Improving Phrase-Based Statistical Translation by modifying phrase extraction and including several features

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

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

Abstract

Nowadays, most of the statistical translation systems are based on phrases (i.e. groups of words). In this paper we study different improvements to the standard phrase-based translation system. We describe a modified method for the phrase extraction which deals with larger phrases while keeping a reasonable number of phrases. We also propose additional features which lead to a clear improvement in the performance of the translation. We present results with the EuroParl task in the direction Spanish to English and results from the evaluation of the shared task "Exploiting Parallel Texts for Statistical Machine Translation" (ACL Workshop on Parallel Texts 2005).

1 Introduction

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 $P(e|f)$,

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

[0]This work has been supported by the European Union under grant FP6-506738 (TC-STAR project).

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 [1] and it consists on a language model $P(e)$ and a separate translation model $P(f|e)$ [5].

In the last few years, new systems tend to use sequences of words, commonly called phrases [8], 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 weigths, λ_i, are typically optimized to maximize a scoring function. It is derived from the Maximum Entropy approach suggested by [13] [14] for a natural language understanding task. It has the advantatge that additional features functions can be easily integrated in the overall system.

This paper addresses a modification of the phrase-extraction algorythm in [11]. It also combines several interesting features and it reports an important improvement from the baseline. It is organized as follows. Section 2 introduces the baseline; the following section explains the modification in the phrase extraction; section 4 shows the different features which have been taken into account; section 5 presents the evaluation framework; and

the final section shows some conclusions on the experiments in the paper and on the results in the shared task.

2 Baseline

The baseline is based on the source-channel approach, and it is composed of the following models which later will be combined in the decoder.

The Translation Model. It is based on bilingual phrases, where a bilingual phrase (BP) is simply two monolingual phrases (MP) in which each one is supposed to be the translation of each other. A monolingual phrase is a sequence of words. Therefore, 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 [17].

During training, the system has to learn a dictionary of phrases. We begin by aligning the training corpus using GIZA++ [6], which is done in both translation directions. We take the union of both alignments to obtain a symmetrized word alignment matrix. This alignment matrix is the starting point for the phrase based extraction.

Next, we define the criterion to extract the set of BP of the sentence pair $(f_{j_1}^{j_2}; e_{i_1}^{i_2})$ and the alignment matrix $A \subseteq J*I$, which is identical to the alignment criterion described in [11].

$$BP(f_1^J, e_1^I, A) = \{(f_{j_1}^{j_2}, e_{i_1}^{i_2}) :$$

$$\forall (j,i) \epsilon A : j_1 \leq j \leq j_2 \leftrightarrow i_1 \leq i \leq i2$$

$$\wedge \exists (j,i) \epsilon A : j_1 \leq j \leq j_2 \wedge i_1 \leq i \leq i2\}$$

The set of BP is consistent with the alignment and consists of all BP pairs where all words within the foreign language phrase are only aligned to the words of the English language phrase and viceversa. At least one word in the foreign language phrase has to be aligned with at least one word of the English language. Finally, the algorithm takes into account possibly unaligned words at the boundaries of the foreign or English language phrases.

The target language model. It is combined with the translation probability as showed in equation (2). It gives coherence to the target text obtained by the concatenated phrases.

3 Phrase Extraction

Motivation. The length of a MP is defined as its number of words. The length of a BP is the greatest of the lengths of its MP.

As we are working with a huge amount of data (see corpus statistics), it is unfeasible to build a dictionary with all the phrases longer than length 4. Moreover, the huge increase in computational and storage cost of including longer phrases does not provide a significant improve in quality [8].

X-length In our system we considered two length limits. We first extract all the phrases of length 3 or less. Then, we also add phrases up to length 5 if they cannot be generated by smaller phrases. Empirically, we chose 5, as the probability of reappearence of larger phrases decreases.

Basically, we select additional phrases with source words that otherwise would be missed because of cross or long alignments. For example, from the following sentence,

Cuando el Parlamento Europeo , que tan frecuentemente insiste en los derechos de los trabajadores y en la debida protección social , (...)

NULL ( ) When (1) the (2) European (4) Parliament (3 4) , (5) that (6) so (7) frequently (8) insists (9) on (10) workers (11 15) ' (14) rights (12) and (16) proper (19) social (21) protection (20) , (22) (...)

where the number inside the clauses is the aligned word(s). And the phrase that we are looking for is the following one.

los derechos de los trabajadores # workers ' rights

which only could appear in the case the maximum length was 5.

4 Phrase ranking

4.1 Conditional probability $P(f|e)$

Given the collected phrase pairs, we estimated the phrase translation probability distribution by relative frecuency.

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

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* [17].

Note that no smoothing is performed, which may cause an overestimation of the probability of rare phrases. This is specially harmful given a BP where the source part has a big frecuency of appearence but the target part appears rarely. For example, from our database we can extract the following BP: "you # la que no", where the English is the source language and the Spanish, the target language. Clearly, "la que no" is not a good translation of "you", so this phrase should have a low probability. However, from our aligned training database we obtain,

$$P(f|e) = P(\text{you}|\text{la que no}) = 0.23$$

This BP is clearly overestimated due to sparseness. On the other, note that "la que no" cannot be considered an unusual trigram in Spanish. Hence, the language model does not penalise this target sequence either. So, the total probability $(P(f|e)P(e))$ would be higher than desired.

