Statistical Machine Translation with a Small Amount of Bilingual Training Data

Maja Popović, Hermann Ney

Lehrstuhl für Informatik VI - Computer Science Department
RWTH Aachen University
Ahornstrasse 55, 52056 Aachen, Germany
{popovic,ney}@informatik.rwth-aachen.de

Abstract

The performance of a statistical machine translation system depends on the size of the available task-specific bilingual training corpus. On the other hand, acquisition of a large high-quality bilingual parallel text for the desired domain and language pair requires a lot of time and effort, and, for some language pairs, is not even possible. Besides, small corpora have certain advantages like low memory and time requirements for the training of a translation system, the possibility of manual corrections and even manual creation. Therefore, investigation of statistical machine translation with small amounts of bilingual training data is receiving more and more attention. This paper gives an overview of the state of the art and presents the most recent results of translation systems trained on sparse bilingual data for two language pairs: Spanish-English, already widely explored with a number of (large) bilingual training corpora available, and Serbian-English - a rarely investigated language pair with restricted bilingual resources.

1. Introduction

The goal of this paper is to give an overview of the state of the art in statistical machine translation using a small amount of bilingual training data and to illustrate it with the most recent results obtained on the Spanish-English and Serbian-English language pairs.

2. Statistical Machine Translation with Sparse Bilingual Training Data

The goal of statistical machine translation is to translate a source language sequence into a target language sequence by maximising the posterior probability of the target sequence given the source sequence. In state-of-the-art translation systems, this posterior probability usually is modelled as a combination of several different models, such as: phrase-based models for both translation directions, lexicon models for both translation directions, target language model, phrase and word penalties, etc. Probabilities that describe correspondences between the words in the source language and the words in the target language are learned from a bilingual parallel text corpus and language model probabilities are learned from a monolingual text in the target language. Usually, the larger the available training corpus, the better the performance of a translation system. Whereas the task of finding appropriate monolingual text for the language model is not considered as difficult, acquisition of a large high-quality bilingual parallel text for the desired domain and language pair requires a lot of time and effort, and for some language pairs is not even possible. In addition, small corpora have certain advantages: the possibility of manual creation of the corpus, possible manual corrections of automatically collected corpus, low memory and time requirements for the training of a translation system, etc. Therefore, the strategies for exploiting limited amounts of bilingual data are receiving more and more attention. In the last five years various publications have dealt with the issue of sparse bilingual corpora.

(Al-Onaizan et al., 2000) report an experiment of Tetun-English translation with a small parallel corpus, although this work was not focused on the statistical approach. The translation experiment has been done by different groups including one using statistical machine translation. They found that the human mind is very capable of deriving dependencies such as morphology, cognates, proper names, etc. and that this capability is the crucial reason for the better results produced by humans compared to corpus based machine translation. If a program sees a particular word or phrase one thousand times during the training, it is more likely to learn a correct translation than if it sees it ten times, or never. Because of this, statistical translation techniques are less likely to work well when only a small amount of data is given.

(Callison-Burch and Osborne, 2003) propose a co-training method for statistical machine translation using the multilingual European Parliament corpus. Multiple translation models trained on different language pairs are used for producing new sentence pairs. They are then added to the original corpus and all translation models are retrained. The best improvements have been achieved after two or three training rounds.

In (Nießen and Ney, 2004) the impact of the training corpus size for stastistical machine translation from German into English is investigated, and the use of a conventional dictionary and morpho-syntactic information for improving the performance is proposed. They use several types of word reorderings as well as a hierarchical lexicon based on the POS tags and base forms of the German language. They report results on the full corpus of about sixty thousand sentences, on the very small part of the corpus containing five thousand sentences and on the conventional dictionary only. Morpho-syntactic information yields significant improvements in all cases and an acceptable translation quality is also obtained with the very small corpus.

Statistical machine translation of spontaneous speech with a training corpus containing about three thousand sentences has been dealt with in (Matusov et al., 2004). They propose acquiring additional training data using a n-gram coverage measure, lexicon smoothing and hierarchical lexicon structure for improving word alignments as well as several types of word reorderings based on POS tags.

