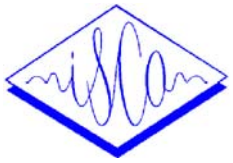
ISCA Archive
http://www.isca-speech.org/archive



INTERSPEECH 2004 - ICSLP
8th International Conference on Spoken Language Processing
ICC Jeju, Jeju Island, Korea
October 4-8, 2004

Finite-state-based and Phrase-based Statistical Machine Translation

Josep M. Crego, José B. Mariño and Adrià de Gispert

TALP Research Center
Universitat Politècnica de Catalunya, Barcelona
{jmcrego|canton|agispert}@talp.upc.es

Abstract

This paper shows the common framework that underlies the translation systems based on phrases or driven by finite state transducers, and summarizes a first comparison between them. In both approaches the translation process is based on pairs of source and target strings of words (segments) related by word alignment. Their main difference comes from the statistical modeling of the translation context. The experimental study has been carried out on an English/Spanish version of the VERBMOBIL corpus. Under the constrain of a monotone composition of translated segments to generate the target sentence, the finite state based translation outperforms the phrase based counterpart.

1. Introduction

Statistical Machine Translation (SMT) is thought as a task where each source sentence f_1^J is transformed into (or generates) a target sentence e_1^I, by means of a stochastic process. The generative model explains how the process is carried out. Thus, translation of a source sentence f_1^J can be formulated as the search of the target sentence e_1^I that maximizes the conditional probability $p(e_1^I|f_1^J)$, which can be rewritten using the Bayes rule as:

$$\max_{e_1^I} \left\{ p(f_1^J|e_1^I) \cdot p(e_1^I) \right\} \quad (1)$$

where $p(f_1^J|e_1^I)$ represents the translation model and $p(e_1^I)$ is the target language model.

Regarding the translation model, the first statistical systems worked at the word level [1], viewing the translation task as a process of translating words and reordering them to build the target sentence.

In the last few years, new systems tend to use sequences of words, commonly called phrases, trying thus to introduce the word context in the translation model, until now only taken into account in the language model. Among these systems, we find those able to deal with phrases extracted from word occurrences [2], phrases syntactically motivated [3], and from a joint probability model [4]. Results show a consistently better performance of these approaches with respect to single word based [5], especially when using phrases extracted from word occurrences.

However, translation can also be seen as a stochastic process maximizing the joint probability $p(f, e)$, typically implemented by means of a Finite-State Transducer (FST):

$$\max_{e_1^I} \{p(e_1^I, f_1^J)\} \quad (2)$$

This approach comes from the speech-to-speech translation task, where in an integrated architecture this joint probability is maximized together with the acoustic model $p(x|f)$ (x being the input acoustic signal).

Despite the theoretic similarity between both approaches (as both maximizations are mathematically equivalent), the actual implementations do produce certain differences that we explore in this paper. By using a unified framework with a practically equivalent search algorithm, a comparison of these models is presented, highlighting their advantages and disadvantages in a real translation task.

The paper is organized as follows. Section 2 describes briefly the particularities of the evaluated SMT systems, section 3 introduces the evaluation framework used to carry out the comparison, whereas the results obtained are discussed in section 4. Finally, section 5 concludes and outlines further research.

2. Translation Model

Before presenting an overview of the Finite-State-based and Phrase-based approaches (hereinafter referred to as FSB and PB, for simplicity), an important distinction must be made between phrases and tuples, as they constitute the core unit from which these systems learn the translation model.

2.1. Phrase and Tuple definition

Given a sentence pair and a corresponding word alignment, a phrase (or bilingual phrase) is any pair of m source words and n target words that satisfies two basic constraints [2]:

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.

FSB systems use a particular case of these phrases, called tuples. Given a parallel sentence, the set of tuples can be defined as the subset of the phrases that fulfills the following conditions:

1. It induces a monotonous segmentation of the pair of sentences.
2. Each tuple cannot be decomposed into smaller phrases without violating the previous constraint.

Note that this subset is unique under these conditions, that is, there is only one possible set of tuples given a parallel sentence.

The example in Figure 1 shows the phrases and tuples extracted from a given pair of aligned source/target sentences. As it can be seen, whereas the sentence pair can be segmented into multiple sets of phrases (ex: [p1+p6+p8+p10+p12], [p1+p7+p11], [p2+p9], etc.), only one segmentation is possible when extracting tuples (ex: [t1+t2+t3+t4]).

