

Improving target language modeling techniques for statistical machine translation

Maxim Khalilov

Departament de Teoria del Senyal i Comunicacions

Centre de Recerca TALP (UPC)

Barcelona 08034, Spain

khalilov@gps.tsc.upc.edu

Abstract

The aim of this study is to find ways of improving target language modeling (TLM) applied to statistical machine translation (SMT). We describe current research activities dedicated to TLM improvement that are applied to the 2007 n-gram-based statistical machine translation system developed in the TALP Research Center at the Technical University of Catalonia (UPC).

We consider two new language modeling improvement techniques: threshold-based TLM pruning and TLM based on statistical classes. Some of the research is still in progress. In this paper we describe some of the major problems faced and outline possible solutions and plans for future research.

We describe the results for the Spanish-English and English-Spanish language pairs from the official TC-STAR¹ 2006 evaluation.

1 Introduction

The statistical approach to machine translation is not a new area of scientific research and was actually one of the earliest fields in computer science. SMT lost popularity during the period from 1960 to 1980 but interest was renewed in the early 1990s and has grown rapidly in recent years due to the potential of the approach.

One of the major advances in the accuracy of translation systems was achieved by changing from early systems based on the *noisy channel* model, which performed word-to-word translation (Brown et al., 1993), to *phrase-based* systems, which are based on the same approach and use aligned bilingual corpora to translate bilingual units (Koehn et al., 2003; Zens et al., 2002).

The *n-gram-based* approach appeared during the same period (Mariño et al., 2006). The main difference between the two modern approaches can be found in the representation of bilingual units defined by word alignment (Crego et al., 2005a).

Language modeling is widely used in a large number of human language technology applications, including SMT. It can either be an integrated component or an additional feature depending on the approach on which the translation system is built. However, it significantly affects the system performance in both cases.

It is well known that LMs can often be very large and sometimes cause memory overflow problems. LM pruning is definitely required, but it reveals an efficiency-performance trade-off which generally causes decreased performance in smaller models. However, carefully determined pruning can reduce system noise and increase translation quality. In this study we consider a possible LM pruning strategy based on rational threshold selection.

We also focus on the use of the word class TLM as a feature of the log-linear model. The system introduces two types of word classes: statistical and linguistic.

The rest of the paper is organized as follows. In Section 2 we briefly outline the UPC n-gram translation system, system models, decoding and opti-

¹ TC-STAR (Technology and Corpora for Speech to Speech Translation). The project web-site is <http://www.tc-star.org>

mization procedures from 2006. In Section 3 we describe the framework of the TC-STAR 2006 evaluation (further details can be found on the ELDA web-site: www.elda.org/en/proj/tcstar-wp4). Section 4 contains the experimental results and in Section 5 we present the conclusions and explain plans for future research.

2 Statistical machine translation

SMT is based on the principle of translating a source sentence f (traditionally French) into a sentence in the target language e (English). The problem is formulated in terms of source and target languages and is defined as an *arg max* operation according to the following equation:

$$\hat{e}^I = \arg \max_{e^I} \{p(e^I | f^J)\} \quad (1)$$

where I and J represent number of words of the sentences in the target and source languages, respectively. Consequently, the translation problem can be reformulated as selecting a translation with the highest probability from among a set of target sentences.

Modern SMT developed from the work carried out by IBM in the early 1990s. This research was largely inspired by experiments in the field of speech recognition. The first SMT systems were based on the *noisy-channel* approach (Brown et al., 1990) and performed word-level translation. The previous equation can be decomposed according to the Bayes rule as follows:

$$\hat{e}^I = \arg \max_{e^I} \{p(f^J | e^I) \cdot p(e^I)\} \quad (2)$$

Therefore, the problem of finding conditional probability becomes an *arg max* operation of the product of two models:

- $P(f|e)$ refers to a bilingual translation model (BTM) probability.
- $P(e)$ refers to a target language model (TLM) probability.

Modern SMT models were enhanced by the maximum entropy approach (Berger et al., 1996) and implemented the posterior probability definition as a log-linear combination of the set of fea-

ture functions (Och and Ney, 2002). Using this technique, it is possible to combine additional feature models in the determination of the translation hypothesis, as shown below (3):

$$\hat{e}^I = \arg \max_{e^I} \left\{ \sum_{m=1}^M \lambda_m h_m(f^J, e^I) \right\} \quad (3)$$

where the feature functions h_m refer to the system models, i.e. BTM, TLM, etc., and λ_m represents the corresponding weights of these models.

