SMT Experiments for Romanian and German Using JRC-ACQUIS

Monica Gavrila
Hamburg University
Faculty of Mathematics, Informatics and Natural Sciences
Vogt-Kölln Str. 30, 20251, Hamburg, Germany
gavrila@informatik.uni-hamburg.de

Abstract

One of the LT[1]-applications that ensures the access to the information, in the user's mother tongue, is machine translation (MT). Unfortunately less spoken languages - a category in which the Balkan and Slavic languages can be included - have to overcome a major gap in language resources, reference-systems and tools. In its simplest form, statistical machine translation (SMT) is based only on the existence of a big parallel corpus and therefore it seems to be a solution for these languages. In this paper the performance of a Moses-based SMT system, for Romanian and German, is investigated using test data from two different domains - legislation (JRC-ACQUIS) and a manual of an electronic device. The obtained results are compared with the ones given by the Google on-line translation tool. An analysis of the obtained translation results gives an overview of the main challenges and sources of errors in translation, in these experimental settings.

Keywords

SMT, Romanian,German, Moses, Google in-line translation tool

1 Introduction

"Less interesting languages"[2] have to overcome a major gap in language resources, reference-systems and tools which ensure the development of an MT-system of higher quality. In its simplest form, statistical machine translation (SMT) is based only on the existence of a big parallel corpus, thereby it seems to be a solution for this kind of languages.

From the currently available corpora for the languages considered in the description of the workshop, JRC-ACQUIS is used for the experiments described in this paper. The languages addressed are Romanian and German. The size of bilingual subsets of JRC-ACQUIS differs strongly from language pair to language pair, e.g. for English-German the size of the corpus is over 1 million sentences, for German-Romanian is less than 350000 sentences. Compared to EUROPARL or to the "News Corpus" used in recent investigations in the EUROMATRIX project [1], bilingual subsets in JRC-ACQUIS have approximately six times less aligned sentences, for the language-pair considered.

In this paper the performance of a simplistic Moses-based SMT-system, when trained and tested on JRC-ACQUIS (version 2.2), is investigated. For one of the test-set, data from a small technical corpus is used. The obtained results are compared with the ones given by the Google SMT on-line translation tool. The outcome shows that for less resourced languages - in this case Romanian - the development of further parallel corpora on broader domains and the improvement of the existing resources seem to be unavoidable. In the case of JRC-ACQUIS, for Romanian, such a step has already been done with JRC-ACQUIS Version 3[3].

The paper is organized as follows: in section 2 the used corpora are presented; sections 3 and 4 describe the experiments performed and their results. The last section concludes the presented results.

2 Data Description

For the experiments described in this paper, German-Romanian was chosen as language pair. The tests were done for both directions of translation.

Romanian is a less-resourced language with a highly inflected morphology and high demand for translation after joining the European Union. Compared to widely spoken languages, few resources and tools were developed for Romanian. An overview of tools for Romanian was made in the CLARIN Project (http://www.clarin.eu). Bilingual resources including Romanian are not so many and with few exceptions (see [11], [10]) relate only to English-Romanian.

Few parallel corpora are available, in which one of the language is Romanian, that have a "*satisfactory*" size, and that do not consider, as the other language, only English, e.g. JRC-ACQUIS, OPUS[4].

One of the reasons for using in this paper JRC-ACQUIS is the fact that, to the author's knowledge, all MT experiments, where Romanian was considered,

[1] LT = Language Technology

[2] In this paper, "*less interesting languages*" means less spoken - as number of people - and politically uninteresting

[3] This last version was not used for the experiments, because, as stated on http://langtech.jrc.it/JRC-Acquis.html, at the moment, for this new version for Romanian, alignment information is not available.

[4] For more details on OPUS please see [9] and http://urd.let.rug.nl/tiedeman/OPUS/

Multilingual Resources, Technologies and Evaluation for Central and Eastern European Languages 2009, Bulgaria, pages 14–18

are done using this corpus - see [4] and [2][5]. Although on-line or commercial translation tools for Romanian exist[6], they are all black-boxes.