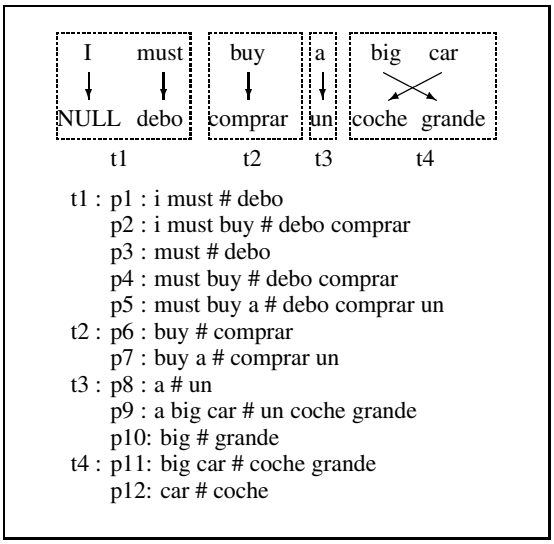


Figure 1: *Tuple and Phrase extraction given an aligned sentence pair. Only phrases of three or less words per side are shown.*

To build phrases from a given alignment, the *extractBP* algorithm has been used [6]. The tuples extraction algorithm is here outlined:

1. Initially, every link is a tuple.
2. If a word is aligned to NULL (ie. the word "*I*" in the example of figure 1). It is added to the next tuple (if it is not possible, the previous tuple is used to be added to).
3. If there is a crossing in the tuples sequence (ex: big # grande, car # coche), the tuples generating the crossing are joined. Step (3) is repeated until no crossings are left.

This method preserves the monotonicity for both languages in the tuples.

2.2. Finite-state-based Translation

Finite-state-based Translation Systems model the translation directly as a composition of tuples. The system learns translations from this bilingual units that are extracted from the word alignments. This way the context used in the translation model is bilingual, it not only takes the target sentence into account, but both languages linked in tuples.

The translation is achieved by a string composition of the most probable sequence of tuples.

$$\hat{e}_1^I = \arg\max_{e_1^I}\{p(e_1^I, f_1^J)\} = \cdots = \tag{3}$$

$$\arg\max_{e_1^I}\{\prod_{n=1}^{N} p((e,f)_n|(e,f)_{n-x+1},...,(e,f)_{n-1})\} \tag{4}$$

The n-*th* tuple of a sentence pair is here referred as $(e,f)_n$.

The translation model can be seen here as a language model, where the language is composed by tuples [7].

2.3. Phrase-based Translation

Phrase-based Translation Systems incorporate the word context into the phrases. Thus, they learn translations not only for single words, but for whole phrases. Translating a given source sentence is accomplished in three steps: segmentation into phrases, translation of the phrases and composition of the target sentence. The search is carried out by the maximization of the equation:

$$\hat{e}_1^I = \arg\max_{e_1^I}\{Pr(e_1^I) \cdot \prod_{k=1}^{K} p(\tilde{f}_k|\tilde{e}_k)\} \tag{5}$$

where $\tilde{f}_k, \tilde{e}_k$ refer to phrase k in each language.

This search can be performed without allowing phrase reordering in the target sentence (monotone decoding), or allowing phrase reordering (non-monotone decoding) [2].

Even though non-monotone phrase-based decoders offer better performance than monotone decoders [2], the monotone constrain defines an equivalent framework for comparing the way PB and FBS systems model translation. In this paper, a monotone search has thus been implemented.

On the other hand, different language models have been proposed and evaluated for the phrase-based approach [2] [8].

2.4. Key issues

Despite the initial mathematical equivalence of both approaches, the implementation of the systems produce differences, mainly found on the structure of phrases and tuples, and how they are used to model the translation.

Regarding the translation model, the FSB approach uses word context through the maximization of the probability of a sequence of tuples. This way, the context appears not only because words are joined inside tuples (sequence of words), but also connected as a sequence of tuples, modeled by an Ngram-style memory for translation (see equation 4). In turn, the PB approach learns this context by joining words in phrases. The connection between phrases is then done taking both the translation probabilities and a language model of the target into account (see equation 5).

Another remarkable difference results from the phrase/tuple extraction procedure. Whereas the FSB approach relies on a small set of tuples (derived from just one segmentation per sentence), expecting them to accurately focus on the translation process, PB systems generate a huge set of phrases, thus relying on statistics to highlight the most appropriate ones by frequency.

Finally, in the PB approach it is possible to introduce large complementary monolingual data to better estimate the target language model (when available), unlike in the FSB approach as currently formulated. However, in this work we have used the same corpus to learn translation and language models.

3. Evaluation Framework

To perform the experiments, we have used the VerbMobil database in English, and its translation into Spanish generated in the framework of the LC-Star project (IST-2001-32216).

3.1. Corpus

The corpus is set up by the transcription of spontaneous dialogs in the appointment and meeting-planning domain, normalized to leave punctuation marks out, and preprocessed categorizing date and time expressions (in the training corpus 2746 time expressions and 897 date expressions were substituted by a unified tag).