Statistical machine translation of the Spanish-English and Catalan-English language pair with sparse bilingual resources in the tourism and travelling domain is investigated in (Popović and Ney, 2005). The use of a phrasal lexicon as an additional language resource is proposed as well as introducing expansions of the Spanish and Catalan verbs. With the help of the phrasal lexicon and morphological information, a reasonable translation quality is achieved with only one thousand sentence pairs from the domain.

The Serbian-English language pair is investigated in (Popović et al., 2005). A small bilingual corpus containing less than three thousand sentences was created and statistical machine translation systems were trained on different sizes of the corpus. The obtained translation results are comparable with results for other language pairs, especially if the small size of the corpus and rich inflectional morphology of the Serbian language are taken into account. Morpho-syntactic information is shown to be very helpful for this language pair.

Statistical machine translation of the Czech-English language pair and the impact of the morphological information are investigated in (Goldwater and McClosky, 2005). As with Serbian-English, morphological transformations have an important role for the translation quality.

The problem of creating word alignments for languages with scarce resources i.e. Romanian-English, Inuktikut-English and Hindi-English has been adressed in (Lopez and Resnik, 2005; Martin et al., 2005).

Training		Spanish	English
1.3M	Sentences	1281427	
	Running Words+PM	36578514	34918192
	Vocabulary	153124	106496
	Singletons [%]	35.2	36.2
13k	Sentences	13360	
	Running Words+PM	385198	366055
	Vocabulary	22425	16326
	Singletons [%]	47.6	43.7
1k	Sentences	1113	
	Running Words+PM	31022	29497
	Vocabulary	5809	4749
	Singletons [%]	60.8	55.3
dict.	Entries	52566	
	Running Words+PM	60964	62011
	Vocabulary	31126	30761
	Singletons [%]	67.7	67.4
Test	Sentences	840	1094
	Running Words+PM	22774	26917
	Distinct Words	4081	3958
	OOVs (1.3M) [%]	0.14	0.25
	OOVs (13k) [%]	2.8	2.6
	OOVs (1k) [%]	10.6	9.4
	OOVs (dict.) [%]	19.1	16.2

Table 1: Corpus statistics for the Spanish-English EPPS task (PM = punctuation marks)

3. Recent Translation Results

The translation system used in our most recent experiments with sparse training data is based on a log-linear combination of seven different models, the most important ones being phrase models (Vilar et al., 2005; Zens et al., 2005). For each language pair, several set-ups with different amount of bilingual data and several types of morpho-syntactic transformations were defined. The morpho-syntactic transformations have been implemented as a preprocessing step, therefore modifications of the training or search procedure were not necessary. For all experiments, the language model has been trained on the largest target language corpus because acquisition of monolingual data is not a particularly difficult issue. The evaluation metrics used for assessment of the systems are WER (Word Error Rate), PER (Position-independent word Error Rate) and BLEU (BiLingual Evaluation Understudy) (Papineni et al., 2002).

3.1. Spanish-English

The translation systems for this language pair are tested on the European Parliament Plenary Sessions (EPPS) corpus which is also used in the TC-Star project evaluation. A description of the corpus can be found in (Vilar et al., 2005). In order to investigate sparse training data scenarios, two sets of a small corpora have been constructed by random selection of sentences from the original corpus. The small corpus referred to as 13k contains about 1% of the original large corpus. The corpus referred to as 1k contains only 1000 sentences - such a corpus basically can be produced manually in relatively short time.

In addition, a conventional Spanish-English dictionary collected from the web which is not related to any particular task is used. The dictionary contains about fifty thousand entries and thirty thousand distinct words for each language.

The statistics for all corpora can be seen in Table 1. For the large corpus, the number of OOVs in the test is very small, much less than 1%. This number grows up to 10% by reducing the bilingual corpus, and for the dictionary alone it reaches 19% for Spanish and 16% for English.