2.1 N-gram-based translation model

The n-gram-based SMT system operates with bilingual units known as tuples (de Gispert and Mariño, 2002). The tuples are extracted from a word-to-word aligned bilingual corpus according to certain constraints (Crego et al., 2004).

The tuple n-gram translation model determines the joint probability of the source and target language units as shown in Equation (2):

$$P(e, f) = \prod_{k=1}^K P(t_k | t_{k-N+1}, \dots, t_{k-1}) \quad (4)$$

where t_k refers to the k^{th} tuple of the given bilingual sentence pair segmented into K tuples and N refers to the n-gram order.

The GIZA++ Toolkit was used to generate word-to-word alignments in source-to-target and target-to-source directions from a bilingual corpus (Och and Ney, 2000). Tuples are then extracted from these alignments. The tuples create a unique segmentation of the sentence pair (Crego et al., 2004). Figure 1 shows an example of tuple extraction from a bilingual sentence pair.



Figure 1. Tuples from a bilingual sentence pair.

The segmentation is unique and is defined by the word-to-word alignment. The n-gram-based approach is considered to be monotonous.

A detailed description of the SMT system that was used as a baseline in this paper can be found in

Mariño et al. (2006) and Costa-jussà and Fonollosa (2006).

2.2 Additional feature models

A translation system follows the maximum entropy approach and implements the log-linear combination of the BTM and five additional feature models:

Target language model

The following equation is used to calculate a standard n-gram TLM:

$$P_{LM}(t_k) \approx \prod_{n=1}^K p(w_n | w_{n-N+1}, \dots, w_{n-1}) \quad (5)$$

where t_k represents the partial translation hypothesis and w_n is the n -th word in this partially translated sentence.

The LM representation differs from a state-of-the-art, phrase-based SMT and a n-gram-based translation system in that it is an integrated component of a phrase-based system whereas LM is used as an additional feature in n-gram-based systems as a way of improving translation accuracy.

This additional feature was implemented in the 3-, 4- and 5-gram TLMs in this study.

Word penalty model

This feature was implemented in order to compensate for the system's preference for short output sentences – a phenomenon that is due to the TLM. Technically, the penalization depends on the total number of words in the partial translation hypothesis and can be determined as follows:

$$P_{wp}(t_k) = \exp(\text{number of words in } t_k) \quad (6)$$

Source-to-target lexicon model

This model uses word-to-word IBM Model 1 probabilities (Och et al., 2004) to estimate the lexical weights of each tuple according to the formula below:

$$P_{IBM1}((e, f)_n) = \frac{1}{(I+1)^J} \prod_{j=1}^J \sum_{i=0}^I p(e_n^i | f_n^j) \quad (7)$$

where f_n^j and e_n^i are the j -th and i -th words in the source and target parts of the tuple $(e, f)_n$, and J and I are the corresponding total numbers of words on either side of it. Giza++ word-to-word source-to-target alignment was used.

Target-to-source lexicon model

This backward lexicon model is the same as the previous model but for the opposite translation direction. We used Giza++ word-to-word target-to-source alignment.

Word class target language model

We introduced the novel feature in the final part of the study. A 5-gram model of the word class TLM was used as a method for reducing data sparseness. We considered two types of word classes: a 5-gram model of linguistic classes using part-of-speech (POS) tags and a 5-gram model of statistical classes extracted from the training corpus.

We used the TnT English POS tagger (Brants, 2000) and the FreeLing Spanish tagger (Carreras et al., 2004) for the monolingual corpus tagging (Popović and Ney, 2006).

The target statistical classes were extracted from the training corpus according to the algorithm outlined in (Och, 1999).

2.3 Word reordering

A linguistically motivated word reordering technique was used for Spanish-to-English translation in order to reduce the number of errors caused by the difference in word order between the two languages. A detailed description of the lexicalized reordering procedure can be found in (Costa-jussà and Fonollosa, 2006).

2.4 Decoding and optimization

The MARIE decoder was used as a search engine for the translation system, the details of which can be found in (Crego et al., 2005b). The decoder implements a beam-search algorithm with pruning capabilities. All of the additional feature models described above were taken into account in the decoding process.

Given the development set and references, the log-linear combination of weights can be adjusted

by the simplex optimization method (Nelder and Mead, 1965) to maximize the score function (see Eq. 1) according to the highest BLEU score (details can be found in Papineni et al., 2002). Experiments on the log-linear combination of BLEU and NIST scores are planned as part of our future research.

