Phrase-Based Statistical Machine Translation using Approximate Matching

Jesús Tomás[1], Jaime Lloret[2], and Francisco Casacuberta[3]

[1] Instituto Tecnolgico de Informtica,
[2] Departamento de Comunicaciones,
Universidad Politcnica de Valencia, 46071 Valencia, Spain

Abstract. Phrase-based statistical models constitute one of the most competitive pattern-recognition approaches to machine translation. In this case, the source sentence is fragmented into phrases, then, each phrase is translated by using a stochastic dictionary. One shortcoming of this phrase-based model is that it does not have an adequate generalization capability. If a sequence of words has not been seen in training, it cannot be translated as a whole phrase. In this paper we try to overcome this drawback. The basic idea is that if a source phrase is not in our dictionary (has not been seen in training), we look for the most similar in our dictionary and try to adapt its translation to the source phrase. We are using the well known edit distance as a measure of similarity. We present results from an English-Spanish task (XRCE).

1 Introduction

The development of a classical *machine translation* (MT) system requires great human effort. *Statistical machine translation* (SMT) has proven to be an interesting pattern-recognition framework for (quasi) automatically building MT systems if adequate parallel corpora are available [1].

The earlier approaches to SMT were *single-word-based* models [2]. The basic assumption of these models is that each source word is generated by only one target word. This does not correspond to the nature of natural language; in some cases, it is necessary to know the context of the word to be translated.

To upgrade this assumption, the so-called *alignment-template* approach was proposed [3]. A template establishes the alignment (possibly through reordering) between a source sequence of word classes and a target sequence of word classes. The lexical model inside the templates is still based on word-to-word correspondences.

A simple alternative to this model has been introduced in recent works: The *phrase-based* (PB) approach [1, 4, 5]. This type of model deals also with the probability that a sequence of contiguous words (*source phrase*) in a source sentence is a translation of another sequence of contiguous words (*target phrase*) in the target sentence. However, in this case, the statistical dictionaries of single-word pairs are substituted by statistical dictionaries of *bilingual phrases*.

Despite its simplicity, the PB approach is one of the most competitive in the present state of the art in SMT [5]. One shortcoming of the PB model is that it does not have an adequate generalization capability. If a sequence of words has not been seen in the training corpus, it cannot be translated as a whole phrase. For example, suppose that the system has in the phrase-dictionary the translation of the English-Spanish bilingual phrase:

"network services user guide" ⇒ *"guia del usuario de servicios de red"* (a)

If a source phrase matches exactly with the left side of (a), the system does not have any problem in obtaining the translation. However, if the source phrase is slightly different,

"network utilities user guide"(b),

the system does not find it in the dictionary, thus the only possibility is translating it using smaller phrases. For example:

"network utilities" ⇒ *"utilidades de red"*
"user guide" ⇒ *"guia del usuario" or "guia del usuario de".*

To obtain the correct translation *"guia del usuario de utilidades de red"* is possible if the system uses the second translation of *"user guide"* and decides to reorder both phrases. However, the correct target phrase has not been seen in training, thus the language model does not have predilection with this output, and the most likely outcome is that the system prefers not to reorder, obtaining the incorrect output *"utilidades de red guia del usuario"*.

Several approaches have been proposed that can overcome the drawback presented by the PB models. In the AT approach, word classes are used. These word classes are learned using an unsupervised method from a bilingual corpus. Another possibility is to use a part-of-speech tagger to determine the word classes [6]. The use of word classes in the AT approach can present a problem of overgeneralization.

Another related work, within the framework of synchronous context-free grammar, is the hierarchical PB model [7]. This model makes it possible to learn a long distance phrase-based reordering. Our goal is different as we focus our attention on a short distance reordering. This kind of reordering is the biggest source of error in English to Spanish machine translation.

In this work, we are interested in obtaining a smooth generalization of the PB approach. For example, if we have the input phrase (b) and we know the translation of a similar source phrase (a), we are interested in using this information to translate (b). The basic idea is that if a source phrase is not in our dictionary (has not been seen in training), we look for the most similar one in our dictionary, and try to adapt its translation to the new source phrase. We use the well known edit distance as a measure of similarity [8].