In order to somehow compensate these unreiliable probabilities we have studied the inclusion of the posterior [12] and lexical probabilities [1] [10] as additional features.

4.2 Feature $P(e|f)$

In order to estimate 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.

$$P(e|f) = \frac{N'(f,e)}{N(f)} \quad (5)$$

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. This results in an important improvement in translation quality.

4.3 IBM Model 1

We used IBM Model 1 to estimate the probability of a BP. As IBM Model 1 is a word translation and it gives the sum of all possible alignment probabilities, a lexical co-ocurrence effect is expected. This captures a sort of semantic coherence in translations.

Therefore, the probability of a sentence pair is given by the following equation.

$$P(f|e; M1) = \frac{1}{(I+1)^J} \prod_{j=1}^{J} \sum_{i=0}^{I} p(f_j|e_i) \quad (6)$$

The $p(f_j|e_i)$ are the source-target IBM Model 1 word probabilities trained by GIZA++. Because the phrases are formed from the union of source-to-target and target-to-source alignments, there can be words that are not in the $P(f_j|e_i)$ table. In this case, the probability was taken to be 10^{-40}.

In addition, we have calculated the IBM^{-1} Model 1.

$$P(e|f; M1) = \frac{1}{(J+1)^I} \prod_{I=1}^{I} \sum_{j=0}^{J} p(e_i|f_j) \quad (7)$$

4.4 Language Model

The English language model plays an important role in the source channel model, see equation (2), and also in its modification, see equation (3). The English language model should give an idea of the sentence quality that is generated.

As default language model feature, we use a standard word-based trigram language model generated with smoothing Kneser-Ney and interpolation (by using SRILM [16]).

4.5 Word and Phrase Penalty

To compensate the preference of the target language model for shorter sentences, we added two

	Spanish	*English*
Train Sentences	1223443	1223443
Words	34794006	33379333
Vocabulary	168685	104975
Dev Sentences	504	504
Words	15353	15335
OOV	25	16
Test Sentences	504	504
Words	10305	10667
OOV	36	19

Table 1: *Statistics of training and test corpus*

simple features which are widely used [17] [7]. The word penalty provides means to ensure that the translations do not get too long or too short. Negative values for the word penalty favor longer output, positive values favor shorter output [7].

The phrase penalty 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.

5 Evaluation framework

5.1 Corpus Statistics

Experiments were performed to study the effect of our modifications in the phrases. The training material covers the transcriptions from April 1996 to September 2004. This material has been distributed by the European Parlament. In our experiments, we have used the distribution of RWTH of Aachen under the project of TC-STAR [1]. The test material was used in the first evaluation of the project in March 2005. In our case, we have used the development divided in two sets. This material corresponds to the transcriptions of the sessions from October the 21st to October the 28th. It has been distributed by ELDA[2]. Results are reported for Spanish-to-English translations.

[1] http://www.tcstar.org/
[2] http://www.elda.org/

5.2 Experiments

The decoder used for the presented translation system is reported in [2]. This decoder is called MARIE and it takes into account simultaneously all the 7 features functions described above. It implements a beam-search strategy.

As evaluation criteria we use: the Word Error Rate (WER), the BLEU score [15] and the NIST score [3].

As follows we report the results for several experiments that show the performance of: the baseline, adding the posterior probability, IBM Model 1 and IBM1^{-1}, and, finally, the modification of the phrases extraction.

Optimisation. Significant improvements can be obtained by tuning the parameters of the features adequately. In the complet system we have 7 parameters to tune: the relatives frecuencies $P(f|e)$ and $P(e|f)$, IBM Model 1 and its inverse, the word penalty, the phrase penalty and the weight of the language model. We applied the widely used algorithm SIMPLEX to optimise [9]. In Table 2 (line 5th), we see the final results.

Baseline. We report the results of the baseline. We use the union alignment and we extract the BP of length 3. As default language model feature, we use the standard trigram with smoothing Kneser-Ney and interpolation. Also we tune the parameters (only two parameters) with the SIMPLEX algorithm (see Table 2).

Posterior probability. Table 2 shows the effect of using the posterior probability: $P(e|f)$. We use all the features but the $P(e|f)$ and we optimise the parameters. We see the results without this feature decrease around 1.1 points both in BLEU and WER (see line 2rd and 5th in Table 2).

IBM Model 1. We do the same as in the paragraph above, we do not consider the IBM Model 1 and the IBM1^{-1}. Under these conditions, the translation's quality decreases around 1.3 points both in BLEU and WER (see line 3th and 5th in Table 2).

Modification of the Phrase Extraction. Finally, we made an experiment without modification of the phrases' length. We can see the comparison between: (1) the phrases of fixed maximum length of 3; and (2) including phrases with a maximum length of 5 which can not be generated by smaller phrases. We can see it in Table 2 (lines 4th and 5th). We observe that there is no much difference between the number of phrases, so this approach does not require more resources. However, we get slightly better scores.

