Training Data in Statistical Machine Translation – The More, the Better? –

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

Cristina Vertan
University of Hamburg, Germany
cristina.vertan@uni-hamburg.de

Abstract

Current statistical machine translation (SMT) systems are stated to be dependent on the availability of a very large training data for producing the language and translation models. Unfortunately, large parallel corpora are available for a limited set of language pairs and for an even more limited set of domains.

In this paper we investigate the behavior of an SMT system exposed to training data of different sizes and types. Our experimental results show that even parallel corpora of modest sizes can be used for training purposes without lowering too much the evaluation scores. We consider two language pairs in both translation directions for the experiments: English-Romanian and German-Romanian.

1 Introduction

Statistical machine translation (SMT) is the most frequently used paradigm, especially when a translation system has to be implemented for a new (less researched) language pair. The pure statistical approach has the advantage that no additional bilingual linguistic expertise is required. Once the training data is available, open-source, language independent systems can be reused. However, the quality of the results is strongly influenced by the size and type of the available training data.

State-of-the-art literature tends to share the opinion that the larger the data, the better the results. (Suresh, 2010) shows that a larger corpus size for training increases the quality of a Moses-based SMT system, for the Europarl corpus for English-French. The same conclusion appears also in (Koehn et al., 2003), for German-English. In (Brants et al., 2007) experiments for Arabic-English data with billions of tokens are presented and a dependency between the output quality and the size of the training data is also demonstrated.

Unfortunately, large amount of parallel training data is available only for a restricted number of language pairs and domains. Additionally, the training step on large corpora is time and (computing-) resources consuming. On the other hand, smaller corpora can be more easily achieved and have the advantage of requiring less time for training. They also offer the possibility of manually correcting and creating the data.

Experiments with smaller data for Serbian-English (approx. 2.6K sentences) are presented in (Popovic and Ney, 2006). In the same paper also experimental results for Spanish-English, with different data sizes are reported. The systems trained on smaller data give acceptable results. However, the trend remains the same: larger data provides better results.

For English-Romanian, SMT systems are presented in (Cristea, 2009) and (Ignat, 2009), with BLEU results of 0.5464 and 0.3208, respectively. Although both systems use as training and test data parts of the JRC-Acquis corpus, the architecture described in (Cristea, 2009) involves the use of linguistic resources and the system implemented in (Ignat, 2009) uses pivot languages. As long as comparisons are not made on identical training and test data, it is difficult to estimate if, overall, the inclusion of linguistic tools increases significantly the performance. The SMT results for Romanian-English, German-Romanian and Romanian-German reported in (Ignat, 2009) are 0.3840, 0.2373 and 0.2415, respectively. For Romanian-English the BLEU score obtained in (Cristea, 2009) is 0.4604.

Especially for MT systems embedded in online applications, which face a dynamic domain change and involve several language pairs, it is extremely important to be aware of the small amount of training data which is available. Such a case

Proceedings of Recent Advances in Natural Language Processing, pages 551–556,
Hissar, Bulgaria, 12-14 September 2011.

is the ATLAS content management system, developed within the EU-Project "Applied Language Technology for Content Management Systems"[1]. In this project a machine translation (MT) engine should be available to translate abstracts from various domains across twelve language pairs.

In this paper we present the results of a Moses-based SMT system, trained on different types of small size corpora (2.2K). For comparison reasons we additionally consider a larger corpus (330K). Especially with respect to the availability of parallel corpora and linguistic resources, Romanian can be considered a lesser resourced language[2].

We chose two language pairs (English-Romanian and German-Romanian) in both directions of translations and, in contrast to (Popovic and Ney, 2006), we use for all experiments the same language pairs. The language pair Romanian (ro)-German (ge) is particularly interesting as both languages present morphological and syntactical features which do not occur in English (en) and make the process of translation even more challenging.

In the following sections we present the Moses-based SMT system used and the data employed in our experiments (Section 2), the translation results and their interpretation (Section 3). Conclusions and further work are described in Section 4.