2.1 Training Data

The training corpus is part of the JRC-ACQUIS (http://wt.jrc.it/lt/Acquis/ - last accessed on 18.04.09). Two types of alignments are available on the corpus homepage: Vanilla and HunAlign. The alignments realized with the Vanilla aligner[7] were used for the experiments presented here. Although not the best solution for MT, the alignment provided is done at paragraph-level. A *paragraph* can be a sentence, a sub-sentential phrase (e.g. noun phrase - NP), a phrase, or more sentences. This has an impact on the translation quality, as most of existing systems recommend sentence alignment.

In order to reduce the number of errors, only 1:1 paragraph alignments were considered for the experiments. This means that from 391972 links in 6558 documents, only 324448 links are used for the Language Model (LM). Due to the cleaning step of the SMT system, which limits the sentence length to 40 words[8], the number of 1:1 alignment links considered for the Language Model (LM) are reduced to 238172 links for the Translation Model (TM). This represents 61.38% of the initial corpus. More details on JRC-ACQUIS can be found in [8].

2.2 Test Data

The experiments were run on two different corpora: one is part of the JRC-ACQUIS corpus and the other is part of a technical manual of an electronic device.

897 sentences (299 from the beginning, 299 from the middle and 299 from the end) were removed from JRC-ACQUIS training data, in order to be used as test sets. Sentences were chosen from different parts of the corpus to ensure a relevant lexical, syntactic and semantic coverage. These 3 sets of 299 sentences represent **Test 1**, **Test 2** and **Test 3** of the experiments. As one of the goal of the experiments was to analyze the reaction of the evaluation scores to data-size, **Test 4** data-set contains all 897 sentences. In oder to see how the translation quality changes inside a corpus, several test-sets of the same size, from the same corpus, were chosen.

In order to evaluate the reaction of the SMT system to other input text type, the second test corpus was considered. It is extracted from a manual of an electronic device. It is sentence-aligned and the translation is manually verified. In the corpus dates, numbers and names were replaced by meta-words, e.g. numbers by NUM. Diacritics were not considered. From this corpus 300 sentences from the middle of the text were used as test data - **Test 5** in the experiments.

The detailed statistics on the data are presented in Table 1.

Corpus	No. of words	Vocabulary size	Average sentence length SL
SL = German			
Training			
JRC-Acquis	3256047	69260	13.6
Test Data			
Test 1	5325	1067	17,8
Test 2	10286	1380	34,4
Test 3	5125	1241	17,23
Test 4	20763	2860	23.14
Test 5	4549	715	15.1
SL = Romanian			
Training			
JRC-Acquis	3453584	48844	14.5
Test Data			
Test 1	5432	1198	18,16
Test 2	11488	1609	38,42
Test 3	5317	1298	17,7
Test 4	22237	3122	24.79
Test 5	4561	767	15.2

Table 1: *Corpora Statistics*

3 Experimental Settings

The SMT system used follows the description of the baseline system given for the EACL 2009 4th Workshop on SMT[9] and it is based on Moses[10] - see [5]. Wanting to see what results can be obtained by a very simple SMT, two parameters were changed: the tuning step is left out and the LM order is 3.

All test data-sets were translated with the Moses-based system and with the Google on-line translation tool[11]. In both cases, the same metrics were used for evaluation: BLEU and TER. For these experiments the use of other linguistic resources was avoided deliberately, in order to be able to evaluate the robustness of a pure SMT-System at domain change. When changing the domain it is expected that out-of-training-vocabulary words (*OOV-Words*) - especially in domain specific vocabulary - play a major role. In the following subsection this aspect is presented.

3.1 Out-of-training-vocabulary Words

The OOV-words were extracted, for both directions of translation, by comparing the training vocabulary and the test vocabulary for the source language (SL).

[5] The language-pair considered in this papers is Romanian-English. For the author was interesting to use the same corpus, as in previous work also Romanian-English experiments were run. These results were compared with the Romanian-German ones. The Romanian-English experiments are not part of the present paper.

[6] An overview of such systems can be found on www.euromatrix.net/euromatrix, last accessed on 17.06.2009.

[7] (http://nl.ijs.si/telri/Vanilla/ - last accessed 18.04.09.

[8] The sentence size limit is the one recommended for the EACL 2009 4th Workshop on SMT

[9] EACL 2009 Workshop on SMT: http://www.statmt.org/wmt09/index.html - last accessed on 18.04.09.

