Tuning a phrase-based statistical translation system for the IWSLT 2005 Chinese to English and Arabic to English tasks

Marta R. Costa-jussà and José A. R. Fonollosa

TALP Research Center
Universitat Politècnica de Catalunya, Barcelona
{mruiz,adrian}@gps.tsc.upc.edu

Abstract

Nowadays, most of the statistical translation systems are based on phrases (i.e. groups of words). We describe a phrase-based system using a modified method for the phrase extraction which deals with larger phrases while keeping a reasonable number of phrases. Also, different alignments to extract phrases are allowed and additional features are used which lead to a clear improvement in the performance of translation. Finally, the system manages to do reordering. We report results in terms of translation accuracy by using the BTEC corpus in the tasks of Chinese to English and Arabic to English, in the framework of IWSLT'05 evaluation.

1. Introduction

From the initial word-based translation models [3], research on statistical machine translation has been strongly improved. At the end of the last decade the use of context in the translation model (phrase-based approach) supposed a clear improvement in translation quality ([17], [16], [8]).

Statistical Machine Translation (SMT) is based on the assumption that every sentence e in the target language is a possible translation of a given sentence f in the source language. The main difference between two possible translations of a given sentence is a probability assigned to each, which has to be learned from a bilingual text corpus. Thus, the translation of a source sentence *f* can be formulated as the search of the target sentence *e* that maximizes the translation probability $P(e|f)$,

$$\tilde{e} = \underset{e}{argmax}\, P(e|f) \tag{1}$$

If we use Bayes rule to reformulate the translation probability, we obtain,

$$\tilde{e} = \underset{e}{argmax}\, P(f|e)P(e) \tag{2}$$

This translation model is known as the source-channel approach [2] and it consists on a language model $P(e)$ and a separate translation model $P(f|e)$ [6].

In the last few years, new systems tend to use sequences of words, commonly called phrases [7], aiming at introducing word context in the translation model. As alternative to the source-channel approach the decision rule can be modeled through a log-linear maximum entropy framework.

$$\tilde{e} = \underset{e}{argmax} \left\{ \sum_{m=1}^{M} \lambda_m h_m(e, f) \right\} \tag{3}$$

The features functions, h_m, are the system models (translation model, language model and others) and weights, λ_i, are typically optimized to maximize a scoring function [12]. It is derived from the Maximum Entropy approach as shown in [1] and has the advantage that additional features functions can be easily integrated in the overall system.

This paper addresses a modification of the phrase-extraction algorithm in [13] and results in Chinese to English and Arabic to English tasks are reported. It also combines several alignments before extracting phrases and interesting features. It is organized as follows. Section 2 explains the SMT system: the phrase extraction, its modification and shows the different features which have been taken into account and, briefly, the decoding; section 3 presents the evaluation framework and the results in Chinese to English and Arabic to English tasks are reported; and the final section shows some conclusions on the experiments and in the evaluation of IWSLT'05.

2. SMT system

As explained in the introduction, the SMT system which is presented is modeled through a log-linear maximum entropy framework. In this section, we explain the models, the feature functions and the decoding that build this system.

The Translation Model is based on bilingual phrase (or phrases). A bilingual unit consists of two monolingual fragments, where each one is supposed to be the translation of its counterpart. During training, the system learns a dictionary of these bilingual fragments, the actual core of the translation systems.

2.1. Phrase-based Translation Model

The basic idea of phrase-based translation is to segment the given source sentence into phrases, then translate each phrase and finally compose the target sentence from these phrase translations [18].

2.1.1. Word alignment

Given a sentence pair, we use GIZA++ [10] to align each of them word-to-word. We can train in both translation directions and we obtain: (1) the alignment in the source to target direction ($s2t$); and (2) the alignment in the target to source direction. If we compose the union of both alignments (sUt), we get a higher recall and a lower precision of the combined alignment.