2 Experimental Setting

2.1 The SMT System

Our MT system follows the description of the baseline architecture provided at the Sixth Workshop on SMT[3] and uses Moses[4]. Moses implements the statistical paradigm and allows the user to train automatically translation models (TM) for the involved language pair. It is assumed that the user has the required training data. The target language model (LM) and the word alignment for the parallel corpus are obtained through external applications. We used for our experiments SRILM[5] and GIZA++ [6], respectively.

Two changes have been made to the specifications of the Workshop on SMT: we left out the tuning step and considered the language model (LM) order 3 (instead of 5). Leaving out the tuning step is motivated by previous experiments we made, in which the tuned system did not always provide the best results. A reason for choosing the order three for the LM was provided by the results shown in the presentation of the SMART[7] project (Rousu, 2008), in which it was stated that "3-grams work generally the best".

2.2 Data Description

We want to study the influence of the training data on the translation results. Therefore, we use for our experiments three corpora of different sizes, which have various compilation methods: **JRC-Acquis_L** (a large-size parallel corpus, automatically aligned at sentence level), **JRC-Acquis_S** (a small-size parallel corpus, automatically aligned at sentence level), and **RoGER_S** (a small-size technical manual, manually compiled and aligned at sentence level).

The first corpus (**JRC-Acquis_L**) is part of the JRC-Acquis[8], a freely available parallel corpus in 22 languages, which consists of European Union documents of legal nature. In order to reduce errors we considered only the one-to-one sentence alignments obtained with Vanilla[9]. In fact, the alignment is realized at paragraph level[10], where a *paragraph* can be a simple or complex sentence, or a sub-sentential phrase (such as a noun phrase). More details on JRC-Acquis can be found in (Steinberger et al., 2006).

Filtering the sentence alignments had different influences on the data-size. For English - Romanian, from 391324 links ($< p >$-alignments) in 6557 documents, only 336509 links were retained. Subsequently, the cleaning step[11] of the SMT system reduced the translation model (TM) to 240219 links. This represents approx. 61.38% of the initial corpus. For German - Romanian, from 391972

[1] http://www.atlasproject.eu.

[2] While the interest for translation from or into German or English appeared in an early stage of MT, an increased demand for automatic translation from and into Romanian was noticed after the enlargement of the European Union in 2007.

[3] http://www.statmt.org/wmt11/baseline.html.

[4] http://www.statmt.org/moses/, (Koehn et al., 2007).

[5] http://www.speech.sri.com/projects/srilm/, (Stolcke, 2002).

[6] Details on GIZA++ can be found in (Och and Ney, 2003).

[7] www.smart-project.eu - last accessed on June 27th, 2011.

[8] The JRC Collection of the Acquis Communautaire: http://wt.jrc.it/lt/Acquis/.

[9] See http://nl.ijs.si/telri/Vanilla/.

[10] The tag $< p >$ from the initial HTML files.

[11] The cleaning step is integrated in Moses and supposes the elimination of sentences longer than 40 words.

links in 6558 documents, only 324448 links were considered for the LM. The TM was reduced to 238172 links (i.e 60.76% of the initial corpus).

The corpus is not manually corrected. Therefore, translation, alignment or spelling errors might influence negatively the output quality.

The tests were run on 897 (3 x 299) sentences, which were not used for training. Sentences were randomly removed from different parts of JRC-Acquis to ensure a relevant lexical, syntactic and semantic coverage. These test sets of 299 sentences represent in the following sections the data sets **Test 1**, **Test 2**, and **Test 3**. **Test 1+2+3** is formed from all 897 sentences. The test data has no sentence length restriction. Some statistical information on JRC-Acquis_L are summarized in Table 1, in which an item represents a word, a number or a punctuation sign.