Morpho-syntactic transformations: Adjectives in the Spanish language are usually placed after the corresponding noun, whereas for English it is the other way round. Therefore, for this language pair we applied local reorderings of nouns and adjective groups in the source language as described in (Popović and Ney, 2006). If the source language is Spanish, each noun is moved behind the corresponding adjective group. If the source language is English, each adjective group is moved behind the corresponding noun. An adverb followed by adjective (e.g. ”more important”) or two adjectives with a coordinate conjuntion in between (e.g. ”economic and political”) are treated as an adjective group. In addition, Spanish adjectives, in contrast to English, have four possible inflectional forms depending on gender and number. This might introduce additional data sparseness problems, especially if only a small amount of training data is available. Thus we replace all Spanish adjectives with their base forms.

Translation results: The following set-ups are defined for the Spanish-English language pair:

Spanish→English		WER	PER	BLEU
dict	baseline	60.4	49.3	19.4
	+reorder adjective	59.4	47.4	20.1
	+adjective base	56.4	46.8	23.8
1k	baseline	52.4	40.7	30.0
	+dictionary	48.0	36.5	36.0
	+reorder adjective	45.0	35.3	39.8
	+adjective base	44.5	34.8	40.9
13k	baseline	41.8	30.7	43.2
	+dictionary	40.6	29.6	46.3
	+reorder adjective	38.5	29.2	48.9
	+adjective base	38.3	29.0	49.6
1.3M	baseline	34.5	25.5	54.7
	+reorder adjective	33.5	25.2	56.4

Table 2: Translation results [%] for Spanish→English

English→Spanish		WER	PER	BLEU
dict	baseline	67.6	55.9	14.1
	+reorder adjective	66.3	55.2	15.7
	+align adjective base	65.7	54.5	16.5
1k	baseline	60.1	47.4	23.9
	+dictionary	56.0	43.2	28.3
	+reorder adjective	54.0	42.0	30.5
	+align adjective base	53.9	42.0	30.6
13k	baseline	49.6	37.4	36.2
	+dictionary	48.6	36.3	37.2
	+reorder adjective	47.4	36.0	38.6
	+align adjective base	47.3	35.7	39.1
1.3M	baseline	39.7	30.6	47.8
	+reorder adjective	39.6	30.5	48.3

Table 3: Translation results [%] for English→Spanish

- training only on a conventional dictionary (dict);
- training on a very small task-specific bilingual corpus (1k);
- training on a small task-specific bilingual corpus (13k);
- training on a large task-specific bilingual corpus (1.3M).

The language model for all set-ups is trained on the large corpus.

Table 2 presents the results for the translation from Spanish to English. It can be seen that the error rates of the system trained only on the dictionary are high and that morpho-syntactic transformations improve the performance. Although the final error rates are still high, they might be acceptable for tasks where only the gist of the translated text is needed, like for example document classification or multilingual information retrieval. Additional morpho-syntactic transformations such as treatment of Spanish verbs could further improve the performance.

When only a very small amount of task-specific bilingual parallel text is used (1k), all error rates are decreased and the BLEU score is increased in comparison to a system trained on the dictionary alone, although they are still rather high. Further, it can be seen that the dictionary is very helpful as an additional training corpus and the morpho-syntactic transformations have a significant impact so that the final error rates are reduced by about 15% relative in comparison to the baseline system. By increasing the size of the task-specific training corpus (13k) all error rates are further decreasing and can be further reduced with help of the dictionary and morpho-syntactic transformations.

The best results obtained with the large corpus are about 12% (relative) better than the best results with the small corpus (13k) and about 25% better in comparison with the very small corpus (1k). These differences seem to be very large, but we have to keep in mind how large the differences between the corpus size are, especially in terms of the time and effort necessary for collection and handling of large corpora.

It should be noted that the impact of a dictionary has not been tested for the full corpus since the corpus itself is sufficiently large. The improvements by replacing Spanish adjectives with their base forms are rather insignificant on this corpus and therefore are not reported.

The translation results for the other direction can be seen in Table 3. All error rates are higher due to the inflectional morphology of the Spanish language, and the effects of the training corpus size, dictionary and morpho-syntactic transformations are very similar. The improvements from the morpho-syntactic transformations are slightly smaller than for the translation into English due to the following reason: noun-adjective reordering is less important for the translation into Spanish because the adjective group is not always situated behind the noun. Therefore some reorderings in English are not really needed. As for the Spanish adjective inflections, for this translation direction alignment has been trained using adjective base forms, whereas the translation models have been trained on the original corpus. This enables better learning from the corpus to some extent, but finding a correct inflection of a Spanish adjective still remains relatively difficult.