Table 1 shows the main statistics of the used data, namely number of sentences, words, vocabulary, and maximum and

mean sentence lengths for each language, respectively.

VMobil	sent.	words	voc.	Lmax	Lmean
Train set					
English	27,995	207,730	3,138	66	7.4
Spanish		199,915	4,848	66	7,1
Test set					
English	2,059	20,585	1,258	57	10
Spanish		19,855	1,704	60	9.6

Table 1: *VerbMobil corpus statistics.*

While the English test set contains 138 words that have not occurred in the training, in Spanish we have 236 unseen words.

Table 1 shows the different sizes of the Spanish and English vocabularies in the corpus. It is remarkable the bigger size of the Spanish vocabulary due to the inflectional characteristic of Spanish, common to all Romance family languages.

3.2. Word alignments

The word alignment has been carried out using GIZA++ [9]. Sentences have been aligned in both translation directions. Afterwards, the combination of source-target and target-source alignments (Union) has been calculated [6].

Table 2 shows the total number of phrases and tuples extracted from the corpus and the number of different items (vocabulary size), given the input alignment and extraction method.

Alignment	Tuples	VocTpl	Phrases	VocPhr
spa2eng	218,071	18,006	569,563	216,536
eng2spa	201,882	20,085	607,474	237,671
union	194,062	18,029	568,773	217,196

Table 2: *Number of tuples and phrases (and their vocabulary sizes) in the training corpus.*

As expected, the number of phrases is bigger than the number of tuples. It is also remarkable the ratio between the items and their vocabulary ($\sim$12:1 for the tuples, $\sim$2:1 for the phrases). A very high percentage of phrases have only appeared a few times in the corpus. Table 3 shows the percentage of phrases and tuples appearing only once and twice in the corpus.

%	Tuples	Phrases
Singletons	68.1	82.1
Doubletons	10.6	8.3

Table 3: *Percentage of singletons and doubletons among the total number of tuples and phrases, for the spa2eng alignment.*

3.3. Language models

Different language models have been tested for the PB approach. All of them have been implemented using the CMU-Cambridge Language Modeling Toolkit. The models are:

1. A trigram model calculated for all the words e_i of the target sentence (SentLM).

$$Pr(e_1^I) = \prod_1^I p(e_i \mid e_{i-2}, e_{i-1}) \quad (6)$$

2. A trigram model calculated for the first word of each phrase $\tilde{e}_k$ (LinkLM) [8].

$$Pr(e_1^I) = Pr(\tilde{e}_1^K) = \prod_{k=1}^K p(e_{k_1} \mid e_{k_1-2}, e_{k_1-1}) \quad (7)$$

3. A model consisting on the conditional probability of each phrase $\tilde{e}_k$ given its previous word (PhraseLM) [2].

$$Pr(e_1^I) = Pr(\tilde{e}_1^K) = \prod_{k=1}^K p(\tilde{e}_k \mid e_{k_1-1}) \quad (8)$$

3.4. Translation models

For the PB system, the translation model probabilities have been calculated using the relative frequencies of the phrases in the training corpus.

$$Pr(\tilde{f}|\tilde{e}) = \frac{N(\tilde{f}, \tilde{e})}{N(\tilde{e})} \quad (9)$$

For the FSB system, the translation probabilities are calculated using the tuples extracted from the different word alignments as a N-gram language model ($N = 3$). We have also used the CMU-Cambridge Language Modeling Toolkit to estimate the probabilities.

3.5. Search

The maximization problem in equations 4 and 5 is solved using an efficient monotone search, implemented as a beam search using dynamic programming. The search algorithm used in the phrase-based approach is described in [2].

The algorithm for the FSB approach defines the quantity $Q(j)$ as the maximum probability of a phrase sequence that covers positions 1 to j of the source sentence. $Q(J+1)$ is the probability of the optimal translation. \$ is the boundary sentence marker:

$$Q(0) = 1 \quad (10)$$

$$Q(j) = \max_{0 \le j' < j, \tilde{e}} \{Q(j') \cdot p(f_{j'+1}^j, \tilde{e})\} \quad (11)$$

$$Q(J+1) = \max \{Q(J) \cdot p(\$, \$)\} \quad (12)$$

4. Results

To evaluate the translation task, we have used the scores WER and BLEU (interpolating unigrams, bigrams, trigrams and fourgrams) with only one reference.

Table 4 and table 5 show the results obtained for each experiment in both translation tasks using the different approaches and language models. From all four alignments (both directions, plus union and intersection), only those producing the best results are shown.