2 Statistical Machine Translation

The goal of SMT is to translate a given source language sentence $s_1^J = s_1...s_J$ into a target sentence $t_1^I = t_1...t_I$. The methodology used [2] is based on the definition of a function $Pr(t_1^I|s_1^J)$ that returns the probability that t_1^I is a translation of a given s_1^J. Following the log-linear approach [9], this function can be expressed as a combination of a series of feature functions, $h_m(t_1^I, s_1^J)$, that are calibrated by scaling factors, λ_m:

$$\hat{t_1^I} = \underset{t_1^I}{\operatorname{argmax}}\, Pr(t_1^I|s_1^J) = \underset{t_1^I}{\operatorname{argmax}} \sum_{m=1}^{M} \lambda_m h_m(t_1^I, s_1^J) \tag{1}$$

This framework allows us a simple integration of several models in the translation system. Moreover, scaling factors allow us to adjust the relative importance of each model. For this objective, Och and Ney propose a minimum error rate criterion [9].

2.1 Phrase-based models

In many state of the art SMT systems, the most important feature function in equation 1 is the PB model. The main characteristic of this model is that it attempts to calculate the translation probabilities of word sequences (phrases) rather than only single words. These methods explicitly estimate the probability of a sequence of words in a source sentence ($\tilde{s}$) being translated as another sequence of words in the target sentence ($\tilde{t}$).

To define the PB model, we segment the source sentence s_1^J into K phrases ($\tilde{s}_1^K$) and the target sentence t_1^I into K phrases ($\tilde{t}_1^K$). A uniform probability distribution over every possible segmentation is assumed. If we assume a monotone alignment, that is, the target phrase in position k is produced only by the source phrase in the same position we get:

$$Pr(t_1^I|s_1^J) \propto \max_{K,\tilde{t}_1^K,\tilde{s}_1^K} \prod_{k=1}^{K} p(\tilde{t}_k|\tilde{s}_k) \tag{2}$$

where the parameter $p(\tilde{t}|\tilde{s})$ estimates the probability of translating the phrase $\tilde{s}$ into the phrase $\tilde{t}$. A phrase can be comprised of a single word (but empty phrases are not allowed). Thus, the conventional word to word statistical dictionary is included. If we permit the reordering of the target phrases, a hidden phrase level alignment variable, α_1^K, is introduced. In this case, we assume that the target phrase in position k is produced only by the source phrase in position α_k.

$$Pr(t_1^I|s_1^J) \propto \max_{K,\tilde{t}_1^K,\tilde{s}_1^K,\alpha_1^K} p(\alpha_1^K) \prod_{k=1}^{K} p(\tilde{t}_k|\tilde{s}_{\alpha_k}) \tag{3}$$

where the distortion model $p(\alpha_1^K)$ establishes the probability of a phrase alignment. Usually a first order model is used, assuming that the phrase-based alignment depends only on the distance of a phrase to the previous one [3].

3 Phrase-Based SMT using Approximate Matching

As section 1 comments, the main weakness of the PB models is the generalization capability. Only phrases that have been seen in a training bilingual corpus can be used in the bilingual dictionary. To deal with this problem, we propose searching the unseen phrases in the bilingual dictionary using approximate matching.

Edit distance (or Levenshtein distance) [10] has proven to be a powerful error-tolerance similarity measure. Moreover, as a subproduct we can determine the minimum edit operations (substitutions, insertions and deletions of words) needed to convert the input phrase into the reference phrase.

The proposed method is used when a source phrase $\tilde{s}$ is not in the dictionary. In this case, we look for bilingual phrases in the dictionary, ($\tilde{s}'$, $\tilde{t}'$), whose source part is very similar to $\tilde{s}$. We consider this case when the edit distance between $\tilde{s}$ and $\tilde{s}'$ is less than a given threshold (typically 1 or 2).

One important issue is to know which words are different in $\tilde{s}$ and $\tilde{s}'$. The difference can be represented as a sequence of edit operations (substitutions, insertions and deletions of words). In many cases there are several minimum edit sequences. For our algorithm, the same result is achieved.

Another important matter is to know which word is the translation of each word in the bilingual phrase that we are generalizing ($\tilde{s}'$, $\tilde{t}'$). For this purpose, we use the most probable word alignment according to the IBM model 1. Word alignments are represented by a, where, $a_{j'}{=}i'$ indicates that the source word $\tilde{s}'_{j'}$ has been aligned with the target word $\tilde{t}'_{i'}$

Using the minimum edit sequence and the word alignments, we are interested in obtaining $\tilde{t}$, the translation of $\tilde{s}$, based on the known bilingual phrase ($\tilde{s}'$, $\tilde{t}'$), as follows: $\tilde{t}'$ is modified according to the differences found between $\tilde{s}$ and $\tilde{s}'$, These differences are represented in the edit sequence. If a *substitution* operation is found associated to a source word $\tilde{s}_j$, we look for the most probable translation of this word (the translation can be a word or a phrase). Then, the target word, $\tilde{t}_i$, is replaced by this word/phrase. The target position, i, is determined using the word alignment, $i{=}a_j$. Similar procedures are followed for the *insertion* and the *deletion* operations. A detailed description of the algorithm is presented in figure 2 and an example is shown in figure 1.

