Experiments with Small-sized Corpora in CBMT

Monica Gavrila
Hamburg University
gavrila@informatik.uni-hamburg.de

Natalia Elita
Hamburg University
elita@informatik.uni-hamburg.de

Abstract

There is no doubt that in the last couple of years corpus-based machine translation (CBMT) approaches have been in focus. Each of the approaches has its advantages and disadvantages. Therefore, hybrid approaches have been developed. This paper presents a comparative study of CBMT approaches, using three types of systems: a statistical MT (SMT) system, an example-based MT (EBMT) system and a hybrid (EBMT-SMT) system. We considered for our experiments three languages, from different language families: Romanian, German and English. Two different types of corpora have been used: while the first is manually created, the latter is automatically built.

1 Introduction

There is no doubt that in the last couple of years corpus-based machine translation (CBMT) approaches have been in focus. Among them, the statistical MT (SMT) approach has been by far more dominant, but the example-based machine translation (EBMT) Workshop at the end of 2009[1] and the new open-source systems (e.g. OpenMaTrEx – see section 2.3) showed a revived interest in the EBMT and hybrid approaches.

The unclear definitions and the mixture of ideas make the difference between the two CBMT approaches difficult to distinguish. In order to show the advantages of one or another method, comparisons between SMT and EBMT (or hybrid) systems have been presented in the literature. To get advantage of positive sides of both CBMT approaches, hybrid systems have been developed. The results, depending on the data type and the systems considered, seem to be positive for various approaches. 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. In both papers the language pair under consideration is English - French.

In this paper we compare several CBMT approaches, using three MT systems: an SMT system (**Mb_SMT**), an EBMT system ($Lin - EBMT^{REC+}$) and a hybrid (EBMT-SMT) system (OpenMaTrEx). MT experiments are run for two language pairs, in both directions of translation: Romanian (ro)-English (en), Romanian (ro)-German (ge). In contrast to other authors, for example (Smith and Clark, 2009), we use small-sized domain-restricted corpora for training. It is usually believed that small-size copora better fit into the EBMT environment. The use of small-sized copora for SMT has been tried before: (Popovic and Ney, 2006) present results for Serbian-English and a training data size of approx. 2.6K sentences. However, to our knowledge, no comparisons among CBMT systems using small-sized data have been published.

Even more, for the language pairs employed in this paper no other comparative studies have been published. Nevertheless, separate results for EBMT and SMT have been presented: EBMT results in (Irimia, 2009)[2] and SMT in (Cristea, 2009) and (Ignat, 2009). All these experiments use for training and testing the JRC-Acquis corpus.

Our paper is organized as follows: the following section presents the MT systems employed. In Section 3 the data used is described and the translation results are interpreted. The paper ends with conclusions and further work.

2 System Description

In this section we present the three CBMT systems we used: an SMT system (**Mb_SMT**), an

[1] computing.dcu.ie/~mforcada/ebmt3/ - last accessed on June 21st, 2011.

[2] Only English and Romanian have been under consideration.

Proceedings of the Student Research Workshop associated with RANLP 2011, pages 67–72,
Hissar, Bulgaria, 13 September 2011.

EBMT system ($Lin - EBMT^{REC+}$) and a hybrid (EBMT-SMT) system (OpenMaTrEx).

2.1 The SMT System: Mb_SMT

The pure SMT system (**Mb_SMT**) follows the description of the baseline architecture given for the Sixth Workshop on SMT[3] at the EMNLP 2011 Conference. **Mb_SMT** uses Moses[4], 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 the SMT system. Not all tests for the system configuration which included tuning showed an improvement. Changing the LM order has been motivated by results reported in the SMART project[5].

2.2 The EBMT System: $Lin - EBMT^{REC+}$

$Lin - EBMT^{REC+}$ is an EBMT system which combines the linear EBMT approach with the template-based one – see (McTait, 2001) for the classification of EBMT approaches and the definition of a template. Before starting the translation, training and test data are pre-processed (such as tokenization and lowercasing) as in the Moses-based SMT system. We use a token[6]-index in order to reduce the search space in the matching process. In case the test sentence is found in the training corpus during the matching procedure, its translation represents the output. Otherwise, the alignment and recombination steps are performed. The matching procedure is an approach based on surface-forms, focusing in recursively finding the longest common substrings. The alignment information is extracted from the GIZA++ output of the **Mb_SMT** system. The longest target language (TL) aligned subsequences are used 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 three types of word-order constraints: First word constraints; Constraints extracted from the target language side of a template; Constraints extracted from both sides of a template. More information about the system, templates and how combinations of constraints influence the evaluation results has been presented in (Gavrila, 2011).

2.3 The Hybrid System: OpenMaTrEx

OpenMaTrEx is a free (open-source) EBMT system based on the marker hypothesis (Dandapat et al., 2010).

