Comparing Corpus-based MT Approaches Using Restricted Resources

Monica Gavrila
Hamburg University
Vogt-Koelln Str 30, 22527
Hamburg, Germany
gavrila@informatik.
uni-hamburg.de

Natalia Elita
Hamburg University
Vogt-Koelln Str 30, 22527
Hamburg, Germany
elita@informatik.
uni-hamburg.de

Abstract

Machine translation (MT) plays an important role in multilingual communication. Dealing with natural language and a diversity of language-pairs, it is not always possible to have sufficient (linguistic) resources for a specific MT approach and a diversity of domains. In this paper we compare a statistical MT system with an example-based one and a hybrid system. For a better overview we include in our comparison also an on-line MT system. We considered for our experiments a small-sized domain-restricted corpus for Romanian and English, in both directions of translation. We also tested which impact part-of-speech information has on the translation results.

1 Introduction

Machine translation (MT) plays an important role in multilingual communication (especially in the World Wide Web environment) and is already an integrated part of current natural language processing (NLP) applications, such as content management systems (CMSs)[1].

Dealing with natural language and a diversity of language-pairs, it is not always possible to have enough (linguistic) resources for a specific MT approach and a large variety of domains. Therefore, we set out focus in this paper on corpus-based MT (CBMT) approaches using a small-size corpus for training. We use for our experiments English-Romanian as language-pair, in both directions of translation.

We present several comparisons between CBMT approaches, in different experimental settings:

- Comparing statistical MT (SMT), example-based MT (EBMT) and hybrid MT (EBMT-SMT) , when no additional linguistic information is added to the corpus. The question which appears is if hybrid systems can overtake the pure CBMT approaches.

- Comparing SMT and EBMT, when part-of-speech (POS) information is added to the data. Usually it is thought that additional linguistic information helps the translation process. The questions we set is what the influence is when small-sized data are involved and which the difference is between the two main CBMT approaches (SMT and EBMT).

For a better overview we compare our results with the ones of an on-line MT system.

Experiments with smaller data (approx. 2.6K sentences) have been presented in the literature in (Popovic and Ney, 2006) for Serbian-English. Comparisons between SMT and hybrid or EBMT approaches are presented in the literature, but usually larger data is used. The marker-based EBMT system described in (Way and Gough, 2005) outperformed the SMT system presented in the same paper. In (Smith and Clark, 2009) the hybrid EBMT-SMT system is outperformed by a Moses-based SMT system. SMT and EBMT approaches for Romanian

[1] For example in the ATLAS (Applied Technology for Language-Aided CMS) project (http://www.atlasproject.eu/).

an English are shown in (Ignat, 2009) and (Irimia, 2009), respectively.

The paper is organized as follows: after the short introduction we will present the MT systems employed. In Section 3 we describe the data used in the experiments and we give a brief description of Romanian. Section 4 shows the automatic evaluation results and their interpretation. The paper ends with conclusions and further work.

2 The MT Systems

In this section we present the CBMT systems used: a Moses-based SMT system (**Mb_SMT**), a pure EBMT system ($Lin - EBMT^{REC+}$) and a hybrid (EBMT-SMT) MT system (OpenMaTrEx). For comparison reasons we also translated our test dataset with an on-line MT system: Google translate.

2.1 The SMT System: Mb_SMT (A)

The pure SMT system (**Mb_SMT**) follows the description of the baseline architecture given for the EMNLP 2011 6th Workshop on SMT[2]. **Mb_SMT** uses Moses[3], 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). We used in our experiments SRILM (Stolcke, 2002) for building the language model (LM) and GIZA++ (Och and Ney, 2003) for obtaining the word alignment information. We made two changes to the specifications of the SMT workshop: we left out the tuning step[4] and we built an LM of order 3, instead of 5[5].

2.2 The EBMT Systems: $Lin - EBMT^{REC+}$ (B)

The EBMT system in this paper ($Lin - EBMT^{REC+}$) has been developed at the University of Hamburg. It combines the linear EBMT approach with the template-based one – see (McTait, 2001) for the definitions of the EBMT approaches and templates. It is based on surface-forms and uses no linguistic resources, with the exception of the parallel aligned corpus. It contains all the three steps of an EBMT system[6]: matching, alignment and recombination. Before starting the translation, training and test data are pre-processed in the same way as in **Mb_SMT**, i.e. tokenization, lowercasing etc. In order to reduce the search space in the matching process, we use a word index. The matching procedure is an approach based on surface-forms, focusing in finding recursively the longest common substrings. If during the matching procedure the test sentence is found in the training corpus, its translation represents the output. Otherwise, the alignment and recombination steps are performed. The alignment information is extracted from the GIZA++ output of the **Mb_SMT** system. The longest TL aligned subsequences are used further in the recombination step, which is based on 2-gram information and word-order constraints. In $Lin - EBMT^{REC+}$ ideas from the template-based EBMT approach are incorporated in the recombination step, by extracting and imposing several types of word-order constraints. More information about the system, templates and how combinations of constraints influence the results is presented in (Gavrila, 2011).

