Text Genre - an Unexplored Parameter in Statistical Machine Translation

Monica Gavrila, Cristina Vertan

University of Hamburg
30 Vogt-Koelln Str, 22527, Hamburg, Germany

gavrila@informatik.uni-hamburg.de, cristina.vertan@uni-hamburg.de

Abstract

It is generally accepted that the performance of a statistical machine translation (SMT) system depends significantly on the concordance between the domain of training and test data. During the last years several methods have been proposed in order to deal with out-of-domain words. Less to no attention has been paid however to text genre within the same domain. In this paper we demonstrate that the style of the training corpus may influence the quality of the translation output even when the domain of the train- and test data remains almost unchanged, but the text genre changes. We use as training data the JRC-Acquis and as test data the Europarl corpus. We include also experiments with an out-of-domain test data, as comparison for the variation of performance of the SMT system.

Keywords: Statistical machine translation, Text genre, Europarl, JRC-Acquis, RoGER, SMT evaluation

1. Introduction

It is generally accepted that statistical machine translation (SMT) provides sufficiently good translation results with in-domain test data and “enough” training data. Results are rapidly decreasing for out-of-domain test data. Therefore, lot of research has been directed in the last years towards domain adaptation of SMT systems - e.g. (Nichus and Waibel 2010). Especially for European languages, current state-of-the-art SMT-engines are trained on one of the two large corpora available: JRC-Acquis or Europarl. Special techniques are applied in a second phase in order to ensure lexical domain adaptation. Less attention is paid to the fact that, even inside one domain, corpora belong to different text genres or, at least, have different discourse structures and, therefore, other types of syntactic structures or semantic frames. These differences may have a bigger influence on the quality of an SMT-system than assumed until now.

1.1. The Context for the Experiments

This aspect is of particular importance in scenarios where a machine translation engine is part of a complex architecture exposed to textual input from heterogeneous domains or text genres. This is the case of a Web-Content-Management System (WCMS) as the ATLAS-System\(^1\). In this system several web-services based on advanced language technology components are built for seven European languages. Among the key technologies which are incorporated, a central role is played by machine translation. Due to lack of enough training data for all possible domains, the data-driven translation engine is trained mostly on the JRC-Acquis corpus and afterwards domain adaptation is performed. For domains for which no training model is available, the user is informed that the translation quality can lack accuracy. As the acceptance of such system depends extensively from the user acceptance we decided to investigate also to which extent the text genre of the input can influence the translation quality.

<table>
<thead>
<tr>
<th>Direction of translation</th>
<th>Paper</th>
<th>BLEU score</th>
</tr>
</thead>
<tbody>
<tr>
<td>English-Romanian</td>
<td>(Cristea, 2009)</td>
<td>0.5464</td>
</tr>
<tr>
<td></td>
<td>(Ignat, 2009)</td>
<td>0.3208</td>
</tr>
<tr>
<td></td>
<td>(Koehn et al., 2009)</td>
<td>0.4900</td>
</tr>
<tr>
<td>Romanian-English</td>
<td>(Cristea, 2009)</td>
<td>0.4604</td>
</tr>
<tr>
<td></td>
<td>(Ignat, 2009)</td>
<td>0.3840</td>
</tr>
<tr>
<td></td>
<td>(Koehn et al., 2009)</td>
<td>0.6080</td>
</tr>
</tbody>
</table>

Table 1. Previous Reported Results.

Usually SMT experiments have been performed and presented with in-domain data, such as (Koehn et al., 2009) or (Ignat, 2009).

An overview of how (rule-based) machine translation (MT) reacts to various text genres is shown in (Calude, 2002), where the MT system used is SYSTRAN\(^2\). The study analyzed machine translated extracts from four text genres with respect to different linguistic errors. Best results were obtained for technical sets of instructions.

Our paper is organized in six sections. After this short introduction we will present our experimental settings: the MT system and the training and test data. In Section 3 we show the evaluation results, followed in Section 4 by presenting factors which influence the results. The paper presents the conclusions and further work in Section 5.

The last part of the paper shows our acknowledgements.

\(^1\) http://www.atlasproject.eu

\(^2\) http://www.systran.com/systran/net.
2. Experiments

In this section we present the SMT system and the training and test data used in the experiments.

2.1. The SMT System

The SMT system follows the description of the baseline architecture given for the EMNLP 2011 Sixth Workshop on SMT. The system uses Moses, 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. We made two changes to the specifications given at the Workshop on SMT: we left out the tuning step and we changed the order of the language model (LM) from 5 to 3. Leaving out the tuning step has been motivated by results we obtained in experiments which are not the topic of this paper, when comparing different settings for SMT: not all tests for the system configuration which included tuning showed improvement in the evaluation scores. Changing the LM order has been motivated by results reported in the SMART project, in which it has been concluded that 3-gram configurations provide best results -- see (Rousu, 2008).