3.2. Serbian-English

The Serbian-English parallel corpus used in our experiments is the electronic form of the Assimil language course described in (Popović et al., 2005). The full corpus is already rather small, containing about three thousand sentences and twenty five thousand running words. In order to investigate extremely sparse training material, a reduced corpus containing 200 sentences reffered to as 0.2k has been randomly extracted from the original corpus. For this corpus, a set of short phrases has been investigated as additional bilingual knowledge.

Table 4 presents the corpora statistics. It can be seen that even for the full corpus the number of OOVs is high, about 5% for English and almost 12% for Serbian (due to the rich inflectional morphology of this language). For the extremely small training corpus, the number of OOVs is about 3 to 4 times higher.

Morpho-syntactic transformations: The inflectional morphology of the Serbian language is very rich for all

Training		Serbian	English
2.6k	Sentences	2632	
	Running Words+PM	22227	24808
	Vocabulary	4546	2645
	Singletons [%]	60.0	45.8
0.2k	Sentences	200	
	Running Words+PM	1666	1878
	Vocabulary	778	603
	Singletons [%]	79.4	65.5
phrases	Entries	351	
	Running Words+PM	617	730
	Vocabulary	335	315
	Singletons [%]	71.3	66.3
Test	Sentences	260	
	Running Words+PM	2100	2336
	Distinct Words	891	674
	OOVs (2.6k) [%]	11.7	4.9
	OOVs (0.2k) [%]	35.2	21.8

Table 4: Corpus statistics for the Serbian-English Assimil task (PM = punctuation marks)

open word classes, but information contained in the inflection usually is not relevant for translation into English. Therefore, converting all Serbian words into their base forms is proposed. Nevertheless, inflections of Serbian verbs might contain relevant information about the person, which is especially important if the pronoun is omitted. Apart from this, there are three Serbian verbs which are negated by appending the negative particle to the verb as a prefix. Thus the following treatment of the Serbian verbs is applied: each verb is converted into a sequence of its base form and the part of the POS tag referring to a person. If the negative form is built by appending a prefix, the prefix i. e. the negative particle is separated.

For the other translation direction, since the articles are one of the most frequent word classes in English, but on the other hand there are no articles at all in Serbian, the articles are removed from the English corpus.

Translation results: For this language pair the following set-ups are defined:

- training on an extremely small task-specific bilingual corpus (0.2k);
- training on a small task-specific bilingual corpus (2.6k).

Since the largest available corpus is already small and the external phrase book is even smaller, we have not investigated translation using only the phrase book, but we used it as additional training material for the extremely sparse training corpus. The language model for all set-ups was trained on the full (2.6k) corpus.

Error rates for the translation from Serbian into English are shown in Table 5. As expected, the error rates of the system trained on an extremely small amount of parallel corpus are high. Performance of such a system is comparable with a system trained only on a conventional dictionary.

Serbian→English		WER	PER	BLEU
0.2k	baseline	65.5	60.8	8.3
	+phrases	65.0	59.8	10.3
	+base forms	59.2	54.8	13.9
	+verb POS+neg	60.0	52.6	14.8
2.6k	baseline	44.5	37.9	32.1
	+base forms	42.9	37.4	35.4
	+verb POS+neg	41.9	34.7	34.6

Table 5: Translation results [%] for Serbian→English

English→Serbian		WER	PER	BLEU
0.2k	baseline	73.4	68.4	6.8
	+phrases	71.9	67.5	9.3
	+remove article	66.7	62.2	9.4
2.6k	baseline	51.8	45.8	23.1
	+remove article	50.4	44.6	24.6

Table 6: Translation results [%] for English→Serbian

Adding short phrases is helpful to some extent, and replacing words with base forms has the most significant impact. Further improvements of PER and BLEU score are obtained by the verb treatment although WER is slightly deteriorated. Increasing the size of the bilingual training corpus to about three thousand sentences and applying morpho-syntactic transformations leads to an improvement of about 30% relative. Using a conventional dictionary and additional morpho-syntactic transformations could further improve the performance.