The probability of a new bilingual phase, $p(\tilde{t}|\tilde{s})$, is calculated multiplying the probability of the generalized phrase, $p(\tilde{t}'|\tilde{s}')$, by the probability of each word inserted or substituted, $p(\tilde{t}_i|\tilde{s}_j)$, by a special probability to penalize each deletion operation, p_{del}. These bilingual phrase probabilities are introduced in equation 1 as a new feature function.

Given a source phrase (not in dictionary), $\tilde{s}$	$\tilde{s}$:	network		new		utilities		user	guide		
Look for a similar source phrase in dictionary, $\tilde{s}'$	e:	E		I		S		E	E		
Obtain e, minimum edit sequence between $\tilde{s}'$ and $\tilde{s}$	$\tilde{s}'$:	network				services		user	guide		
Look for $\tilde{t}'$, translation of $\tilde{s}'$	a:	7				5		3	1		
Obtain a, word alignment between $\tilde{t}'$ and $\tilde{s}'$	$\tilde{t}'$:	guia	del	usuario	de			servicios	de	red	
Use e and a to transform $\tilde{t}'$ in $\tilde{t}$, the translation of $\tilde{s}$	$\tilde{t}$:	guia	del	usuario	de	nuevas		utilidades	de	red	

Fig. 1. Algorithm and example of phrase generalization using approximate matching.

INPUT:	bilingual phrase dictionary: $p(\tilde{t}'\|\tilde{s}')$ source phrases in test: $\tilde{S}$
OUTPUT:	bilingual phrase dictionary: $p(\tilde{t}\|\tilde{s})$
PARAMETERS:	minimum probability of a bilingual phrase for generalizing : p_{min} maximum edit distance between source phrases for generalizing: e_{max} probability to penalize deletion operation: p_{del}

$\forall \tilde{s} \in \tilde{S}$ / $\forall \tilde{t}\ p(\tilde{t}|\tilde{s}) = 0$
  $\forall(\tilde{s}', \tilde{t}')$ / $p(\tilde{t}'|\tilde{s}') \geq p_{min} \wedge EditDistance(\tilde{s}', \tilde{s}) \leq e_{max}$
    Let be e a minimum edit sequence between $\tilde{s}'$ and $\tilde{s}$ ($e_k \in$ {E,S,I,D})
    Let be a the more probable IBM1 word alignment between $\tilde{t}'$ and $\tilde{s}'$
    $\tilde{t} = \tilde{t}'$; $p = p(\tilde{t}'|\tilde{s}')$; j=1; j'=1
    For $k = 1$ to $|e|$
      Case e_k
        E: j++; j'++
        S: $\tilde{t}_{a_{j'}} = \text{argmax}_t\, p(t|\tilde{s}_j)$; $p = p \cdot max_t p(t|\tilde{s}_j)$; j++; j'++
        I: $t = \text{argmax}_t\, p(t|\tilde{s}_j)$; insert t at left of $\tilde{t}_{a_{j'}}$; $p = p \cdot max_t p(t|\tilde{s}_j)$; j++
        D: $\tilde{t}_{a_{j'}} = \varnothing$; $p = p \cdot p_{del}$; j'++
    Insert new bilingual phrase: $p(\tilde{t}|\tilde{s})$=p

Fig. 2. Detailed algorithm used for generalizing bilingual phrases based on approximate matching. $EditDistance(\tilde{s}', \tilde{s})$ is the minimum number of substitution, insertion, and deletion operations needed to convert $\tilde{s}'$ into $\tilde{s}$. This minimum edit sequence is represented as e, using the symbols: E-*equal*, S-*substitution*, I-*insertion*, and D-*deletion*. $|e|$ is the number of symbols in e.

3.1 Algorithm implementation

The application of the proposed method should be carried out using three restrictions/modifications. The first obvious restriction is to use this approach only when a very similar phrase is found in the phrasal dictionary. That is, only when the edit distance between both phrases is one (e_{max}=1).

The second restriction is motivated by the following argument: In many cases, the unseen phrase can be correctly translated as a monotone concatenation of two phrases in the dictionary. When this occurs, it is not a good idea to use the proposed algorithm. If there are no other alignments that cross it, we consider this phrase can be monotonely generalized, thus, we do not apply the algorithm to this phrase. Figure 3 (a) shows an example.