3 TC-STAR evaluation framework

The translation results reported as a baseline system were evaluated in the framework of the TC-STAR 2006 evaluation.

The data provided for shared tasks are from the European Parliament Plenary Sessions (EPPS) Final Text Edition (FTE) data set for English-Spanish and Spanish-English language pairs. The FTE condition corresponds to the official transcripts of the parliamentary sessions and is actually a written language translation condition.

Two reference translations were used for the development and test sets and the first 500 sentences of the official development corpora for both directions were used to maximize the score function. Basic corpus statistics are shown in Table 1.

EPPS	Spanish	English
Training set (EPPS-FTE)		
Sentences	1.3 M	1.3 M
Words	36.57 M	34.9 M
Vocabulary	153 K	107 K
Development set		
Sentences	500	500
Words	15 K	12 K
Vocabulary	2.5 K	2.3 K
Test set		
Sentences	699	1.155
Words	31 K	30 K
Vocabulary	3.9 K	4 K

Table 1. EPPS corpora (M = millions, K = thousands).

4 Experiments and results

The experiments presented in this section are divided into two groups: first we review the impact of TLM threshold pruning on translation accuracy and model size; we then enhance the translation system by incorporating TLM word classes and assess its performance.

4.1 TLM pruning experiments

The LM estimation procedure was performed using the SRI Language Modeling Toolkit (Stolke, 2002) which enables users to set a minimal count of n-grams included in the LM.

It is known that the n-gram order, i.e. n-gram *history length*, has a strong influence on the LM perplexity and the final translation score in a SMT application. We do not thoroughly investigate the impact of the reduction in perplexity, but it is obvious that a significant perplexity reduction will improve the translation accuracy.

We investigated 3-, 4- and 5-gram TLMs in the experiments performed.

The aim of threshold pruning is not to incorporate all of the n-grams that appear in the training corpus fewer times than a cut-off threshold value into the LM. We defined a set of threshold values for each n-gram order (in a “complete” system, the threshold would be 1 for all the n-grams). Obviously the unigram threshold is permanently set to 1 as we do not intend to reduce vocabulary.

Unfortunately, we were not able to perform experiments on the totally unpruned, high-ordered models due to the lack of memory resources and the translation decoder limitations. Consequently, the minimally pruned system configuration includes Threshold 2 for 4- and 5-gram LMs and Threshold 1 for low-order n-grams.

The experiments performed can essentially be considered as an attempt to constrain the n-gram vocabulary in order to reduce system noise and to accelerate the decoding process. The experimental results are shown in Tables 2 and 3.

We analyzed the translations generated by the SMT system and found the results obtained for the development corpus as a result of the model weight optimization (refer to Dev), while the final BLEU scores (case-insensitive) obtained for the test corpora (refer to Test) when the same system configurations and optimized model weights are considered.

SMT decoding can be considered a computationally intensive process in which model size is a crucial factor that can significantly influence the decoding time. The threshold setting considerably reduces the model size, while the case-insensitive BLEU score, which is only associated with the system performance measure in this study, remains constant or even increases.

N-gram order	Pruning threshold				BLEU		Model size, millions				
	2	3	4	5	Dev	Test	1	2	3	4	5
3	1	1	-	-	65.26	56.41	0.11	2.29	9.43	-	-
	1	2	-	-	65.18	56.84	0.11	2.29	2.95	-	-
	2	2	-	-	65.20	56.38	0.11	0.99	2.95	-	-
4	1	1	2	-	65.50	56.66	0.11	2.29	9.43	3.74	-
	1	2	2	-	65.59	56.81	0.11	2.29	2.75	3.74	-
	2	2	2	-	65.55	56.74	0.11	0.99	2.75	3.74	-
5	1	1	2	2	65.32	56.85	0.11	2.29	9.43	3.45	3.40
	1	2	2	2	65.10	56.82	0.11	2.29	2.75	3.45	3.40
	2	2	2	2	65.18	56.42	0.11	0.99	2.75	3.45	3.40

Table 2. Final BLEU score for Spanish-to-English translation and pruned target LMs.