2.3 The Hybrid System: OpenMaTrEx (C)

The hybrid EBMT-SMT system we used is OpenMaTrEx: a free open-source EBMT system based on the marker hypothesis. This hypothesis (Green, 1979) is a universal psycholinguistic constraint which states that natural languages are '*marked*' for complex syntactic structure at surface form by a closed set of specific lexemes and morphemes.

OpenMaTrEx consists of a marker-driven chunker, several chunk aligners, and two engines: one is based on the simple proof-of-concept monotone recombinator (called Marclator[7]) and the other uses a Moses-based decoder (called MaTrEx[8]).

From the two modes (Marclator and MaTrEx)

[2] www.statmt.org/wmt11/baseline.html.

[3] www.statmt.org/moses/.

[4] Leaving out the tuning step is motivated by the size of the data in this paper and the results we obtained in experiments which are not the topic of this paper, when comparing SMT with and without tuning. Not all tests with tuning showed an improvement.

[5] The change has been motivated by results presented in (Rousu, 2008)

[6] The steps of an EBMT system are firstly described in (Nagao, 1984).

[7] www.openmatrex.org/marclator/.

[8] www.sf.net/projects/mosesdecoder/.

in which OpenMaTrEx can be run, we chose for this paper the hybrid MT architecture, the MaTrEx mode. In this mode the system wraps around the Moses statistical decoder, using a hybrid translation table containing marker-based chunks as well as statistically extracted phrase pairs. For our experiments we followed the training and translation steps as described in (Dandapat et al., 2010).

The markers for English have been already contained in OpenMaTrEx. They were derived from the Apertium English-Catalan dictionaries[9]. The markers for Romanian were created from scratch during the experiments presented in this paper. Morpho-syntactic specifications from MULTEXT-East[10] and Wikipedia[11] were used to derive the markers. There are currently 366 Romanian and 307 English makers. More about the Romanian markers can be found in (Gavrila and Elita, 2011).

2.4 The On-line System: Google Translate (D)

For comparison reasons we included an on-line MT System – Google Translate (`translate.google.com`) – in our experiments. The system is a free statistically-based machine translation service, provided by Google Inc. It translates a section of text, document or webpage, from one source language (SL) into the target language (TL). While Google Translate is nominated as an SMT system on `Wikipedia.org`, on the Google support webpage[12] it is only stated that it uses the "*state-of-the-art technology*", without reference to any specific MT approach.

3 The Corpus

For our experiments we used a domain restricted, small-sized corpus: RoGER. It 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 (ro), English (en), German and Russian. The corpus has been manually compiled and verified. It is not annotated and diacritics are ignored. The initial text was preprocessed by replacing numbers, websites and images with "*meta-notions*" as follows: numbers by *NUM*, pictures by *PICT* and websites by *WWWSIT*E. In order to simplify the translation process, some abbreviations were expanded.

The corpus contains 2333 sentences for each language. The average sentence length is eleven tokens for English, Romanian and German and nine for Russian. More statistical data about the corpus is presented in Table 1. Punctuation signs are considered as tokens. More about the RoGER corpus can be found in (Gavrila and Elita, 2006)

From the corpus, 133 sentences have been randomly extracted as the test data, the remaining 2200 sentences being used as training data.

We considered two experimental settings: one when no additional linguistic information is added to the corpus (Experimental setting I) and one when part-of-speech (POS) information is incorporated in the corpus (Experimental setting II). While former setting uses all four MT system mentioned in Section 2, the latter employs only **Mb_SMT** and $Lin - REC^{REC+}$. This happens as only these two MT systems work with the modified corpus, with no real impact on the algorithms or other resources. However, some POS information is indirectly included in the OpenMaTrEx algorithm in the form of markers.

For the Experimental setting II we annotated the corpus by means of the text processing web services described on the website of the Research Institute for Artificial Intelligence of the Romanian Academy (RACAI)[13]. The website provides on-line web services for text processing (such as tokenization, sentence splitting, POS Tagging and lemmatization), factored translation and language identification. More information about the web-services can be found in (Tufis et al., 2008). We concatenated the POS information to the word as **WORD+"*POS*"+POS**, where "*POS*" is a delimiter. A word with POS information (**WORD+"*POS*"+POS**) is considered during the translation as one token for the corpus-based MT ap-

[9] `www.apertium.org/?id=whatisapertium\&lang=en`.