The third restriction is due to the observation that the replacement of a single word in a phrase produces many concordance errors. As observed in figure 3 (b), when the Spanish word "escaner" is replaced by "impresora" the indefinite article "un" must be replaced by "una" for a correct gender concordance.

In order to solve this kind of error we proceeded as follows: If the previous word of a given word is aligned with a previous word in the target phrase, and we know the translation of this bigram, then we replace the two aligned words with this translation in the target phrase. Figure 3 (c) shows an example. This procedure is also tried using a group of three words: previous word, word to be replaced, and next word. Also using a group of two words: word to be replaced and next word.

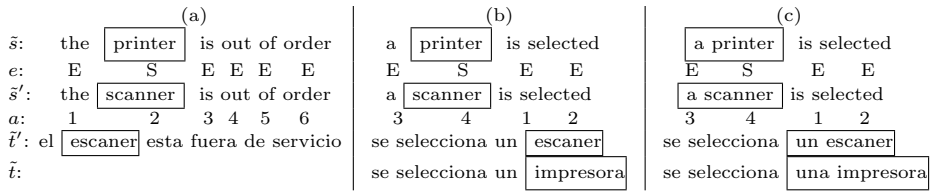

Fig. 3. Examples of several phrase generalizations: (a) The generalization is rejected. (b) Gender concordance error. "un impresora" must be replaced by "una impresora". (c) Error of b is corrected by replacing in $\tilde{t}'$, "un escaner" by the more probable translation of "a printer", that is "una impresora."

3.2 Search

The generalization procedure proposed in this work has been incorporated into a search engine previously developed for the PB models. In our implementation we use a simple solution: When phrases in the source sentence are searching in the bilingual dictionary, if a phrase is not found, we use the proposed procedure to obtain a set of possible translations of this phrase.

Given a source phrase, the algorithm described in figure 2 must find all the phrases in the dictionary with edit distance less than a certain threshold, which can be an expensive computational problem.

In a preliminary implementation we have used a serial search. Although the number of phrases can be very high, serial search can be performed in a reasonable time using several restrictions when edit distance threshold is set to one. For example, in the reported experiments, a test of more than a thousand sentences is translated in less than one hour (with a dictionary of two million of phrases).

More efficient search algorithms for this problem are described in [11]. Some of these algorithms can achieve a computational cost of square root of the number of phrases in the dictionary.

4 Experimental Results

In order to validate the approach described in this paper, a series of experiments was carried out using the XRCE corpus [12]. They involve the translation of technical Xerox manuals from English to Spanish [3].

As evaluation criteria we use *Word Error Rate* (WER) [13] and *BiLingual Evaluation Understudy* (BLEU) [14]. Statistical significance of the results is calculated using paired bootstrap [13]. In table 1, we highlight the statistical significance as follows. A result labelled with a "▲" ("△") means that the system is better than the baseline with a confidence of 99% (95%). A "−" means no significant differences.

In the experiments, following log-linear model combination is used:

[3] Train (English/Spanish): 56 K sentences, 665/753 K words, 26/30 K vocabulary.

$$\hat{t}_1^{\hat{I}} = \underset{t_1^I, \tilde{t}_1^K, \tilde{s}_1^K}{\operatorname{argmax}} \sum_{i=1}^{I} \left[c_1 + \lambda_1 \log p(t_i|t_{i-2}^{i-1}) + \lambda_2 \log \sum_{j=1}^{J} p(t_i|s_j) + \lambda_3 \log \sum_{j=1}^{J} p(s_j|t_i) \right] + \sum_{k=1}^{K} \left[c_2 + \lambda_4 \log p(\tilde{t}_k|\tilde{s}_k) + \lambda_5 \log p_{am}(\tilde{t}_k|\tilde{s}_k) \right]$$

This integrates the following knowledge sources: Target language model (trigram model) $p(t_i|t_{i-2}^{i-1})$. Single word translation models (IBM model 1), both direct ($p(t_i|s_j)$) and inverse ($p(s_j|t_i)$). Conventional phrase based translation model ($p(\tilde{t}|\tilde{s})$). Phrase based translation model using approximate matching ($p_{am}(\tilde{t}|\tilde{s})$). Two penalties, c_1 and c_2, are included to control I and K values.

Default parameters in the experiments were: maximum phrase length, which was 14 words; parameter estimation, which was the relative frequency and the search, which was monotone (equation 2). Results with non-monotone search were similar to the results with monotone search.