The marker hypothesis (Green, 1979) is a universal psycholinguistic constraint which states that natural languages are '*marked*' for complex syntactic structures at surface form by a closed set of specific lexemes and morphemes. That is, a basic phrase-level segmentation of an input sentence can be achieved by exploiting a closed list of known marker words to signalize the start and end of each segment.

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]).

The system uses GIZA++ for word alignments and IRSTLM[9] to obtain the LM. The complete architecture of OpenMaTrEx is described in (Dandapat et al., 2010) and (Stroppa et al., 2006). OpenMaTrEx can be run in two modes: Marclator and MaTrEx. In the MaTrEx mode it 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). Only

[3] `www.statmt.org/wmt11/baseline.html` - last accessed on July 14th, 2011.

[4] `www.statmt.org/moses/` - last accessed on July 14th, 2011.

[5] `www.smart-project.eu` - last accessed on July 14th, 2011.

[6] A token is represented by a word form, a number or a punctuation sign.

[7] `www.openmatrex.org/marclator/` - last accessed on July 1st, 2011.

[8] `www.sf.net/projects/mosesdecoder/` - last accessed on July 1st, 2011.

[9] `http://hlt.fbk.eu/en/irstlm` - last accessed on July 21sth, 2011.

the results of the run in MaTrEx mode (the hybrid MT architecture) are shown in the current article, as this is the usual way to use OpenMaTrEx, according to its developers.

2.3.1 Marker Words Files

In this subsection we present the marker words files for Romanian developed during this research. The markers for English and German have been already contained in the system: The English markers were derived from the Apertium English-Catalan dictionaries[10]; The German markers were extracted from the "Ding" dictionary by Sarah Ebling[11].

We extracted the markers for Romanian during the experiments presented in this paper by considering the morpho-syntactic specifications from MULTEXT-East[12] and Wikipedia[13].

The set of markers for Romanian consists of the chunking and non-chunking punctuation that has been acquired from the English marker words file. The other word categories included in the file are: determiners, pronouns (personal, demonstrative, possessive, interrogative, relative), prepositions, conjunctions (coordinative and subordinative), (cardinal) numerals, adverbs and auxiliary verbs.

Definite articles and weak forms of the personal pronouns are two examples of clitic forms in Romanian. We have not considered the definite articles as markers, as they appear within the word as endings (e.g. ro: *dosareLE* – en: *THE files*). Personal pronouns separated by a hyphen have not been included in the set of markers (e.g. ro: *LE-am citit* – en: *I read THEM*).

Some of the determiners are ambiguous, as they can also be pronouns or numerals (e.g. ro. *O fată*) – en: *A girl*; ro: *ia-O* – en: *take IT*; ro: *O pară şi două mere* – en: *A pear and two apples*). Only given the context it can be determined whether the word is a determiner, a numeral or a pronoun. In order to avoid ambiguity, indefinite articles were introduced as determiners in the set of markers and the category *determiner pronoun* was included only once under the category of pronouns.

There are currently 366 Romanian, 307 English and 656 German markers. Both German and Romanian have diacritics: in case of German - both versions (with and without diacritics) of the same marker word are included in the file. In case of Romanian, we created two separate sets of markers: one with and one without diacritics.

3 Evaluation

In this section, before the evaluation results are presented, we describe the training and test data used in the experiments.

3.1 Data Description

We used for the evaluation two different types of corpora, both having the same size: RoGER, a manual of an electronic device, and JRC-Acquis$_{SMALL}$, a sub-part of JRC-Acquis which contains regulations of the European Union (EU).

RoGER is a domain-restricted parallel corpus, including four languages (**Ro**manian, **E**nglish, **G**erman and **R**ussian). It is manually aligned at sentence level. Moreover, the text is manually preprocessed, by replacing concepts such as numbers and web pages, with '*meta-notions*' – for example numbers with *NUM*. It contains no diacritics. More information about the RoGER corpus can be found in (Gavrila and Elita, 2006).

Its small size (2333 sentences) is compensated by the correctness of the translations and sentence alignments. We randomly extracted 133 sentences, which we used as test data for all three MT systems. The rest of 2200 sentences represent the training data. Statistical information about RoGER is shown in Table 1.

Data SL	No. of tokens	Voc.	Average sent. length
English-Romanian			
Training	27889	2367	12.68
Test	1613	522	12.13
Romanian-English, Romanian-German			
Training	28946	3349	13.16
Test	1649	659	12.40
German-Romanian			
Training	28361	3230	12.89
Test	1657	604	12.46

Table 1: RoGER statistics (SL= source language, voc.=vocabulary, sent.=sentence or sentences).

The second corpus considered, JRC-

[10] `www.apertium.org/?id=whatisapertium\&lang=en` - last accessed on June 21st, 2011.