[10] `nl.ijs.si/ME/V4/msd/html/msd-ro.html`.

[11] `ro.wikipedia.org/wiki/Parte_de_vorbire`.

[12] `translate.google.com/support/?hl=en`.

[13] `http://www.racai.ro/webservices/TextProcessing.aspx` - last accessed on June 27th, 2011.

Feature	English	Romanian	German	Russian
No. tokens	26096	25850	27142	22383
Voc.* size	2012	3104	3031	3883
Voc. (*Word-frequency higher than two*)	1231	1575	1698	1904

Table 1: The RoGER corpus – Some statistics (**Voc.=vocabulary*).

proaches involved. From the information provided by the web services we only used one of the POS tags[14]

Statistical information on the data for Experimental setting I is shown in Table 2. The statistical information about the training and test data which contains POS information is presented in Table 3 (Experimental setting II).

Data SL	No. of words	Voc. size	Average sentence length
en-ro			
Training	27889	2367	12.68
Test	1613	522	12.13
ro-en			
Training	28946	3349	13.16
Test	1649	659	12.40

Table 2: RoGER statistics (Experimental setting I).

Data SL	No. of words	Voc. size	Average sentence length
en-ro			
Training	27816	2815	12.64
Test	1610	564	12.11
ro-en			
Training	28954	4133	13.16
Test	1651	735	12.41

Table 3: RoGER statistics when additional POS information is added (Experimental setting II).

3.1 Language Characteristics: Romanian

As English is the language mostly used in NLP, we will present several characteristics of Romanian in this subsection.

Romanian is a morphologically rich language, having less resources when compared with other European languages. It is an Eastern Romance language, with grammar and basic vocabulary closely related to those of its relatives (e.g. Italian, Spanish, French). It has been influenced by several other languages, such as the Slavic languages, Hungarian and Turkish. This influence is encountered especially at lexical level.

Among the language-specific characteristics induced by its Latin origin are the following: a 3-gender system, double negation and pronoun-elliptic sentences. Also, as in all Romance languages, Romanian verbs are highly inflected (according to person, number, tense, etc.) Another Latin element that has survived in Romanian while having disappeared from other Romance languages is the morphological case differentiation in nouns, albeit reduced from the original seven to only three forms (nominative/accusative, genitive/dative and vocative).

It is the only Romance language where definite articles are attached to the end of the noun or the adjective as enclitics, depending on the position of the adjective before or after the noun. This phenomenon is encountered in some Slavic languages (Bulgarian, Macedonian), in Scandinavian languages and in Albanian.

4 Experimental Results

We evaluated our translations using two automatic evaluation metrics based on n-grams: BLEU and NIST. 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 de-

[14] The C-TAG: The first tag after the lemma provided by the web services.

tails 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 evaluation scores for all four MT systems (Experimental setting I) are shown in Table 4. In this table several explanations are needed: **A** is **Mb_SMT**, **B** $Lin-EBMT^{REC+}$, **C** OpenMaTrEx and **D** Google translate.

Score	**A**	**D**	**C**	**B**
		en-ro		
BLEU	0.4386	0.4782	0.3934	0.3085
NIST	6.5599	6.9334	5.9725	5.5322
		ro-en		
BLEU	0.4765	0.5241	0.4428	0.3668
NIST	6.8022	7.4478	6.4124	6.2991

Table 4: Evaluation results for RoGER (no POS Information).

It can be seen that for all cases the pure SMT system is better than the hybrid system. The EBMT system is the last. The on-line MT system overtakes all MT systems we trained.

Table 5 shows how POS information influences the translation results of **Mb_SMT** (System **A**) and $Lin-EBMT^{REC+}$ (System **B**)

Score	**A**	**B**
	en-ro	
BLEU	0.3879	0.2916
NIST	5.8047	5.0893
	ro-en	
BLEU	0.4618	0.3559
NIST	6.3533	6.0039

Table 5: Evaluation results for RoGER (additional POS information).

A comparison between the results of the two settings (with and without additional POS in the corpus) is shown in Figure 1.