[10] http://www.statmt.org/moses/ (last accessed on 18.04.09)

[11] More on Google: http://translate.google.de/translate_t?hl=de# - last accessed on 08.05.09
See also [1]

As expected, the percentage of OOV-words, for the technical manual data-set, is higher - see Table 2. As seen in Section 4, the higher number of OOV-words leads to worse translation scores.

When manually analyzing the extracted words, it was noticed that, in the first corpus, due to segmentation and spelling errors and not-replacement of numbers, dates etc with meta-words, sometimes the extracted words are not correct, e.g. "*dreptulde*" (correct: "*dreptul de*" - English: "*the right of*"), or just symbols are extracted, e.g. "*2ev*", "*0155*", "***". After the removal of the wrong extracted words, the number of OOV-words for **Test 4** was reduced to almost 50% for Romanian-German and to 83% for German-Romanian. In Table 2 the number of OOV-words, after the removal procedure, are shown.

The words extracted for the second corpus were 99% right.

Corpus	No. of words	Percentage
SL = German		
Test 1	47	4.4%
Test 2	37	2.68 %
Test 3	185	14.9 %
Test 4	267	9.3%
Test 5	280	39.16%
SL = Romanian		
Test 1	17	1.41%
Test 2	20	1.24%
Test 3	82	6.36%
Test 4	130	4.16%
Test 5	279	36.327%

Table 2: *OOV-Words*

4 Experimental Results

In the experiments, due to the lack of multiple references, the comparison with only one reference translation is considered. The following metrics are used:

- BLEU (**bi**lingual **e**valuation **u**nderstudy) - The NIST/BLEU implementation, version 12 [12] is used. Although criticized, BLEU is 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 references. More details about BLEU can be found in [6]. As for previous developed systems BLEU is one of the evaluation metrics, for comparison reasons, it is still important to calculate it.
- TER (**t**ranslation **e**rror **r**ate)[13] - It calculates the minimum number of edits needed to get from a obtained translation 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 one can find in [7].

The obtained results are shown in **Table 3**, **Table 4** and **Table 5**.

Score	Test 1	Test 2	Test 3	Test 4
German - Romanian				
BLEU	**0.2955**	**0.4244**	**0.2884**	**0.3644**
TER	**0.6198**	**0.5905**	**0.6438**	**0.6112**
Romanian - German				
BLEU	0.2953	**0.4411**	0.2939	**0.3726**
TER	0.6437	**0.5588**	0.6791	**0.6112**

Table 3: *Evaluation Results for the SMT System for the JRC-ACQUIS Test Data*

Score	Test 1	Test 2	Test 3	Test 4
German - Romanian				
BLEU	0.2853	0.2809	0.274	0.2838
TER	0.6397	0.6707	0.6642	0.6612
Romanian - German				
BLEU	**0.3277**	0.3301	**0.3208**	0.3332
TER	**0.5971**	0.6590	**0.6576**	0.6425

Table 4: *Evaluation Results for the Google On-line Translation System for the JRC-ACQUIS Test Data*

Score	SMT	Google
German - Romanian		
BLEU	0.0192	**0.1041**
TER	0.9318	**0.836**
Romanian - German		
BLEU	0.0223	**0.2242**
TER	0.9358	**0.7434**

Table 5: *Evaluation Results for the for the Manual Test Data - Test 5*

The BLEU scores from Table 3, 4 and 5 are graphically represented in Figure 1.

It is seen from Table 3, 4 and 5 that the BLEU and the TER scores are in all cases correlated.

The interpretation of the results is focused on three directions

1. variations of the evaluation metrics across sets of test data;
2. the comparison with the Google MT on-line tool;
3. manual evaluation.

A Moses-based system, that considers also Romanian and German, is described in [3]. Although not comparable, as the experimental settings are not the same, the BLEU scores reported in this paper are 0.2789 for Romanian-German and 0.2695 for German-Romanian.

[12] mteval_v12, as implemented on www.itl.nist.gov/iad/mig//tests/mt/2008/scoring.html - last accessed on 18.04.09

[13] TER as implemented on www.cs.umd.edu/ snover/tercom - last accessed on 18.04.09

4.1 Variation of the scores across different sets of test data

An interesting aspect of the evaluation is the variation of scores across sets of test data from the same corpus, using the same system. The corpus contains data in the time interval 1958-2006. Although terminology might have changed, both languages, Romanian and German, did not suffer major transformations, e.g. at syntactic level.