[11] `www-user.tu-chemnitz.de/~fri/ding/` - last accessed on June 21st, 2011.

[12] `nl.ijs.si/ME/V4/msd/html/msd-ro.html` - last accessed on July 1st, 2011.

[13] `ro.wikipedia.org/wiki/Parte_de_vorbire` - last accessed on July 1st, 2011.

Acquis$_{SMALL}$, is a sub-corpus of the JRC-Acquis (Steinberger et al., 2006). To analyze how the systems behave in case of another type of small-sized corpus, 2333 sentences have been randomly extracted from the center of the whole JRC-Acquis data. These sentences form the JRC-Acquis$_{SMALL}$ corpus. From this data, 133 sentences have been randomly selected as test data. The rest of 2200 remain as training data. JRC-Acquis$_{SMALL}$ has not been manually verified or modified. More information about the corpus can be found in Table 2.

Data SL	No. of tokens	Voc.	Average sent. length
English-Romanian			
Training	75405	3578	34.27
Test	4434	992	33.33
Romanian-English			
Training	72170	5581	32.80
Test	4325	1260	32.51
German-Romanian			
Training	69735	5929	31.69
Test	3947	1178	29.67
Romanian-German			
Training	75156	6390	34.16
Test	4366	1320	32.82

Table 2: JRC-Acquis$_{SMALL}$ statistics.

The three languages used in this paper present different morphological and syntactical characteristics. As English has been used quite often in MT experiments, for a better understanding of the translation challenges, we will briefly describe Romanian and German in the following paragraphs .

Romanian is a lesser resourced language with a highly infected morphology and high demand for translation after joining the European Union in 2007. It is a Romance language, with influence from Slavic languages especially on vocabulary and phonetics. Features, such as its inflectional system, or the three genders, make difficult the adaptation of language technology systems for other family-related languages.

German is a Germanic language, which is also inflected and presents a 3-gender system and well defined inflection classes. Two special features are represented by the verbs with particles (the separation of the particle from the verb inside the sentence and the challenge that the particle can be ambiguous) and the compounds. Compounds in German are normally written as single words, without spaces or other word boundaries.[14]

Analyzing Tables 1 and 2 ifferences in the text style can be also noticed in the average length of the sentences: between 12 and 13 tokens for RoGER and between 29 and 34 for JRC-Acquis$_{SMALL}$. The total number of tokens and the vocabulary size reinforce the differences between the languages: the vocabulary size for the inflected languages is higher as the one for English; the total numbers of tokens for German is lower, as German uses more compounds.

3.2 Automatic Evaluation Results

We evaluated the obtained translations automatically by using the BLEU (**bil**ingual **e**valuation **u**nderstudy) score. BLEU measures the number of n-grams, of different lengths, of the system output that appear in a set of references. More information on BLEU can be found in (Papineni et al., 2002). We considered the twelfth version of the BLEU implementation from the National Institute of Standards and Technology (NIST)[15]: *mteval_v12*.

Although BLEU is criticized in the research environment, the choice of the metrics is motivated by our resources (software, linguistic resources, etc.) and, for comparison reasons, by results reported in the literature. Due to lack of data and further translation possibilities, the comparison with only one reference translation is considered.

The obtained results are presented in Tables 3 (for RoGER)and 4 (for JRC-Acquis$_{SMALL}$). In the following subsection we will analyze these results.

3.3 Interpretation of the Results

In order to be able to analyze better the results, we examined the test data set from two points of view: the number of out-of-vocabulary words (OOV-words) and the number of test sentences already found in the training data. Both aspects have a direct influence on the translation quality and

[14]The longest German word verified to be actually in (albeit very limited) use is Rindfleischetikettierungsüberwachungsaufgabenübertragungsgesetz, which, literally translated, is "beef labelling supervision duty assignment law" [from Rind (cattle), Fleisch (meat), Etikettierung(s) (labelling), Überwachung(s) (supervision), Aufgaben (duties), Übertragung(s) (assignment), Gesetz (law)].

[15]`www.itl.nist.gov/iad/mig//tests/mt/2008/scoring.html` - last accessed on June 14th, 2011.

Mb_SMT	$Lin - EBMT^{REC+}$	Open-MaTrEx
English – Romanian		
0.4386	0.3085	0.4320
Romanian – English		
0.4765	0.3668	0.4663
German – Romanian		
0.3240	0.2646	0.2564
Romanian – German		
0.3405	0.2894	0.3058

Table 3: BLEU results (RoGER).