5.3 Shared Task

This section explains the participation of "Exploiting Parallel Texts for Statistical Machine Translation". We used the EuroParl data provided for this shared task [4]. A word-to-word alignment was performed in both directions as explained in section 2. The phrase-based translation system which has been considered implements a total of 7 features (already explained in section 4). Notice that the language model has been trained with the training provided in the shared task. However, the optimization in the parameters has not been repeated, and we used the parameters obtained in the subsection above. We have obtained the results in the Table 3.

6 Conclusions

We reported a new method to extract longer phrases without increasing the quantity of phrases (less than 0.5%).

We also reported several features as $P(e|f)$ which in combination with the functions of the source-channel model provides significant improvement. Also, the feature IBM1 in combination with IBM1^{-1} provides improved scores, too.

Finally, we have optimized the parameters, and we provided the final results which have been presented in the Shared Task: Exploiting Parallel Texts for Statistical Machine Translation (June 30, 2005) in conjunction with ACL 2005 in Ann Arbor, Michigan.

7 Acknowledgements

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

References

[1] P.F. Brown, J. Cocke, S.A. Della Pietra, V.J. Della Pietra, F. Jelinek, J.D. Lafferty, R.L. Mercer, and P.S. Rossin. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85, June 1990.

[2] Josep M. Crego, José B. Mariño, and Adrià de Gispert. An Ngram-based Statistical Machine Translation Decoder. In *Draft*, 2005.

[3] G. Doddington. Automatic evaluation machine translation quality using n-gram co-ocurrence statistics. In *Proc. ARPA Workshop on Human Language Technology*, 2002.

[4] EuroParl: European Parliament Proceedings Parallel Corpus. Available on-line at: http://people.csail.mit.edu/koehn/publications/europarl/. 1996-2003.

[5] I. García-Varea. *Traducción Automática estadística: Modelos de Traducción basados en Máxima Entropía y Algoritmos de Búsqueda* . UPV, Diciembre 2003.

[6] Giza++. http://www-i6.informatik.rwth-aachen.de/∼och/software/giza++.html/, 1999.

[7] P. Koehn. A Beam Search Decoder for Phrase-Based Statistical Machine Translation Models. 2003.

[8] P. Koehn, F. J. Och, and D. Marcu. Statistical phrase-based translation. In *Proceedings of the Human Language Technology Conference (HLT-NAACL)*, pages 127–133, May 2003.

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

Phr Length	λ_{LM}	$\lambda_{p(f\|e)}$	$\lambda_{p(e\|f)}$	λ_{IBM1}	$\lambda_{IBM1^{-1}}$	λ_{PP}	λ_{WP}	*WER*	*BLEU*	*NIST*	*# frases*
3	0.788	0.906	0	0	0	0	0	33.98	57.44	10.11	67.7M
3+5length	0.788	0.941	0	0.771	0.200	3.227	0.448	28.97	64.71	11.07	68M
3+5length	0.788	0.824	0.820	0	0	3.430	-0.083	29.17	64.59	10.99	68M
3	0.746	0.515	0.979	0.514	0.390	1.537	-1.264	27.94	65.70	11.18	67.7M
3+5length	0.788	0.617	0.810	0.635	0.101	1.995	-0.296	27.88	65.82	11.23	68M

Table 2: *Results for the different experiments with optimized parameters in the direction SPA->ENG*

Phr Length	λ_{LM}	$\lambda_{p(f\|e)}$	$\lambda_{p(e\|f)}$	λ_{IBM1}	$\lambda_{IBM1^{-1}}$	λ_{PP}	λ_{WP}	*BLEU*	*# frases*
3+5length	0.788	0.617	0.810	0.635	0.101	1.995	-0.296	29.84	34.8M

Table 3: *Results for the ACL training and ACL test (SPA->ENG)*

[10] 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. In *Proceedings of the Human Language Technology Conference (HLT-NAACL)*, May 2004.

[11] F. J. Och and H. Ney. The Alignment Template Approach to Statistical Machine Translation. *Computational linguistics*, 30:417–449, December 2004.

[12] Franz Josef Och and Hermann Ney. Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. In *ACL*, pages pages 295–302, July 2002.

[13] Papineni, S.Roukos, and R.T. Ward. Feature-based language understanding. In *European Conf. on Speech Communication and Technology*, pages 1435–1438, September 1997.

[14] Papineni, S.Roukos, and R.T. Ward. Maximum likelihood and discriminative training of direct translation models. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Proceedings*, pages 189–192, May 1998.

[15] K.A. Papineni, S. Roukos, T. Ward, and W.J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Technical Report RC22176 (W0109-022), IBM Research Division*, 2001.

[16] A. Stolcke. SRILM - An Extensible Language Modeling Toolkit. In *Proceedings Intl. Conference Spoken Language Processing*, September 2002.

[17] R. Zens and H. Ney. Improvements in Phrase-Based Statistical Machine Translation. In *Proceedings of the Human Language Technology Conference (HLT-NAACL)*, pages 257–264, May 2004.