Alignment	Model	WER	BLEU
giza++ union	PhrsLM	32.55	0.4640
giza++ union	SentLM	32.96	0.4532
giza++ union	LinkLM	34.98	0.4437
giza++ spa2eng	FSB	31.50	0.4981

Table 4: *Spanish to English translation task.*

Alignment	Model	WER	BLEU
giza++ union	PhrsLM	38.18	0.4566
giza++ union	SentLM	38.35	0.4531
giza++ union	LinkLM	39.83	0.4438
giza++ eng2spa	FSB	35.70	0.5038

Table 5: *English to Spanish translation task.*

Four results are highlighted:

- The Spanish to English translation task achieves consistently better results than the English to Spanish translation task. This can be explained by the bigger size of the Spanish vocabulary.
- When comparing the different language models proposed in the PB approach, the LinkLM rates the worst, as it only takes one trigram into account. A slight preference for the PhrsLM language model instead of the SentLM is found.
- Comparing the approaches FSB and PB, consistent better results are achieved by the FSB approach.
- Regarding the alignments, the FSB approach achieves the highest performance using one-to-many alignments, specially for the alignment direction matching the translation direction. However, the PB approach achieves optimal performance with the alignment resulting from the union, usually generating more links but less accurate. This strategy turns harmful for the FSB approach, as it tends to produce much longer tuples that might overfit to training data.

Except for the PhrsLM and SentLM comparison, the rest are over the $\pm 0,6$ threshold of the confidence margin (given the number of words in the test). All results obtained with the BLEU score correlate with WER measure.

It is worth mentioning the use of only one reference to calculate WER and BLEU scores. This fact leads to a worse performance score than it is real, as described in the example of table 6, where although a good translation has been produced, the score is as high as $WER = 40\%$. These sentences are extracted from the Verbmobil corpus.

Source	te va bien el lunes
Target REF	is monday fine for you
Translation	does monday suits you

Table 6: *Example of correct translation different from the reference.*

5. Conclusion and Further work

Under the monotonicity restriction of the decoder algorithm (see section 2), the FSB approach outperforms the PB approach in both translation tasks (see tables 4 and 5), specially when translating from English to Spanish, where the bigger vocabulary size of the target language (Spanish) makes the task harder.

It seems that the PB approach does not benefit from the huger generation of phrases before the translation modeling. By estimating from a smaller but more accurate set of tuples, the FSB approach is able to better model the translation task. The different use of context might also explain this difference in performance. However, this should be confirmed when using other corpora, expanding the scope of the evaluation framework.

As introduced in 3, the current FSB approach is restricted to the monotonicity condition. Two extensions of this approach want to be investigated in future work to overcome this restriction:

- Introducing a word reordering scheme inside the tuple [10]. The reordering would be learned in the tuples extraction process, and used after the decoding process.
- Using a language model in the target sentence, to help the translation model when a translation probability can only be calculated using a unigram. That is, when the translation probability depends only on one tuple.

$$\arg\max_{e}\{p(e,f)\cdot p(e)\} \tag{13}$$

These extensions of the FSB approach would allow for a comparison of the PB and FSB approaches under a new evaluation framework using a non-monotone decoder.

6. References

[1] P. Brown, S. Della Pietra, V. Della Pietra, and R. Mercer, "The mathematics of statistical machine translation," *Computational Linguistics*, vol. 19, no. 2, pp. 263–311, 1993.

[2] R. Zens, F. Och, and H. Ney, "Phrase-based statistical machine translation," *Proc. Conference on Empirical Methods for Natural Language Processing*, 2002.

[3] K. Yamada and K. Knight, "A syntax-based statistical translation model," *Proceedings of ACL 39th meeting*, pp. 6–11, July 2001.

[4] 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)*, pp. 133–139, July 2002.

[5] P. Koehn, F. Och, D. Marcu, and W. Wong, "Statistical phrase-based translation," *HLT-NAACL 2003*, 2003.

[6] F. Och and H. Ney, "Improved statistical alignment models," *38th Annual Meeting of the Association for Computational Linguistics*, pp. 440–447, October 2000.

[7] A. de Gispert and J. Mariño, "Using X-grams for speech-to-speech translation," *Proc. of the 7th Int. Conf. on Spoken Language Processing, ICSLP'02*, September 2002.

[8] C. Tillmann and F. Xia, "A phrase-based unigram model for statistical machine translation," *HLT-NAACL*, 2003.

[9] Giza++ software, "http://www-i6.informatik.rwth-aachen.de/~och/software/giza++.html," *Statistical machine translation, final report, JHU workshop*, 1999.

[10] A. de Gispert and J. Mariño, "Experiments in word-ordering and morphological preprocessing for transducer-based statistical machine translation," *IEEE Automatic Speech Recognition and Understanding Workhsop, ASRU'03*, November 2003.