2.2. Training Data

The training data is part of the JRC-Acquis corpus for English - Romanian. JRC-Acquis is a freely available parallel corpus in 22 languages, which consists of European Union documents of a legal nature. It is based on the Acquis Communautaire (AC), the total body of European Union (EU) law applicable the EU Member States. This collection of legislative text changes continuously and currently comprises selected texts written between the 1950s and now.

From the two types of sentence alignments available (Vanilla and HunAlign), we used the Vanilla alignments. The same alignments have been used in (Ignat, 2009). The sentence alignment is done at paragraph level, where a paragraph can be a simple or complex sentence, a sub-sentential phrase (e.g. noun phrase) or even more sentences. 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 has not been (manually) corrected. Therefore, translation, alignment or spelling errors can have an influence on the output quality.

For the SMT experiments, from 391324 links in 6557 documents, only 336509 links (the one-to-one alignments) have been considered. Due to the cleaning step of the SMT system, the number of one-to-one alignment links used for the LM was reduced to 240219 links for the Translation Model (TM). This represents 61.38% of the initial corpus. The average sentence length is around 14.5 tokens. In this paper token means a word, a number and a punctuation sign.

2.3. Test Data

We used test data from three different corpora:

- JRC-Acquis itself (Case A) – in-domain data;
- Europarl (Case B) – ‘similar’ data (in-domain, different genre – out-of-genre data);
- RoGER (Case C) – out-of-domain data.

The first two corpora could be considered in the same domain, as both refer to EU matters, but they are of a different genre: JRC-Acquis contains EU regulations; Europarl is extracted from the literal reports of the debates in the European Parliament. RoGER represents a totally different domain, as it contains text from a manual of an electronic device. The separation of these tests has been done by inspection and intuition.

2.3.1. A: JRC-Acquis

The tests were run on parts of the JRC-Acquis, which were not used for training. 897 sentences (three sets of 299 sentences – A: Test 1, A: Test 2, A: Test 3) were removed before the training step from the initial corpus, in order to be used as test data. Sentences were removed from different parts of the corpus to ensure a relevant lexical, syntactic and semantic coverage. A: Test 1+2+3 data set contains all the sentences. The test data has not been cleaned, this means that no length restriction is considered and sentences might be repeated. For example, the paragraph “Article NUMBER” repeats itself 53, 44 and 11 times in A: Test 1, A: Test 2 and A: Test 3, respectively. The data is in-domain data. The average sentence length is around 21 tokens.

2.3.2. B: Europarl

The Europarl parallel corpus (Koehn, 2005) is extracted from the proceedings of the European Parliament (the literal reports of the debates) dating back to 1996 and contains in its last version twenty-one languages.

We extracted from version 6 of the corpus three different test data sets, each of 299 sentences from the English-Romanian data. As for JRC-Acquis, we extracted the data from different parts of the corpus: from the beginning, middle and the end of the corpus. Small corrections have been done, as sometimes also sentences in other languages have been encountered. The test data sets from this corpus are: B: Test 1, B: Test 2, B: Test 3 and B: Test 1+2+3. The average sentence length is around 13 tokens. However, for B: Test 1 and 2 it is around 7.5 and for B: Test 3 it is 24.5.

The data is in-domain, but it has a different genre when compared with the training data: the structure and

---

7. The tag < p > from the initial HTML files.
8. In the Moses description, all sentences longer than forty tokens are excluded.
discourse of the text are totally different than the ones of the JRC-Acquis. The text refers to similar matters as the training data: European regulations. We consider these test data sets as ‘similar’ test data.

2.3.3. RoGER
In order to analyze the performance of SMT systems to a total different type of text input, we used the RoGER corpus. RoGER is a parallel corpus, aligned at sentence level. It is domain-restricted, as the texts are from a users’ manual of an electronic device. The languages included in the development of this corpus are Romanian, English, German and Russian. The corpus was manually compiled. It is not annotated and diacritics are ignored. The corpus was manually verified: the translations and the (sentence) alignments were manually corrected. More about the RoGER corpus can be found in (Gavrila and Elita, 2006). From the 2333 sentences, we extracted 300 sentences from the middle of the corpus and used them as test data (C: Test). The average sentence length is around 15 tokens. The data is totally out-of-domain.

3. Evaluation Results
We evaluated our translations using three automatic evaluation metrics: BLEU, NIST and TER. The choice of the metrics is motivated by the (linguistic) resources we had available 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.

TER calculates the minimum number of edits needed to get from obtained translations to the reference translations, normalized by the average length of the references. It considers insertions, deletions, substitutions of single words and an edit-operation which moves sequences of words. More information about TER can be found in (Snover et al., 2006).

The obtained evaluation results are presented in the Tables 2 and 3. A graphical representation is shown in Figure 1.

The results for in-domain data are similar to other BLEU scores published in the literature (with the exception of the test data set A: Test 1 for Romanian-English)\(^{10}\). The out-of-domain data provides quite low results. The results for ‘similar’ data, somehow surprisingly, are closer to the ones of the out-of-domain data.