Several parameters can influence the results of automatic evaluation:

- **The creation of the test data**. As mentioned in Section 2, the test data was extracted from different parts of the aligned corpus. As there is no equal distribution of sentences per year included in the corpus, it might be possible that all sentences related to e.g. 1978 EU-Regulations are in the test data but not in the training data. OOV-words (see Section 3.1) and differences in lexical semantics among years can be in this case source for the variations of the scores.

- **Sentence limitation in the Moses translation model**. This was set to 40 words / sentence. Test data had no restrictions in this sense. The average sentence-length of the test data is higher for both translation-directions (see Table 1)

- **Variation of paragraph length in the alignment**. The 1:1 alignments vary strong in length, some of them are NPs, some of them are 1-verb sentences and some contain more than one sentence.

- **Verification of the test data**. The test data is not manually checked, so that only good and "relevant" test paragraphs are used. In some test data-sets, paragraphs like "*Article article_number*" are repeated several times. Sometimes, due to the automatic extraction of the test-sets, the reference translation is wrong (error of the alignment in JRC-ACQUIS).This reduces the BLEU score.

- **Rephrasing**. When manually analyzing part of the translations (see 4.3), it was noticed that some of the translations were correct from the human evaluation point of view, but they rephrased the reference translations. As BLEU calculation is based on n-grams, this leads to a decrease of the score.

4.2 Comparison with Google MT-System

The Google system is stable, i.e. the scores are close to each-other. The BLEU score varies between 0.274 and 0.2853 for German - Romanian (0.0113 score difference) and between 0.3208 and 0.3332 for Romanian - German (0.0124 score difference). The SMT system has the difference between the scores approximately ten times higher, e.g. the BLEU score difference for German - Romanian is 0.136, and for Romanian - German is 0.1472. In order to interpret the results a more detailed manual analysis of the translations is necessary.

For German-Romanian the Moses-based system has a higher BLEU (lower TER) score than the Google one. For Romanian-German on the test-sets of 299 sentence, in two cases out of three, Google has better scores. On the 897 test-set the scores of the Moses-based system are better.

However Google is not a reliable comparison as the system evolves dynamically, by contributions of users and there is no deep information about the architecture of the system. It is estimated that the training data is huge, comparable with the one used for the experiments reported in [1]. In favor of this argument is also the scores obtained for the electronic device corpus. The Google BLEU score is very similar to the one obtained in [1] when changing the domain. In conclusion, the availability of a larger training data set would increase the performance and robustness of a pure SMT-System. Also correcting the training and test data can lead to better results.

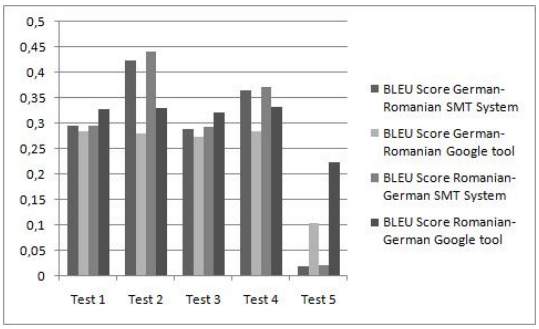

Fig. 1: *BLEU Score*

4.3 Manual Evaluation for German-Romanian

In order to extract the sources of errors, the translations of 100 paragraphs from **Test 4** data-set, obtained by the Moses-based SMT system, were manually analyzed. In order to have different paragraph-types, 50 were chosen from the beginning and 50 from the end. As the human evaluator has as mother tongue Romanian, the translation direction considered was German-Romanian.

If some paragraphs consist of only one word, it was observed that the last 50 paragraphs are longer: e.g. *paragraph 863* has 82 words and consists of one phrase and two sentences. There are 49 paragraphs shorter than 6 words.

The eight sources of translation errors are presented in **Table 6**. Some errors (e.g. OOV-words) presented in **Table 6** are due to the limited training data. Due to the German compounds and syntax, an important source of errors is the word alignment. These errors can be solved by adding more data or a bilingual dictionary.