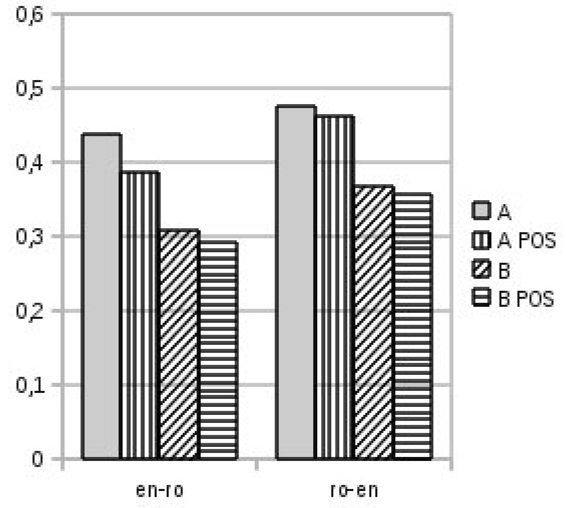

Figure 1: Comparison of the Evaluation Results

For this specific data the results which contain POS information are lower than the ones without additional information. There are two reasons for these results: either POS information is affecting negatively the translations or the automatic scores cannot capture the improvement. Therefore, we should manually analyze part of the results. The negative impact can be due to incorrect results of the web-services (incorrect POS attached) or increase of data sparseness, which has a direct impact on the statistical approaches and the word alignment.

For a better overview on the results we compared the tokens[15] of the translations with those in the references. The results are shown in Table 6 in which "*Common tokens*" (CT) are tokens which the reference and the translation have in common and "*Ordered common tokens*" (O.CT) are common tokens between the translation and its reference, which have the same order in both sentences.

For example, the following two sentences:
*I decided **to go** home **by** bus.*
*We **go to** the theater **by** car.*
have three "*common tokens*" (*to*, *go*, *by*) and two "*ordered common tokens*" (*go*, *by*).

The percentage values in Table 6 are calculated from the total number of tokens in the reference translation. The results for **Mb_SMT** are closer to the reference translation. Moreover, the use of POS information influences negatively the values.

We manually analyzed the results of **Mb_SMT**

[15] In this context token means word, number or punctuation sign.

Desc.	**Ref.**	**A**	**B**
en-ro			
Total	495	490	466
CT	-	352 (71.11%)	302 (61.01%)
O. CT	-	343 (69.29%)	244 (49.29%)
en-ro and POS			
Total	490	472	480
CT	-	273 (55.71%)	257 (52.45%)
O. CT	-	267 (54.49%)	211 (43.06%)

Table 6: Comparison between the translations and their references (Ref.=reference, Desc.=description).

and $Lin - EBMT^{REC+}$ from the point of view of adequacy[16] and fluency[17]. Although not fully relevant,as only one human evaluator was available, but still with possible impact on further research, the average results for adequacy and fluency are presented in Table 7. The evaluation scale for adequacy and fluency is the one described in (LDC, 2005):

Adequacy: 1=None, 2=Little, 3=Much, 4=Most, 5=All.

Fluency: 1=Incomprehensible, 2= Disfluent, 3=Non-native, 4=Good, 5=Flawless

Evaluation	**A**	**B**
en-ro		
Adequacy	4.22	3.64
Fluency	4.08	3.44
en-ro and POS		
Adequacy	4.1	3.66
Fluency	3.74	3.3

Table 7: System analysis: adequacy and fluency (average values).

These results confirm the automatic evaluation scores and previous analyses.

The test scenario was kept as realistic as possible. Therefore, we have not excluded test sentences already in the training corpus: common users do not analyze the texts before translating them. Next to tests sentences included in the training data, also out-of-vocabulary (OOV) words have a direct impact on the translation results. An overview of these two aspects in our data is shown in Table 8.

Corpus	**No. of OOV-Words (% from voc.* size)**	**Sentences in the corpus**
en-ro		
Test	60 (11.49%)	37 (27.81%)
Test (POS)	74 (13.12%)	37 (27.81%)
ro-en		
Test	84 (12.75%)	34 (25.56%)
Test POS	116 (15.78%)	34 (25.56%)

Table 8: Analysis of the test data sets (Experimental settings I and II) (**voc.=vocabulary*).

As expected, the number of OOV-words increases when POS information is included in the data. Also the number increases when Romanian is the source language. This happens due to the characteristics of the language.

5 Conclusions and Further Work

In this paper we presented several CBMT experiments with different approaches using a small-sized domain-restricted corpus.

Analyzing the results it can be concluded that not always additional linguistic information improves the MT results. Also combining different approaches does not always lead to better results. The training and test data themselves, the impact of additional information (such as increase of data sparseness) directly influence the translations. For under-resourced language-pairs or lower-resourced domains it can be enough just the use of a pure SMT system.

For a better understanding of the results further (manual) analysis is required. Moreover, we need to run more tests with different language-pairs and corpora. Some further results in this direction can be found in (Gavrila and Elita, 2011).