Data	**No. of items**	**Voc.* size**	**Average sent.* length**
en – ro			
Training (SL)	3579856	39784	14.90
LM Romanian	9572058	81616	28.45
Test 1 (SL)	6424	1048	21.48
Test 2 (SL)	7523	735	25.16
Test 3 (SL)	5609	1111	18.76
Test 1+2+3 (SL)	19556	2345	21.80
ro – en			
Training (SL)	3386495	55871	14.10
LM English	9955983	55856	29.59
Test 1 (SL)	5672	1245	18.97
Test 2 (SL)	7194	923	24.06
Test 3 (SL)	5144	1355	17.20
Test 1+2+3 (SL)	18010	2717	20.08
ge – ro			
Training (SL)	3256047	76600	13.67
LM Romanian	9122333	80484	28.12
Test 1 (SL)	5325	1140	17.81
Test 2 (SL)	10286	1439	34.40
Test 3 (SL)	5125	1292	17.23
Test 1+2+3 (SL)	20763	3000	23.15
ro – ge			
Training (SL)	3453586	56219	14.50
LM German	8469146	121969	26.10
Test 1 (SL)	5432	1294	18.17
Test 2 (SL)	11488	1663	38.42
Test 3 (SL)	5317	1388	17.78
Test 1+2+3 (SL)	22237	3336	24.79

Table 1: Corpus statistics for JRC-Acquis_L (* voc = vocabulary, sent=sentence).

The second corpus we used is **JRC-Acquis_S**, a sub-corpus of JRC-Acquis_L, which consists of 2333 sentences. The sentences were extracted from the middle of JRC-Acquis_L. From these, 133 sentences were randomly selected as test data. The remaining 2200 sentences represent the training data. The statistics on this corpus are presented in Table 2.

Data SL	**No. of items**	**Voc.**	**Average sent. length**
en – ro			
Training	75405	3578	34.27
Test	4434	992	33.33
ro – en			
Training	72170	5581	32.80
Test	4325	1260	32.51
ge – ro			
Training	69735	5929	31.69
Test	3947	1178	29.67
ro – ge			
Training	75156	6390	34.16
Test	4366	1320	32.82

Table 2: Statistics for JRC-Acquis_S.

RoGER_S, the third corpus in this paper, is a parallel corpus, consisting of technical texts in four languages[12], which is manually aligned at sentence level. The text is preprocessed by replacing concepts such as numbers or web pages with '*meta-notions*': numbers = NUM, websites = WWW etc. More about the RoGER corpus can be found in (Gavrila and Elita, 2006). RoGER_S has the same number of training and test sentences as JRC-Acquis_S. The main difference to JRC-Acquis_S is the correctness of the translations and sentence alignments. The statistical information about this corpus is presented in Table 3.

Data SL	**No. of items**	**Voc.**	**Average sent. 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
ge – ro			
Training	28361	3230	12.89
Test	1657	604	12.46
ro – ge			
Training	28946	3349	13.16
Test	1649	659	12.40

Table 3: Statistics for RoGER_S.

3 Evaluation and Interpretation of Translation Results

3.1 Automatic Evaluation

The obtained translations have been evaluated using two automatic metrics: BLEU and TER. The choice of the metrics is motivated by the available

[12] **Ro**manian, **G**erman, **E**nglish, **R**ussian.

resources and, for comparison reason, by the results reported in the literature. The comparison was done with only one reference translation, as we work in a realistic scenario with dynamic domain change (see section 1.)

Although criticized, **BLEU** (**bil**ingual **e**valuation **u**nderstudy) is the score mostly used for MT evaluation in the last couple of years. 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[13] can be found in (Papineni et al., 2002).

TER[14] calculates the minimum number of edits required 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).

In Table 4 we present the results we obtained for all three corpora. The boldface numbers represent the highest scores for the specific language combination and evaluation metric.

Score	RoGER_S	JRC-Acquis_S	JRC-Acquis_L (Test 1+2+3)
		en – ro	
BLEU	0.4386	**0.4801**	0.4015
TER	**0.3784**	0.5032	0.5023
		ro – en	
BLEU	0.4765	**0.4904**	0.4255
TER	**0.3465**	0.4509	0.4457
		ge – ro	
BLEU	0.3240	0.2811	**0.3644**
TER	**0.5239**	0.6658	0.6113
		ro – ge	
BLEU	0.3405	0.2926	**0.3726**
TER	**0.5570**	0.6816	0.6112

Table 4: Evaluation results (all three corpora).