2.1.2. Phrase-extraction

Phrases are extracted from sentence pairs and theirs corespondents word alignments following the criterion in [13] and the modification in phrase length in [4]. A phrase is any pair of m source words and n target words that satisfies two basic constraints:

1. Words are consecutive along both sides of the bilingual phrase,

2. No word on either side of the phrase is aligned to a word out of the phrase.

It is unfeasible to build a dictionary with all the phrases. That is why we limit the maximum size of any given phrase. Also, the huge increase in computational and storage cost of including longer phrases does not provide a significant improve in quality [7] as the probability of reappearance of larger phrases decreases.

In our system we considered two length limits.The length of a monolingual phrase is defined as its number of words. The length of a phrase is the greatest of the lengths of its monolingual phrases. We first extract all the phrases of length X or less. Then, we also add phrases up to length Y (Y greater than X) if they cannot be generated by smaller phrases. Basically, we select additional phrases with source words that otherwise would be missed because of cross or long alignments [4].

Given the collected phrase pairs, we estimate the phrase translation probability distribution by relative frequency.

$$P(f|e) = \frac{N(f,e)}{N(e)} \tag{4}$$

where N(f,e) means the number of times the phrase *f* is translated by *e*. If a phrase *e* has $N > 1$ possible translations, then each one contributes as *1/N* [18].

2.2. Additional features

- Firstly, we consider the target language model. It actually consists of an n-gram model, in which the probability of a translation hypothesis is approximated by the product of word n-gram probabilities:

$$p(T_k) \approx \prod_{n=1}^{k} p(w_n|...w_{n-3}, w_{n-2}, w_{n-1}) \tag{5}$$

  where T_k refers to the partial translation hypothesis and w_n to the n^{th} word in it.

- As translation model we use the conditional probability. Note that no smoothing is performed, which may cause an overestimation of the probability of rare phrases. This is specially harmful given a bilingual phrase where the source part has a big frequency of appearance but the target part appears rarely. That is why we use the posterior phrase probability, we compute again the relative frequency but replacing the count of the target phrase by the count of the source phrase [11].

$$P(e|f) = \frac{N'(f,e)}{N(f)} \tag{6}$$

  where N'(f,e) means the number of times the phrase *e* is translated by *f*. If a phrase *f* has $N > 1$ possible translations, then each one contributes as *1/N*.

  Adding this feature function we reduce the number of cases in which the overall probability is overestimated.

- The following two feature functions correspond to a forward and backward lexicon models. These models provides lexicon translation probabilities for each tuple based on the word-to-word IBM model 1 probabilities [11]. These lexicon models are computed according to the following equation:

$$p((t,s)_n) = \frac{1}{(I+1)^J} \prod_{j=1}^{J} \sum_{i=0}^{I} p_{IBM1}(t_n^i|s_n^j) \tag{7}$$

  where s_n^j and t_n^i are the j^{th} and i^{th} words in the source and target sides of tuple $(t,s)_n$, being J and I the corresponding total number words in each side of it.

  For computing the forward lexicon model, IBM model 1 probabilities from GIZA++ source-to-target alignments are used. In the case of the backward lexicon model, GIZA++ target-to-source alignments are used instead.

- We consider the widely used word penalty model. This function introduces a sentence length penalization in order to compensate the system preference for short output sentences. This penalization depends on the total number of words contained in the partial translation hypothesis, and it is computed as follows:

$$wp(T_k) = exp(\text{number of words in } T_k) \quad (8)$$

where, again, T_k refers to the partial translation hypothesis.

- Finally, the last feature is the phrase penalty [18] which is a constant cost per produced phrase. Here, a negative weight, which means reducing the costs per phrase, results in a preference for adding phrases. Alternatively, by using a positive scaling factors, the system will favor less phrases.

2.3. Decoding

In SMT decoding, translated sentences are built incrementally from left to right in form of hypotheses, allowing for discontinuities in the source sentence.

A Beam search algorithm with pruning is used to find the optimal path. The search is performed by building partial translations (hypotheses), which are stored in several lists. These lists are pruned out according to the accumulated probabilities of their hypotheses.