Several experiments were carried out to assess the approach presented. Table 1 compares the results obtained using the baseline PB model with the approximate matching algorithm. Several modifications of this algorithm have been proposed in section 3.1. Table 1 also shows the improvements obtained with these modifications, which are essential to obtain good results.

	WER	BLEU
baseline	24.7	64.9
+ approximate matching	24.5^{-}	65.1^{-}
+ excluding monotone	$24.1^{\blacktriangle}$	$65.5^{\triangle}$
+ replace contiguous words	$23.9^{\blacktriangle}$	$65.7^{\blacktriangle}$

Table 1. Results using several tuning algorithm in English-Spanish XRCE task.

In a second experiment, we analyze how often approximate matching is used and how often the results are improved: 64% of the test sentences contain almost one generalized phrase. Many of the generalized phrases are not used in the final output. In the experiment, only 10% of the sentences had different output after adding the new phrases. When the output is different, we compare it with the base-line output in terms of WER obtaining better results in 6% of the sentences and worse results in 2% of the sentences. The WER decreases 0.8 points.

5 Conclusions and Further Work

This work investigates how to deal with the sparsity problem within the PB model. In order to overcome the generalization capability in this model, a new method to adapt the bilingual phrases in the dictionary to unseen phrases has been proposed. The method uses an approximate matching based on the well-known edit distance. In the experimental phase, we have demonstrated that we can significantly statistically reduce the translation errors in the XRCE task.

In the future, we plan to validate this approach with other corpora and language pairs. A more efficient search algorithm must be used. In [15], edit distance is generalized using the permutation operation. We are interested in incorporating this new operation in the algorithm.

Acknowledgments

This work has been partially supported by the Spanish project TIC2003-08681-C02-02 and the IST Programme of the European Union IST-2001-32091.

References

1. Tomás, J., Casacuberta, F.: Monotone statistical translation using word groups. In: Procs. of the Machine Translation Summit VIII, Santiago, Spain (2001) 357–361
2. Brown, P.F., Della Pietra, S.A., Della Pietra, V.J., Mercer, R.L.: The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics **19**(2) (1993) 263–311
3. Och, F., Ney, H.: The alignment template approach to statistical machine translation. Computational Linguistics, **30**(4) (2004) 417–450
4. Zens, R., Och, F.J., Ney, H.: Phrase-based statistical machine translation. Advances in Artificial Inteligence **LNAI 2479**(25) (2002) 18–32
5. Koehn, P., Och, F.J., Marcu, D.: Statistical phrase-based translation. In: Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAACL), Edmonton, Canada (2003) 48–54
6. Tomás, J., Casacuberta, F.: Combining phrase-based and template-based models in statistical machine translation. Volume 2652 of Lecture Notes in Computer Science. Springer-Verlag (2003) 1021–1031
7. Chiang, D.: A hierarchical phrase-based model for statistical machine translation. In: Proc. of ACL 2005, Michigan, USA (2005) 263–270
8. Mandreoli, F., Martoglia, R., Tiberio, P.: Searching similar (sub)sentences for example-based machine translation. In: Proc. Atti del Decimo Convegno Nazionale su Sistemi Evoluti per Basi di Dati, Isola d'Elba, Italy (2002)
9. Och, F.J., Ney, H.: Discriminative training and maximum entropy models for statistical machine translation. In: Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, PA (2002)
10. Levenstein, V.: Binary codes capable of correcting deletions, insertions, and reversals. Cybernetics and Control Theory **10**(8) (1965) 707–710
11. Hall, P.A.V., Dowling, G.R.: Approximate string matching. ACM Comput. Surv. **12**(4) (1980) 381–402
12. SchlumbergerSema S.A., Inst. Tec. de Informática, R.W.T.H. Aachen, University of Montreal, Celer Soluciones, Société Gamma, Xerox Research Centre Europe: TT2. TransType2 - computer assisted translation. Project technical annex. (2001)
13. Bisani, M., Ney, H.: Bootstrap estimates for confidence intervals in asr performance evaluation. In: IEEE International Conference on Acoustics, Speech, and Signal Processing. Volume 1., Montreal, Canada (2004) 409–412
14. Papineni, K.A., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. Technical Report RC22176, IBM Research Division, Thomas J. Watson Research Center, Yorktown Heights, NY (2001)
15. Leusch, G., Ueffing, N., Ney, H.: A novel string-to-string distance measure with applications to machine translation evaluation. In: Proc. of Machine Translation Summit IX, New Orleans, USA (2003) 240–247