The results from Table 4 for Romanian-English are overall similar with state-of-the art evaluation described in Section 1. For Romanian-German our result overtake the system presented in (Ignat, 2009). However, a truly one-to-one comparison is not possible, as we do not work with identical test and training data as the referred systems.

Score	Test 1	Test 2	Test 3	Test 1+2+3
		en – ro		
BLEU	0.3997	**0.4179**	0.3797	0.4015
TER	0.5007	**0.4898**	0.5208	0.5023
		ro – en		
BLEU	0.2545	**0.5628**	0.4271	0.4255
TER	0.5020	**0.3756**	0.4684	0.4457
		ge – ro		
BLEU	0.2955	**0.4244**	0.2884	0.3644
TER	0.6200	**0.5905**	0.6438	0.6113
		ro – ge		
BLEU	0.2953	**0.4411**	0.2939	0.3726
TER	0.6437	**0.5588**	0.6791	0.6112

Table 5: Evaluation results for JRC-Acquis_L

Even for same training data evaluation results may vary across test sets, as presented in Table 5. Here we show how dependent are the SMT results on the test data. As the size and domain-type of the test data (**Test 1** - **Test 3**) is identical, the differences in BLEU and TER score can be explained only through lexical and syntactical variation across test-sets. Some sources for these variations are represented by out-of-vocabulary words (OOV-words) and the number of test sentences already found in training data. An overview of these two aspects in all the three corpora can be seen in Tables 6 and 7. As expected, best results are obtained for the test data set which has less OOV-words and which contains most sentences in the training data: **Test 2**. As it is not the topic of this paper, we will not extend the explanation for these variations or present any possible solutions.

Corpus	No. of OOV-Words (% from voc. size)	Sentences in the corpus
	JRC-Acquis_L **en – ro**	
Test 1	33 (3.15%)	69 (23.07%)
Test 2	2 (0.27%)	134 (44.81%)
Test 3	96 (8.64%)	85 (28.42%)
Test 1+2+3	131 (5.59%)	288 (21.10%)
	ro – en	
Test 1	51 (4.10%)	69 (23.07%)
Test 2	7 (0.76%)	117 (39.13%)
Test 3	111 (8.19%)	81 (27.09%)
Test 1+2+3	169 (6.22%)	267 (29.76%)
	ge – ro	
Test 1	69 (6.05%)	73 (24.41%)
Test 2	53 (3.68%)	121 (40.46%)
Test 3	187 (14.47%)	83 (27.75%)
Test 1+2+3	309 (10.30%)	277 (30.88%)
	ro – ge	
Test 1	44 (3.40%)	76 (25.41%)
Test 2	97 (5.83%)	109 (36.45%)
Test 3	105 (7.56%)	79 (26.42%)
Test 1+2+3	246 (7.37%)	264 (29.43%)

Table 6: Analysis of the test data sets (JRC-Acquis_L)

[13] We considered the NIST/BLEU implementation *mteval_v12*, as on `http://www.itl.nist.gov/iad/mig/tests/mt/2008/scoring.html`.

[14] TER (**t**ranslation **e**rror **r**ate.) as implemented on `http://www.cs.umd.edu/~snover/tercom/` - last accessed on 12.01.2010.

Corpus	No. of OOV-Words (% from voc. size)	Sentences in the corpus
RoGER_S **en – ro**		
Test	60 (11.49%)	37 (27.81%)
ro – en		
Test	84 (12.75%)	34 (25.56%)
ge – ro		
Test	101 (16.72%)	31 (23.30%)
ro – ge		
Test	84 (12.75%)	34 (25.56%)
JRC-Acquis_S **en – ro**		
Test	72 (7.25%)	38 (28.57%)
ro – en		
Test	129 (10.23%)	33 (24.81%)
ge – ro		
Test	171 (14.51%)	41 (30.82%)
ro – ge		
Test	160 (12.12%)	40 (30.07%)

Table 7: Analysis of the test data sets (RoGER and JRC-Acquis_S)

In the next subsection we will show more detailed the sensitivity of SMT systems to training and test data size and type.