N-gram order	Pruning threshold				BLEU		Model size, millions				
	2	3	4	5	dev	test	1	2	3	4	5
3	1	1	-	-	55.94	49.63	0.14	2.52	9.52	-	-
	1	2	-	-	55.38	50.23	0.14	2.52	2.93	-	-
	2	2	-	-	55.64	49.79	0.14	1.03	2.93	-	-
4	1	1	2	-	55.79	49.71	0.14	2.52	9.52	3.86	-
	1	2	2	-	56.07	50.13	0.14	2.52	2.68	3.86	-
	2	2	2	-	55.90	49.57	0.14	1.03	2.68	3.86	-
5	1	1	2	2	55.84	50.07	0.14	2.52	9.52	3.52	3.67
	1	2	2	2	55.54	49.69	0.14	2.52	2.68	3.52	3.67
	2	2	2	2	55.49	49.93	0.14	1.03	2.68	3.52	3.67

Table 3. Final BLEU score for English-to-Spanish translation and pruned target LMs.

We can see from the results that the TLM configuration that provides the best trade-off between BLEU score and model size in terms of BLEU score is the *4-gram* configuration with the thresholds set to 2 for 4- and 3-grams and 1 for 2- and unigrams (*4-2211* system configuration).

4.2 TLM based on word class experiments

The next step in our study deals with TLM word class implementation as a method for reducing data sparseness. The feature implements a 5-gram language model of target linguistic and statistical classes.

The tuple translation unit is redefined as a triplet which includes:

- A source sequence of words containing the source side of the bilingual tuple.
- A target string containing the target side of the tuple.
- A class string containing the linguistic or statistical class sequence corresponding to the words in the target string.

This information is only used in the decoding process in order to find an alternative class sequence associated with the competing partial-translation hypothesis. It is not directly used to calculate bilingual translation model probabilities (Mariño et al., 2006).

The TnT English tag set used contains 36 POS tags, while the FreeLing 1.5 Spanish tag set provides greater morphological diversity and 331 different tags. A total of 200 different statistical classes were extracted from the training corpus with the freely available software *mkcls* (Och, 1999).

Table 4 shows the translation results produced by the SMT systems, including the TLM that incorporates linguistic and statistical classes. The *4-2211* (see 4.1) system configuration refers to a baseline.

System configuration	<i>BLEU</i>	
	dev	test
Es-En baseline	65.59	56.81
Es-En POS TLM	65.62	56.78
Es-En Statistical TLM	65.84	56.99
En-Es baseline	56.07	50.13
En-Es POS TLM	56.11	50.05
En-Es Statistical TLM	56.32	50.27

Table 4. Effect of the supporting source POS-tags and statistical tags on translation accuracy.

One clear advantage provided by the statistical approach is that the statistical word classes do not depend on the language.

5 Conclusions and future work

In this paper we presented a study of possible simple methods for improving TLM. The baseline system is the translation system used for TC-STAR 2006 evaluation. We investigated two possible methods: simple experiments pertaining to the size of language models; and a linguistically and statistically motivated word class TLM.

There were no significant gains in the reported translation results due to the LM pruning threshold, but there was a significant reduction in the number of stored n-grams.

The main conclusion of this study is that n-gram-based systems are not very sensitive to the decrease in the cut-off value for the appearance of high-order n-grams in the LM. Furthermore, in general, recently appearing n-grams can be discounted to zero as they did not appear in the training corpus.

A more discouraging conclusion is that increasing the history length does not improve system performance. This is explained by the specific character of the n-gram-based system, which includes LM as an additional feature with variable weight. We will therefore perform further research into this phenomenon in a future study.

The second part of the study introduced a new feature: a TLM based on linguistic and statistical classes. There was no significant gain, but we did observe a slight improvement in performance when Spanish is set as the target language.

There is undoubtedly a great deal of work still to be done in this area. There are already several ideas for statistical and linguistic word class com-

binations as an additional feature, such as the factored language model (Kirchoff and Yang, 2005).

In contrast to the phrase-based LM factorization, in this study we analyzed the n-gram-based SMT system Moses². Moses is an open-source statistical machine translation system that applies automatically trained translation models to words which may have a factored representation. Factored modeling may be a good solution to the data sparseness problem because it provides a new, intelligent method for combining information sources. We plan to study the use of different morphologically or statistically determined classes as information sources.

Another idea that we intend to cover in future work is to apply the techniques described to the n-gram bilingual translation model, as it is ultimately a language model that deals with bilingual units.

6 Acknowledgments

This work was partially funded by the European Union under the integrated project TC-STAR – Technology and Corpora for Speech to Speech Translation (IST-2002-FP6-506738, <http://www.tc-star.org>) and by the Spanish Government under grant TEC2006-13964-C03 (AVIVAVOZ project).