Worst hypotheses with minor probabilities are discarded to make the search feasible. Also the decoder allows reordering. The use of the reordering strategies suppose a necessary trade-off between quality and efficiency. That is why two reordering strategies are used:

- A distortion limit (m). A source word (phrase or tuple) is only allowed to be reordered if it does not exceed a distortion limit, measured in words.

- A reorderings limit (j). Any translation path is only allowed to perform j reordering jumps.

See [5] for further details.

3. Evaluation Framework

3.1. Corpus Statistics

Experiments have been carried out in two tasks of the IWSLT'05 evaluation[1]: Chinese to English (BTEC Corpus [15]) and Arabic to English.

The BTEC is a small corpus translation task. Table 1 shows the main statistics of the used data, namely number of sentences, words, vocabulary, and mean sentence lengths for each language.

[1] www.slt.atr.jp/IWSLT2005

BTEC	*Chinese*	*English*
Training Sentences	20 k	20 k
Words	176.2 k	182.3 k
Vocabulary	8.7 k	7.3 k
Development Sentences	1006	1006
Words	7.3 k	6 k
Vocabulary	1.4 k	1.3 k
Test Sentences	506	506
Words	3.7 k	-
Vocabulary	963	-

Table 1: *Chinese to English task. BTEC Corpus: Training, Development and Test data sets. The Development data set has 16 references, (k stands for thousands)*

BTEC	*Arabic*	*Arabic'*	*English*
Training Sentences	20 k	20 k	20 k
Words	131.7 k	180.5 k	182.3 k
Vocabulary	25.2 k	16 k	7.3 k
Development Sentences	1006	1006	1006
Words	5.3 k	7.2 k	6 k
Vocabulary	2.4 k	1.9 k	1.3 k
Test Sentences	506	506	506
Words	2.6 k	3.6 k	-
Vocabulary	1.4 k	1.2 k	-

Table 2: *Arabic to English task. There are shown both the original Arabic and the Arabic' (re-tokenized) statistics. BTEC Corpus: Training, Development and Test data sets. The Development data set has 16 references, (k stands for thousands)*

At the same time, Table 2 shows the same statistics, but for the Arabic to English task. The Arabic', which is also showed in the statistics, has been preprocessed. The preprocessing stage was only performed on the Arabic side of the corpus, and apart from standard punctuation marks, it aims at separating prefixes (such as the article) that highly increase the vocabulary size. In detail, we produce a hard separation of all words starting by ال and بال (as ب + ال), in order to separate articles from words. Note that this process is neither informed (it does not use any tagging software) nor complete (several other Arabic particles are usually attached to words). However, it already produces a significative vocabulary reduction leading to improved performance.

3.2. Units

We used GIZA++ to perform the word alignment of the whole training corpus. We use the union alignment and, as an improvement, we add the alignment in the source to target direction to the union alignment (hereinafter, sAt), which seems to reach better accuracy in translation (as

Alignment	Chinese to English	Arabic to English
sUt	122.9 k	186.2 k
s2t	288.8 k	421.2 k
sAt	294.4 k	424.2 k

Table 3: *Vocabulary of phrases for each alignment (source to target, union and the addition of both) and for each task. The phrases parameters are X=4 and Y=7 for Chinese phrases and X=5 and Y=7 for Arabic sentences (this parameters are studied in next subsection).*

we will see in the following subsection). In fact, the algorithm of phrase-extraction obtains a higher vocabulary when using the source to target alignment (see Table 3) as there are less cross words and more phrases follow the rule of not having aligned words out of the phrase.

In the Chinese to English task, we experiment with the phrases' length as seen in Table 4. We compare them by building the baseline with each set of phrases. The models in the baseline are: translation model, language model, word penalty, phrase penalty, IBM1 in both directions and reordering (using $m = 5$ and $j = 3$). We reach the best result in BLEU while extracting phrases up to length 4 (X) and, in addition, those phrases up to length 7 (Y) which could not be generated by smaller phrases.