3.2 Interpretation of the Results

In Table 4 we presented the variation of BLEU and TER scores across the three corpora. In (Koehn et al., 2003) a log-linear dependency between the size of the training corpora and the BLEU scores was observed. In contrast, our results cannot confirm this dependency for all language pairs investigated[15]. While for German-Romanian the log-linear dependency seem to be preserved, for English-Romanian the BLEU scores for JRC-Acquis_S are better than the ones for JRC-Acquis_L. Also worth to remark is that the BLEU scores for the other small corpus – ROGER_S –, are in the case of English-Romanian between the other two BLEU scores, and in the case of Romanian-English closer to the BLEU score for JRC-Acquis_S. This leads us to the conclusion that the hypothesis of log-linear dependency has to be tested before one decides to invest a lot of work in collecting large data sets. Giving the fact that in both of our experiments, as well as in (Koehn et al., 2003), the log-linear dependency was noticed in case of language pairs involving German, it could be an indication that the German specific morphological features, in special the dynamic word composition, could be a reason for this behavior. The high number of compounds in German may imply a higher data-sparseness, which can be compensated only through large amounts of training data.

Another interesting observation can be done regarding the TER Scores. The best TER scores were obtained, independent of the chosen language pair, for the ROGER_S corpus. One explanation is the particular syntax of this corpus: technical short sentences, in which the translation usually preserves the SL word order, as far as the syntax in both source and target languages allows. In contrast, in JRC-Acquis one finds often reformulations or shorter sentences. As TER measures the differences between output and reference translation in number of insertions, deletions and replacements, this may be cause of alternation of the TER scores.

Given the fact that the BLEU scores for the ROGER_S corpus are also in line with current state-of-the-art systems, we can conclude that for technical domains a small, manually corrected corpus can be successfully used for obtaining a reasonable translation output.

All the results we have presented reinforce the idea that SMT is fully dependent on the training and test data size and type and on the evaluation procedure. We will further show how dependent the results are to all the steps involved in the translation and evaluation processes by presenting the results in Table 8. We evaluated the results for the JRC-Acquis_S corpus, when no detokenization or recasing in the post-processing has been done. In contrast to the information from Table 4, in this last case, the translation evaluation scores are better. This shows that, next to the training and test data itself, sometimes pre- or post-processing steps affect (negatively) the evaluation scores.

Language Pair	BLEU	TER
en – ro	0.5359	0.3586
ro – en	0.5573	0.3279
ge – ro	0.3051	0.5808
ro – ge	0.3279	0.5796

Table 8: Results for JRC-Acquis_S (no recasing, no detokenization)

4 Conclusions

The results presented and discussed in this paper let us conclude that there is not always an a pri-

[15] We also do not exclude the difference in the results due also to different evaluation methodology. However, this aspect is not analyzed in this paper

ori size which can be recommended for developing a standard SMT systems independent of language pair and domain. The experiments we made showed (again) how dependent SMT results are on training and test data and on all processing steps. Especially for on-line applications which embed MT systems, where translation domain changes dynamically and a large number of language pairs is involved, a framework criteria for the training and test data is necessary. Our further work includes more experiments with different data (type and size) and language pairs. Also the associated statistical confidence intervals need to be calculated to have a better view on the evaluation results.

Acknowledgments

Ideas and results presented in this paper are part of Monica Gavrila's PhD research, conducted at the University of Hamburg, and of the ATLAS EU-Project (`www.atlasproject.eu`), supported through the ICT-PSP-Programme of the EU-Commission.

References

Thorsten Brants, Ashok C. Popat, Peng Xu, Franz J. Och, and Jeffrey Dean. 2007. Large language models in machine translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 858–867, Prague, Czech Republic.

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.

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

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

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 48–54, Morristown, NJ, USA. Association for Computational Linguistics.

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*, pages 177–180, Prague, Czech Republic, June.

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.

Juho Rousu. 2008. Workpackage 3 advanced language models. Online, January. SMART Project.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*, pages 223–231, August.

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

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

Bipin Suresh. 2010. Inclusion of large input corpora in statistical machine translation. Technical report, Stanford University.