References

Dandapat, Sandipan and Mikel L. Forcada and Declan Groves and Sergio Penkale and John Tinsley and Andy Way. 2010 OpenMaTrEx: A Free/Open-Source

[16] Adequacy refers to the degree to which information present in the original is also communicated in the translation.

[17] Fluency refers to the degree to which the output is well formed according to the rules of the target language.

Marker-Driven Example-Based Machine Translation System *IceTAL'10*, pages 121–126.

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

Irimia, Elena. 2009 EBMT Experiments for the English-Romanian Language Pair *In Proceedings of the Recent Advances in Intelligent Information Systems*, 91–102, ISBN 978-83-60434-59-8.

Ignat, Camelia. 2009 Improving Statistical Alignment and Translation Using Highly Multilingual Corpora *PhD Thesis*, INSA - LGeco- LICIA, Strasbourg, France.

Gavrila, Monica and Natalia Elita. 2006 Roger - un corpus paralel aliniat *In Resurse Lingvistice şi Instrumente pentru Prelucrarea Limbii Române Workshop Proceedings*, 63–67 December, Publisher: Ed. Univ. Alexandru Ioan Cuza, ISBN: 978-973-703-208-9.

Gavrila, Monica. 2011 Constrained recombination in an example-based machine translation system *Proceedings of the EAMT-2011: the 15th Annual Conference of the European Association for Machine Translation*, May, Leuven, Belgium.

Gavrila, Monica and Natalia Elita. 2011 Experiments with Small-size Corpora in CBMT *Proceedings of RANLP Student Research Workshop* September, Hissar, Bulgaria.

Green, T.R.G. 1979 The necessity of syntax markers: Two experiments with artificial languages *Journal of Verbal Learning and Verbal Behavior*, Volume 18, Number 4, 481 – 496. ISSN 0022-5371.

Koehn, Philipp and Hieu Hoang and Alexandra Birch and Chris Callison-Burch and Marcello Federico and Nicola Bertoldi and Brooke Cowan and Wade Shen and Christine Moran and Richard Zens and Chris Dyer and Ondrej Bojar and Alexandra Constantin and Evan Herbst. 2007 Moses: Open Source Toolkit for Statistical Machine Translation *Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session*, June, Prague, Czech Republic

Linguistic Data Consortium. 2005 Linguistic data annotation specification: Assessment of fluency and adequacy in translations. Revision 1.5. Technical report `http://projects.ldc.upenn.edu/TIDES/Translation/TransAssess04.pdf`

McTait, Kevin. 2002 *Translation Pattern Extraction and Recombination for Example-Based Machine Translation* PhD Thesis, Center for Computational Linguistics, Department of Language Engineering, PhD Thesis, UMIST.

Nagao, Makoto. 1984 A Framework of a Mechanical Translation between Japanese and English by Analogy Principle *Proceedings of the international NATO symposium on Artificial and human intelligence*, 173–180 New York, NY, USA, Elsevier North-Holland, Inc., ISBN 0-444-86545-4, Lyon, France.

Och, Franz Josef and Hermann Ney. 2003 A Systematic Comparison of Various Statistical Alignment Models *Journal of Computational Linguistics*, Volume 29, Number, pages 19–51

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

Popovic, Maja and Hermann Ney. 2006 Statistical machine translation with a small amount of bilingual training data. *LREC-2006: Fifth International Conference on Language Resources and Evaluation. 5th SALTMIL Workshop on Minority Languages: Strategies for developing machine translation for minority languages*, 25-29, Genoa, Italy, May.

Rousu, Juho. 2008 SMART Project: Workpackage 3 advanced language models. On-line material, `http://www.smart-project.eu/files/SMART-Y1-review-WP3.v1.pdf` (last accessed on September 5th, 2011).

Smith, James and Stephen Clark. 2009 EBMT for SMT: A new EBMT-SMT hybrid. *Proceedings of the 3rd International Workshop on Example-Based Machine Translation* 3–10, Editors Mikel L. Forcada and Andy Way, Dublin, Ireland.

Stolcke, Andreas. 2002 SRILM - An Extensible Language Modeling Toolkit *Proc. Intl. Conf. Spoken Language Processing*, 901–904 September, Denver, Colorado.

Tufis, Dan and Radu Ion and Alexandru Ceausu and Dan tefnescu. 2008 RACAI's Linguistic Web Services *Proceedings of the 6th Language Resources and Evaluation Conference - LREC 2008* Marrakech, Morocco, May ELRA - European Language Resources Association. ISBN 2-9517408-4-0

Andy Way and Nano Gough. 2005 Comparing example-based and statistical machine translation *Natural Language Engineering*, 1:295309, September.