis and EUconst, we could notice the use of direct questions and/or exclamations, of personal pronouns (”ich”,”mir” etc.) and of shorter phrases, due to spoken style. Other particularities which are worth mentioning are[14]:

- Strong pronominal coherency chains ([noun] -it - it - it);

- Syntactic parentheses and syntactic ambiguities;

- Longer asyntactic segments (ellipses);

- Many metalinguistic elements;

- Rhetorical idioms.

2.4.4. Linguistic comparison of the three corpora

We structure the comparison along three criteria:

1. Differences in the syntactic construction:

JRC-Acquis and EUconst are juridical texts; however the EU-Constitution German text follows more the German style and aesthetics of written texts, while JRC-Acquis has an impersonal style. With this respect some of the syntactical construction in EU-Constitution are not retrieved in the JRC-Acquis texts. Europarl has a spoken discourse style however with some particularities: sentences are less elliptic as in spontaneous speech, and are relatively short.
All of the three texts contain many metalinguistic segmentation signals. Headers, paragraphs, and enumerations are not syntactically framed but marked by line feeds.

[13] Capitals in the original

[14] List is not exhaustive.

2. Differences in the Lexicon:

With respect to the vocabulary, JRC-Acquis is full of terminology, one sentence containing several specific terms. Problematic are the big number of spontaneous domain specific compound words. These words are usually used once or twice in text i.e. they are not statistically relevant and usually a proper alignment is therefore missing. The terminology in Europarl is quite different from the one in JRC-Acquis and Eu-Constitution. It is parliamentary, interspersed with some technical terms of the subject, which is being discussed. Speakers always respect quick understandability by the listeners. We found only a moderate number of spontaneous compounds, but frequently personal addressing and personal statements.

3. Comparison of phrase length:

Regarding the phrase length, the following phenomena could be noticed:

- JRC-Acquis: 50 lines have 30 full stops, 5 logical, 19 syntactic phrases;
- EUConst: 50 lines have 16 full stops and 18 phrases;
- Europarl: 50 lines have 26 full stops and 21 logical and syntactic phrases.

3. Results and Discussion

We performed several experiments, all having the JRC-Acquis as training data. We considered German as source language (SL) and English as the target one (TL). The choice is motivated by the following reasons:

- the linguistic analysis of source language phenomena (availability of a native speaker);
- the availability for the given language pair of three different corpora belonging to same domain, but having different discourse style.

We evaluated our translations using two automatic evaluation metrics: BLEU and NIST. The choice of the metrics is motivated by the available resources (software) and the results reported in the literature. Due to lack of data and further translation possibilities, the comparison with only one reference translation is considered in these experiments.
Although criticized, BLEU (bilingual evaluation understudy) is the score mostly used in the last years for MT evaluation. It measures the number of n-grams, of different lengths, of the system output that appear in a set of reference translations. More details about BLEU can be found in (Papineni et al., 2002).
The NIST Score, described in (Doddington, 2002), is similar to the BLEU score in that it also uses n-gram co-occurrence precision. If BLEU considers a geometric mean of the n-gram precision, NIST calculates the arithmetic mean. Another difference is that n-gram precisions are weighted by the n-gram frequencies.
The obtained results are presented in Table 2.

Test Data	BLEU	NIST
Test 1: JRC-Acquis (tuning JRC-Acquis)	0.5325	8.9775
Test 2: EU-Constitution (tuning JRC-Acquis)	0.3626	7.0983
Test 3: Europarl (tuning JRC-Acquis)	0.1979	5.8594
Test 4: EU-Constitution (tuning EU-Constitution)	0.3712	7.3489
Test 5: JRC-Acquis (tuning EU-Constitution)	0.5070	9.0888

Table 2: BLEU and NIST scores for different test and tuning data-sets

From the above table it can observed that there is a strong variation of the BLEU score even when the discourse genre variation is not so large (**Test 1** vs. **Test 2**). A tuning with data from the new discourse brings only slight improvement (**Test 4**). The radical decrease of the BLEU score in **Test 3** corresponds to a complete different discourse style in the test corpus. The number of out-of-domain-words (6.82% in **Test 1**; 14,09% in **Test 2** and 18,48% in **Test 3**) are not directly proportional with the decrease of the BLEU score. This is a clear indication for deeper linguistic phenomena which influence the translation quality.
Out-of-vocabulary words (OOV-words) and sentences already included in the training data influence the evaluation results. An overview of such aspects is shown in Table 3.