Table 6 shows results for the translation from English into Serbian. As expected, all error rates are significantly higher than for the other translation direction since the translation into the morphologically richer language always has poorer quality.

The importance of the phrases seems to be larger for this translation direction. Removing English articles improves the translation quality for both set-ups. As for the other translation direction, increasing the size of the training corpus results in up to 30% relative improvement.

4. Conclusion

Strategies for statistical machine translation with limited amount of bilingual training data are receiving more and more attention. Past and recent experiences have shown that an acceptable translation quality can be achieved with a very small amount of task-specific parallel text, especially if conventional dictionaries, phrasal books, as well as morpho-syntactic knowledge are available. Translation systems built only on a conventional dictionary or on extremely small task-specific corpora might be usefull for applications such as document classification or multilingual information retrieval.

5. Acknowledgement

The reported work is based on the projects that have been suported by the German Science Foundation (DFG) and by the European Project TC-Star (FP6-506738).

6. References

Yaser Al-Onaizan, Ulrich Germann, Ulf Hermjakob, Kevin Knight, P. Koehn, Daniel Marcu, and Kenji Yamada. 2000. Translating with scarce resources. In *The Seventeenth National Conf. on Artificial Intelligence*, pages 672–678, Austin, TX, July.

Chris Callison-Burch and Miles Osborne. 2003. Co-training for statistical machine translation. In *Proc. of the 6th Annual CLUK Research Colloquium*, Edinburgh, UK, January.

Sharon Goldwater and David McClosky. 2005. Improving stastistical machine translation through morphological analysis. In *Proc. of the Conf. on Empirical Methods for Natural Language Processing (EMNLP)*, Vancouver, Canada, October.

Adam Lopez and Philip Resnik. 2005. Improved hmm alignment for languages with scarce resources. In *43rd Annual Meeting of the Assoc. for Computational Linguistics: Proc. Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond*, pages 83–86, Ann Arbor, MI, June.

Joel Martin, Rada Mihalcea, and Ted Pedersen. 2005. Word alignment for languages with scarce resources. In *43rd Annual Meeting of the Assoc. for Computational Linguistics: Proc. Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond*, pages 65–74, Ann Arbor, MI, June.

Evgeny Matusov, Maja Popović, Richard Zens, and Hermann Ney. 2004. Statistical machine translation of spontaneous speech with scarce resources. In *Proc. of the Int. Workshop on Spoken Language Translation (IWSLT)*, pages 139–146, Kyoto, Japan, September.

Sonja Nießen and Hermann Ney. 2004. Statistical machine translation with scarce resources using morpho-syntactic information. *Computational Linguistics*, 30(2):181–204, June.

Kishore Papineni, Salim Roukos, Todd Ward, and Wie-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318, Philadelphia, PA, July.

Maja Popović and Hermann Ney. 2005. Exploiting phrasal lexica and additional morpho-syntactic language resources for statistical machine translation with scarce training data. In *Proc. of the 10th Annual Conf. of the European Association for Machine Translation (EAMT)*, pages 212–218, Budapest, Hungary, May.

Maja Popović and Hermann Ney. 2006. POS-based word reorderings for statistical machine translation. To appear in *Proc. of the Fifth Int. Conf. on Language Resources and Evaluation (LREC)*, Genova, Italy, May.

Maja Popović, David Vilar, Hermann Ney, Slobodan Jovičić, and Zoran Šarić. 2005. Augmenting a small parallel text with morpho-syntactic language resources for Serbian–English statistical machine translation. In *43rd Annual Meeting of the Assoc. for Computational Linguistics: Proc. Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond*, pages 41–48, Ann Arbor, MI, June.

David Vilar, Evgeny Matusov, Saša Hasan, Richard Zens, and Hermann Ney. 2005. Statistical machine translation of european parliamentary speeches. In *Proc. MT Summit X*, pages 259–266, Phuket, Thailand, September.

Richard Zens, Oliver Bender, Saša Hasan, Shahram Khadivi, Evgeny Matusov, Jia Xu, Yuqi Zhang, and Hermann Ney. 2005. The RWTH phrase-based statistical machine translation system. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, pages 155–162, Pittsburgh, PA, October.