We observe that the number of phrases when using both lengths (X and Y) does not grow up as quickly as when using only one length. In fact, it keeps similar to the size of the smaller length (X), while the accuracy in translation has been improved.

In the case of Arabic to English, Table 5 shows the equivalent comparison. Here, we extract phrases up to length 5 (X) and, in addition, the phrases up to length 7 (Y) which could not be generated by smaller phrases.

As default language model feature, we use a standard word-based 4gram language model generated with smoothing Kneser-Ney and interpolation of higher and lower order ngrams (by using SRILM [14]).

3.3. Experiments

The evaluation in the BTEC task has been carried out using references and translations in lowercase and without punctuation marks. We applied the widely used algorithm SIMPLEX to optimize the different weights (using the development set) [9]. Results in the test set with 16 references are reported.

The experiments in Table 6 correspond to the Chinese to English translation task under the phrase-based SMT system. The baseline considers the models and the phrase lengths mentioned in the subsection above. The improved system considers both the phrases extracted from the source to target alignment and the union alignment, and, also, adds the posterior probability feature. Here, the posterior probability seems not to add anything to the system with only sAt.

The experiments in Table 7 correspond to the Arabic to English translation task under the phrase-based SMT system. The baseline considers again the models and the phrase lengths mentioned in the subsection above. Note that in this case the posterior probability feature function combined with the inclusion of the phrases from the additional alignment, makes the translation more accurate. The inclusion of posterior probability provides a significant increase in performance in this case because the $P(f|e)$ tends to be more overestimated in phrases that come from the source to target alignment.

4. Conclusions

We reported a phrase-based system. The translation model is set in the log-linear maximum entropy framework, and uses several features functions. Finally, the decoder which is based on a beam search allows for distortion.

This phrase-based system has been improved in different ways: the alignment (sAt) used outperforms the union alignment when using the additional feature of posterior probability; and the variation in phrase length allows better results while keeping reasonable the number of phrases.

As future work, we will analyze the difference in behaviors between both tasks in order to propose a more accurate optimizer and a more complex combination of features functions (instead of the linearity).

5. Acknowledgments

This work has been 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), the Spanish government, under grant TIC-2002-04447-C02 (Aliado Project), Universitat Politècnica de Catalunya and the TALP Research Center under TALP-UPC-RECERCA grant.

The authors want to thank Josep M. Crego, José B. Mariño, Adrià de Gispert, Patrik Lambert and Rafael E. Banchs (members of the TALP Research Center) for their contribution to this work.

6. References

[1] A. Berger, S. Della Pietra, and V. Della Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–72, March 1996.

[2] P. Brown, J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, J.D. Lafferty, R. Mercer, and P.S. Roossin. A statistical approach to machine trans-

X	Y	SIZE	mWER	BLEU	NIST	PER
3	3	220.7 k	48.11	43.26	8.312	38.65
4	4	268.7 k	47.88	43.46	8.337	39
5	5	309 k	48.11	43.51	8.491	39.16
4	7	275.6 k	47.75	43.47	8.356	38.89

Table 4: *Analysis of the parameters of phrase length (X, Y) in the Chinese to English task using the union alignment. Each option shows its size (number of phrases extracted) and its bleu optimized and evaluated in the development set (considering the baseline)*

X	Y	SIZE	mWER	BLEU	NIST	PER
4	4	285.7 k	38.05	52.95	9.093	33.02
5	5	337.4 k	38.01	53.04	9.124	32.96
6	6	381.6 k	37.83	53.46	9.154	32.93
5	7	340 k	37.86	53.61	9.098	32.75

Table 5: *Analysis of the parameters of phrase length (X, Y) in the Arabic to English task using the union alignment. Each option shows its size (number of phrases extracted) and its bleu optimized and evaluated in the development set (considering the baseline)*