Data	OOV-words	Sent. in the training data
Test JRC-Acquis	6.82	38.20
Test EUconst	14.09	20.70
Test Europarl	18.48	0.00
Tuning JRC-Acquis	4.99	38.67
Tuning EUconst	10.46	22.93

Table 3: Data analysis (The results are in %.)

OOV-words in the tuning data means new words are added to the training data. Sentences already in the tuning data means that part of the training data is doubled. The difference between the total number of sentences in the tuning data and the number of sentences already in the corpus is here more relevant, as this represents new information added to the training data.
The results of the automatic analysis are in-line with the differences in discourse style between the training and test data. However this experiment shows also the limitations of current automatic metrics. It s impossible to say which linguistic differences among the training and test data determine the decrease of BLEU score.

4. Conclusion and Further Work

Our experiments show that only considering test and training data within the same domain does not guarantee a system performance comparable with the golden standard.

More important is the choice of data with same discourse genre. We intend to realize a linguistic analysis of the translation results and to extend our experiments to the reverse translation direction. Further experiments should consider test data from different domain but similar discourse style and employment of domain adaptation methods.

5. Acknowledgements

The present work is part of the EU-Project ATLAS, supported through the ICT-PSP-Programme of the EU Commission for "Multilingual Web".

6. References

Andreea S. Calude. 2002. Machine translation of various text genres. In *Proceedings of the 7th Language and Society Conference of the New Zealand Linguistic Society*, page 12pp, Hamilton,New Zealand, November.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Monica Gavrila. 2011. Constrained recombination in an example-based machine translation system. In Vincent Vondeghinste Mikel L. Forcada, Heidi Depraetere, editor, *EAMT-2011: the 15th Annual Conference of the European Association for Machine Translation*, pages 193–200, Leuven, Belgium, May. EAMT. Accepted for oral presentation (Research Track). ISBN 9789081486118.

Nizar Habash and Fatiha Sadat. 2006. Arabic preprocessing schemes for statistical machine translation. In *HLT-NAACL 2006: Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL*, pages 49–52, New York, NY, USA, June.

Camelia Ignat. 2009. *Improving Statistical Alignment and Translation Using Highly Multilingual Corpora*. Ph.D. thesis, INSA - LGeco- LICIA, Strasbourg, France, June, 16th. It can be found on: http://sites.google.com/site/cameliaignat/home/phd-thesis - last accessed on 3.08.09.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session*, pages 177–180, Prague, Czech Republic, June.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the MT Summit*.

Christof Monz. 2011. Statistical machine translation with local language models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 869–879, Edinburgh, Scotland, UK, July 27-31.

Jan Niehues and Alex Waibel. 2010. Domain adaptation in statistical machine translation using factored translation models. In *Proceedings of EAMT 2010*, Saint-Raphael.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Session: Machine translation and evaluation*, pages 311 – 318, Philadelphia, Pennsylvania. Publisher: Association for Computational Linguistics Morristown, NJ, USA.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaz Erjavec, Dan Tufis, and Daniel Varga. 2006. The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)*, pages 2142–2147, Genoa, Italy, May, 24-16.

Andreas Stolcke. 2002. Srilm - an extensible language modeling toolkit. In *Proc. Intl. Conf. Spoken Language Processing*, pages 901–904, Denver, Colorado, September.

Jörg Tiedemann. 2009. News from opus - a collection of multilingual parallel corpora with tools and interfaces. In *Recent Advances in Natural Language Processing (vol V)*, pages 237–248, John Benjamins, Amsterdam/Philadelphia.