Mb_SMT	$Lin - EBMT^{REC+}$	Open-MaTrEx
English – Romanian		
0.4801	0.3550	0.4446
Romanian – English		
0.4904	0.3910	0.4771
German – Romanian		
0.2811	0.2167	0.2468
Romanian – German		
0.2926	0.2458	0.2433

Table 4: BLEU results (JRC-Acquis$_{SMALL}$).

evaluation results. These results for RoGER and JRC-Acquis$_{SMALL}$ are presented in Tables 5 and 6.

No. OOV-words (% from voc. size)	Sent. in the training corpus
English-Romanian	
60 (11.49%)	37 (27.8%)
Romanian-English	
84 (12.75%)	34 (25.5%)
German-Romanian	
101 (16.72%)	31 (23.3%)
Romanian-German	
84 (12.75%)	34 (25.5%)

Table 5: Analysis of the test data set (RoGER).

It could be noticed that all systems work better for English-Romanian (both directions of translations) than for German-Romanian (both directions of translations). The lower results for the translation direction German-Romanian can be also explained by the number of OOV-words and sentences found in the training data. We notice a similar behavior for both corpora for Romanian-English, in both directions of translation. For

No. OOV-words (% from voc. size)	Sent. in the training corpus
English-Romanian	
72 (7.25%)	38 (28.5%)
Romanian-English	
129 (10.23%)	33 (24.8%)
German-Romanian	
171 (14.51%)	41 (30.82%)
Romanian-German	
160 (12.12%)	40 (30.0%)

Table 6: Analysis of the test data set (JRC-Acquis$_{SMALL}$).

all three MT systems the results for Romanian-English are better than for English-Romanian. Generally, also the results for Romanian-German are better than the ones for German-Romanian. This behavior could mean that building the output for Romanian is more difficult than for the other two languages. Moreover, the German compound nouns could cause data-sparsity.

Compared with the other systems **Mb_SMT** works the best. OpenMaTrEx has the results quite close to the ones of **Mb_SMT**. It is better than the EBMT system with only two exceptions: for German-Romanian and the RoGER data or for Romanian-German and the JRC-Acquis$_{SMALL}$ data, $Lin - EBMT^{REC+}$ gives slightly better results than OpenMaTrEx. While comparing the **Mb_SMT** and OpenMaTrEx, we obtained results similar to the ones in (Smith and Clark, 2009)[16]. The difference is only the corpus size: (Smith and Clark, 2009) used a large-sized corpus (the Europarl corpus) in their experiments.

4 Conclusions and Further Work

In this paper three corpus-based MT systems have been compared using the same test and training data. MT experiments were made for two language pairs (Romanian-English, Romanian-German), in both directions of translation. Two small-sized domain-restricted corpora of different types were used in the experiments – a framework which is thought to better fit the EBMT approach.

In order to establish which system is really the best, as the BLEU score has been criticized in the last couple of years, a manual analysis of the results is currently being made. Splitting Ger-

[16] A one-to-one comparison is not possible, as the training and test data are different.

man compounds to avoid data sparsity is our next action point. We also need to test the systems with larger copora to analyze how the quality of translation changes when the size of the corpus is progressively incremented. Other interesting aspects we consider is running OpenMaTrEx under the Marclator mode and testing how changing (increasing) the list of markers influences the results.

References

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

Sandipan Dandapat, Mikel L. Forcada, Declan Groves, Sergio Penkale, John Tinsley, and Andy Way. 2010. Openmatrex: A free/open-source marker-driven example-based machine translation system. In *IceTAL'10*, pages 121–126.

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

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

Thomas R.G. Green. 1979. The necessity of syntax markers: Two experiments with artificial languages. *Journal of Verbal Learning and Verbal Behavior*, 18(4):481 – 496.

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.

Elena Irimia. 2009. Ebmt experiments for the english-romanian language pair. In *Proceedings of the Recent Advances in Intelligent Information Systems*, pages 91–102. ISBN 978-83-60434-59-8.

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 *Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session*, Prague, Czech Republic, June.

Kevin McTait. 2001. *Translation Pattern Extraction and Recombination for Example-Based Machine Translation*. Ph.D. thesis, Centre for Computational Linguistics, Department of Language Engineering, PhD Thesis, UMIST.

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.

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

James Smith and Stephan Clark. 2009. Ebmt for smt: A new ebmt-smt hybrid. In Mikel L. Forcada and Andy Way, editors, *Proceedings of the 3rd Intyernational Workshop on Example-Based Machine Translation*, pages 3–10, Dublin, Ireland, November, 12-13.

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)*, 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.

Nicolas Stroppa, Declan Groves, Andy Way, and Kepa Sarasola. 2006. Example-based machine translation of the basque language. In *Proceedings of AMTA 2006 – 7th Conference of the Association for Machine Translation in the Americas*, pages 232–241, Cambridge, MA, USA., August.

Andy Way and Nano Gough. 2005. Comparing example-based and statistical machine translation. *Natural Language Engineering*, 11:295–309, September.