\(^{10}\) A one-to-one comparison is not possible, as the training and test data are not the same.
the training data as lemma, but a specific word-form\(^{11}\) might be missing.

<table>
<thead>
<tr>
<th>Test Data</th>
<th>OOV-words</th>
<th>Sentences in the corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>A: Test 1</td>
<td>51 (4.10%)</td>
<td>69 (23.07%)</td>
</tr>
<tr>
<td>A: Test 2</td>
<td>7 (0.76%)</td>
<td>117 (39.13%)</td>
</tr>
<tr>
<td>A: Test 3</td>
<td>111 (8.19%)</td>
<td>81 (27.09%)</td>
</tr>
<tr>
<td>A: Test 1+2+3</td>
<td>169 (6.22%)</td>
<td>267 (29.76%)</td>
</tr>
<tr>
<td>B: Test 1</td>
<td>59 (11.32%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>B: Test 2</td>
<td>697 (33.88%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>B: Test 3</td>
<td>94 (15.40%)</td>
<td>2 (0.66%)</td>
</tr>
<tr>
<td>B: Test 1+2+3</td>
<td>837 (30.41%)</td>
<td>2 (0.22%)</td>
</tr>
<tr>
<td>C: Test</td>
<td>330 (37.67%)</td>
<td>0 (0%)</td>
</tr>
</tbody>
</table>

Table 4. Data Description (Romanian-English)

Comparing the OOV-words for Test 1+2+3 for Europarl and Test 1+2+3 for JRC-Acquis, we could conclude that these two sets of OOV-words are (almost) totally different: only three words for English-Romanian and two for Romanian-English are in common in these two sets of OOV-words.

As expected, for the translation direction Romanian-English, the highest number of OOV-words appear in data C (RoGER; out-of-domain data) data (37.67\%). However, for English-Romanian, Test 2 from Europarl (data B; ‘similar’ data) contains the highest number of OOV-words: 18.68\%. The out-of-domain data (data C) has only 14.65\% of OOV-words.

<table>
<thead>
<tr>
<th>Test Data</th>
<th>OOV-words</th>
<th>Sentences in the corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>A: Test 1</td>
<td>33 (3.15%)</td>
<td>69 (23.07%)</td>
</tr>
<tr>
<td>A: Test 2</td>
<td>2 (0.27%)</td>
<td>134 (44.81%)</td>
</tr>
<tr>
<td>A: Test 3</td>
<td>96 (8.64%)</td>
<td>85 (28.42%)</td>
</tr>
<tr>
<td>A: Test 1+2+3</td>
<td>131 (5.59%)</td>
<td>288 (21.10%)</td>
</tr>
<tr>
<td>B: Test 1</td>
<td>30 (7.5%)</td>
<td>21 (7.02%)</td>
</tr>
<tr>
<td>B: Test 2</td>
<td>288 (18.68%)</td>
<td>3 (1%)</td>
</tr>
<tr>
<td>B: Test 3</td>
<td>60 (11.62%)</td>
<td>22 (7.35%)</td>
</tr>
<tr>
<td>B: Test 1+2+3</td>
<td>366 (17.99%)</td>
<td>46 (5.12%)</td>
</tr>
<tr>
<td>C: Test</td>
<td>93 (14.65%)</td>
<td>0 (0%)</td>
</tr>
</tbody>
</table>

Table 5. Data Description (English-Romanian)

A better analysis of the OOV-words in different test data-sets should be made to have a more realistic overview. For example, it could be possible that in data B, due to the text genre, more declination or conjugation forms have been used, when compared with data A. Therefore, the use of a lemmatizer in the translation process could improve the translation results.

Concerning the number of test sentences already found in the training data, excluding in-domain data, more sentences have been found for English-Romanian and ‘similar’ data. For Romanian-English the results for this aspect is similar for both out-of-domain and ‘similar’ data: under 1%.

5. Conclusions

In this paper we showed several SMT experiments with different test data (in-domain vs. out-of-domain vs. ‘similar’ data) using the JRC-Acquis (English-Romanian) corpus for training. The results for in-domain and out-of-domain data are as expected. Somehow surprisingly, the results for ‘similar’ data are closer to the results for out-of-domain data. The differences in discourse and vocabulary lowered the translation scores for the Europarl tests, although we find ourselves in the same European framework as in the training data. This shows that having only ‘similar’ data for a specific domain, we cannot always expect good translation results. We can consider the conclusion of this paper limited to the data we used and only as a starting point for further analyses. A manual analysis of the translations should bring a better overview on the automatic scores and the sources of errors. Further experiments with various corpora and language pairs are needed before drawing a final (more general) conclusion.

6. Acknowledgements

Part of the work in this paper is part of the EU-Project ATLAS, supported through the ICT-PSP-Programme of the EU-Commission (Topic “Multilingual Web”) and the PhD research conducted by Monica Gavriţa at the University of Hamburg.

References


the Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, pp. 177-180, Prague, Czech Republic, June.