In around 10% of the paragraphs, the translation was adequate and fluent, but it was the reference translation rephrased - e.g. passive voice translated

Error	Frequency	Explanation / Example
OOV-words	35 cases	Compounds or part of compounds "*Forschungsfonds*" ("*Research fonds*") Sometimes only half of the compound word is translated "*anpassungsprotokoll*" ("*the Protocol adjusting...*") translated as "***protocolul*** *anpassungs****protokoll***" instead of "*protocolul de adaptare*"
Punctuation		wrong position of "*)*"
Prepositions	10%	wrong or word-to-word translation" "*in das Abkommen*" ("*into the Agreement*") translated as "*din acord*" ("*from the agreement*")
Agreement, case	12%	
Missing words	23 cases nouns, articles or prepositions	Missing definite article for genitive
Missing verb	14%	This is due to the German syntax Distance between the auxiliary and main verb Subordinate sentences
Extra words	less than 5%	
Word order	around 15%	
Wrong translation (semantics)	20 cases	

Table 6: *Manual Evaluation: Sources of Errors (% means percentage from the number of paragraphs; case means the appearance of the phenomenon (i.e. in one paragraph there can be more cases)*

as active voice - or it contained synonyms. This influences negatively the automatic evaluation

5 Conclusion

In this paper the performance of an SMT system based on Moses is investigated on test data from different domains for German - Romanian, in both directions. No additional linguistic tools were used. The article presents the comparison between the results of the Moses-based SMT system and the ones given by the Google on-line translation tool. The training corpus used is the JRC-ACQUIS. The test data are taken from the JRC-ACQUIS corpus and from a manual of an electronic device.

In the described experimental settings, in all cases for German-Romanian and in some cases for Romanian-German, the Moses-based SMT system, trained and tested on the same data type, scores better than the Google on-line tool. This, in spite of the fact that both languages are inflected, and that the corpus (JRC-ACQUIS) is small and includes errors.

In the other cases, with increased and better - i.e. sentence-aligned - training data, the Google performance can be reached with a Moses-based SMT-System. As it is not a black-box system, one has the possibility to control the workflow, and introduce in a targeted way linguistic components when available.

References

[1] C. Callison-Burch, P. Koehn, C. Monz, and J. Schroeder. Findings of the 2009 workshop on statistical machine translation. In *In Proceedings of the 4th EACL Workshop on Statistical Machine Translation*, pages 1–28, Athens, Greece, March, 30-31 2009.

[2] D. Cristea. Romanian language technology and resources go to europe. Presentated at the FP7 Language Technology Informative Days, January, 20-11 2009. To be found at: ftp://ftp.cordis.europe.eu/pub/fp7/ict/docs/language-technologies/cristea_en.pdf - last accessed on 10.04.2009.

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

[4] E. Irimia. Experimente de traducere automata bazată pe exemple pentru limbile engleza/romana. In *In Resurse Lingvistice şi Instrumente pentru Prelucrarea Limbii Române Workshop Proceedings*, pages 131–140, Iaşi, Romania, November 2008. Publisher: Ed. Univ. Alexandru Ioan Cuza, ISSN: 1843-911X.

[5] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. Moses: Open source toolkit for statistical machine translation. In *Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session*, Prague, Czech Republic, June 2007.

[6] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 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, 2002. Publisher: Association for Computational Linguistics Morristown, NJ, USA.

[7] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*, 2006.

[8] R. Steinberger, B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, D. Tufis, and D. Varga. 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)*, Genoa, Italy, May, 24-16 2006.

[9] J. Tiedemann and L. Ngygaard. The opus corpus - parallel and free. In *Proccedings of the 4th International Conference of Language Resources and Evaluation*, Lisbon, Portugal, May 26-28 2004.

[10] D. Tufiş, S. Koeva, T. Erjavec, M. Gavrilidou, and C. Krstev. Building language resources and translation models for machine translation focused on south slavic and balkan languages. In *Proceedings of the Sixth International Conference Formal Approaches to South Slavic and Balkan Languages (FASSBL 2008)*, pages 145–152, Dubrovnik, Croatia, September 25-28 2008. In Marko Tadi, Mila Dimitrova-Vulchanova and Svetla Koeva (eds.).

[11] C. Vertan, W. von Hahn, and M. Gavrila. Designing a parole/simple german-english-romanian lexicon. In *In Language and Speech Infrastructure for Information Access in the Balkan Countries Workshop Proceedings - RANLP 2005*, Borovets, Bulgaria, September 2005.