References

- A. Berger, S. Della Pietra and V. J. Della Pietra, 1996, “A maximum entropy approach to natural language processing”, *Computational Linguistics*, Vol. 22, no. 1, pp. 39-72.
- T. Brants, 2000, “TnT – a statistical part-of-speech tagger”, *Proceedings of the Sixth Applied Natural Language Processing*.
- P. Brown, S. Della Pietra, V. J. Della Pietra and R. L. Mercer, 1993, “The mathematics of statistical machine translation: parameters estimation”, *Computational Linguistics*, Vol. 19, no. 2, pp. 263-311.
- X. Carreras, I. Chao, L. Padró and M. Padró, 2004, “Freeling: An open-source suite of language analyzers”, *Fourth Int. Conf. on Language Resources and Evaluation, LREC’04*.
- M. R. Costa-jussà and J. A. R. Fonollosa, 2006, “Statistical Machine Reordering”, *Proceedings of EMNLP 2006*, pp. 70-76.

² <http://www.statmt.org/moses/>

- J. M. Crego, J. B. Mariño and A. de Gispert, 2004, "Finite-state-based and phrase-based statistical machine translation", *Proc. of the Eighth Int. Conf. on Spoken Language Processing*, pp. 37-40.
- J. M. Crego, M. R. Costa-jussà, J. B. Mariño and J. A. R. Fonollosa, 2005, "N-gram-based versus Phrase-based Statistical Machine Translation", *In Proc. of the Int. Workshop on Spoken Language Translation - IWSLT 2005*, C-STAR Consortium, 2005a, pp. 177-184.
- J. M. Crego, J. B. Mariño and A. de Gispert, 2005, "An n-gram-based statistical machine translation decoder", *Proc. of the ninth Int. Conf. on Spoken Language Processing, ICSLP'05*, 2005b.
- A. de Gispert and J. B. Mariño, 2002, "Using x-grams for speech-to-speech translation", *Proc. of the Seventh Int. Conf. on Spoken Language Processing*.
- K. Kirchhoff and M. Yang, 2005, "Improved Language Modeling for Statistical Machine Translation", *Proceedings of the ACL Workshop on Building and Using Parallel Texts*.
- P. Koehn, F. J. Och and D. Marcu, 2003, "Statistical phrase-based translation", *Proc. of the 2003 Meeting of the North American chapter of the ACL*, Edmonton, Alberta.
- J. A. Nelder and R. Mead, 1965, "A simplex method for function minimization", *The Computer Journal*, Vol. 7, pp. 308-313.
- J. B. Mariño, R. Banchs, J. M. Crego, A. de Gispert, P. Lambert, J. A. R. Fonollosa, M. R. Costa-jussà and M. Khalilov, 2006, "UPC's Bilingual N-gram Translation System", *TC-Star Speech to Speech Translation Workshop (TC-Star'06/Wkshp)*, Barcelona (Spain).
- F. J. Och, 1999, "An Efficient Method for Determining Bilingual Word Classes", *Ninth Conf. of the Europ. Chapter of the Association for Computational Linguistics*, EACL'99, pp. 71-76, Bergen (Norway).
- F. J. Och, D. Gildea, S. Khudanpur, A. Sarkar, K. Yamada, A. Fraser, S. Kumar, L. Shen, D. Smith, K. Eng, V. Jain, Z. Jin and D. Radev, 2004, "A smorgasbord of features for statistical machine translation", *Proc. of the Human Language Technology Conference, HLT-NAACL'2004*, pp. 161-168..
- F. J. Och and H. Ney, 2000, "Improved statistical alignment models", *Proc. of the 38th Ann. Meeting of the ACL*, Hong Kong, China.
- F. J. Och and H. Ney, 2002, "Discriminative training and maximum entropy models for statistical machine translation", *40th Annual Meeting of the Association for Computational Linguistics*, pp. 295-302.
- K. Papineni, S. Roukos, T. Ward and W. J. Zhu, 2002, "Bleu: a method for automatic evaluation of machine translation", *Proc. of the 40th Ann. Conf. of the ACL*, Philadelphia, PA.
- M. Popović and H. Ney, 2006, "POS-based Word Reorderings for Statistical Machine Translation", *Fifth International Conference on Language Resources and Evaluation (LREC)*, pp. 1278-1283, Genoa (Italy).
- A. Stolcke, 2002, "SRILM: an extensible language modelling toolkit", *Proc. of the Int. Conf. on Spoken Language Processing*, pp. 901-904, Denver, CO.
- R. Zens, F. J. Och and H. Ney, 2004, "Improvements in phrase-based statistical machine translation", *Proc. of the HLT Conference, HLT-NAACL'2004*, pp. 257-264.