Phrase-based	mWER	BLEU	NIST	PER
Baseline (X=4, Y=7)	47.75	43.47	8.356	38.89
Baseline (X=4, Y=7) + P(e|f)	46.73	44.22	7.6602	37.80
Baseline (X=4, Y=7) + sAt	45.69	45.68	7.9603	37.88
Baseline (X=4, Y=7) + (P(e|f) + sAt)	45.91	45.23	7.974	37.96

Table 6: *Results for the Chinese to English translation task using different features. The last row shows the best system*

Phrase-based	mWER	BLEU	NIST	PER
Baseline (X=5, Y=7)	37.86	53.61	9.098	32.75
Baseline (X=5, Y=7) + P(e|f)	38.04	53.58	9.1601	32.33
Baseline (X=5, Y=7) + sAt	36.64	55.87	9.5561	30.62
Baseline (X=5, Y=7) + (P(e|f) + sAt)	35.0	57.26	9.331	30.30

Table 7: *Results for the Arabic to English translation task using the phrase-based translation model and different features. The last row shows the best system*

lation. *Computational Linguistics*, 16(2):79–85, 1990.

[3] P. Brown, S. Della Pietra, V. Della Pietra, and R. Mercer. The mathematics of statistical machine translation. *Computational Linguistics*, 19(2):263–311, 1993.

[4] M. R. Costa-jussà and J.A.Rodriguez Fonollosa. Improving the phrase-based statistical translation by modifying phrase extraction and including new features. *Proceedings of the ACL Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond*, June 2005.

[5] J. M. Crego, J. Mariño, and A. de Gispert. An ngram-based statistical machine translation decoder. *EUROSPEECH 05*, September 2005.

[6] I. García Varea. *Traducción automática estadística: modelos de traducción basados en máxima entropía y algoritmos de búsqueda*. PhD Thesis in Informatics, Dep. de Sistemes Informàtics i Computació, Universitat Politècnica de València, 2003.

[7] P. Koehn, F.J. Och, and D. Marcu. Statistical phrase-based translation. *Proc. of the Human Language Technology Conference, HLT-NAACL'2003*, May 2003.

[8] D. Marcu and W. Wong. A phrase-based, joint probability model for statistical machine translation. *Proc. of the Conf. on Empirical Methods in Natural Language Processing, EMNLP'02*, pages 133–139, July 2002.

[9] J.A. Nelder and R. Mead. A simplex method for function minimization. *The Computer Journal*, 7:308–313, 1965.

[10] F.J. Och. Giza++ software. http://www-i6.informatik.rwth-aachen.de/~och/ software/giza++.html. 2003.

[11] 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. A smorgasbord of features for statistical machine translation. *Proc. of the Human Language Technology Conference, HLT-NAACL'2004*, pages 161–168, May 2004.

[12] F.J. Och and H. Ney. Discriminative training and maximum entropy models for statistical machine translation. *40th Annual Meeting of the Association for Computational Linguistics*, pages 295–302, July 2002.

[13] F.J. Och and H. Ney. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417–449, December 2004.

[14] A. Stolcke. Srilm - an extensible language modeling toolkit. *Proc. of the 7th Int. Conf. on Spoken Language Processing, ICSLP'02*, September 2002.

[15] T. Takezawa, E. Sumita, F. Sugaya, H Yamamoto, and S. Yamamoto. Toward a broad-coverage bilingual curpus for speech translation of travel conversations in the real world. *LREC 2002*, pages 147–152, May 2002.

[16] K. Yamada and K. Knight. A syntax-based statistical translation model. *39th Annual Meeting of the Association for Computational Linguistics*, pages 523–530, July 2001.

[17] R. Zens, F.J. Och, and H. Ney. Phrase-based statistical machine translation. In M. Jarke, J. Koehler, and G. Lakemeyer, editors, *KI - 2002: Advances in artificial intelligence*, volume LNAI 2479, pages 18–32. Springer Verlag, September 2002.

[18] R. Zens, F.J. Och, and H. Ney. Improvements in phrase-based statistical machine translation. *Proc. of the Human Language Technology Conference, HLT-NAACL'2004*, pages 257–264, May 2004.