N-gram-based Machine Translation

José B. Mariño[*]
Rafael E. Banchs[*]
Josep M. Crego[*]
Adrià de Gispert[*]
Patrik Lambert[*]
José A. R. Fonollosa[*]
Marta R. Costa-jussà[*]
Universitat Politècnica de Catalunya

This article describes in detail an n-gram approach to statistical machine translation. This approach consists of a log-linear combination of a translation model based on n-grams of bilingual units, which are referred to as tuples, along with four specific feature functions. Translation performance, which happens to be in the state of the art, is demonstrated with Spanish-to-English and English-to-Spanish translations of the European Parliament Plenary Sessions (EPPS).

1. Introduction

The beginnings of statistical machine translation (SMT) can be traced back to the early fifties, closely related to the ideas from which information theory arose (Shannon and Weaver 1949) and inspired by works on cryptography (Shannon 1949, 1951) during World War II. According to this view, machine translation was conceived as the problem of finding a sentence by decoding a given "encrypted" version of it (Weaver 1955). Although the idea seemed very feasible, enthusiasm faded shortly afterward because of the computational limitations of the time (Hutchins 1986). Finally, during the nineties, two factors made it possible for SMT to become an actual and practical technology: first, significant increment in both the computational power and storage capacity of computers, and second, the availability of large volumes of bilingual data.

The first SMT systems were developed in the early nineties (Brown et al. 1990, 1993). These systems were based on the so-called noisy channel approach, which models the probability of a target language sentence T given a source language sentence S as the product of a translation-model probability $p(S|T)$, which accounts for adequacy of translation contents, times a target language probability $p(T)$, which accounts for fluency of target constructions. For these first SMT systems, translation-model probabilities at the sentence level were approximated from word-based translation models that were trained by using bilingual corpora (Brown et al. 1993). In the case of target language probabilities, these were generally trained from monolingual data by using n-grams.

Present SMT systems have evolved from the original ones in such a way that mainly differ from them in two respects: first, word-based translation models have been

[*] Department of Signal Theory and Communications, Campus Nord, Barcelona 08034, Spain.

Submission received: 9 August 2005; revised submission received: 26 April 2006; accepted for publication: 5 July 2006

replaced by phrase-based translation models (Zens, Och, and Ney 2002; Koehn, Och, and Marcu 2003) which are directly estimated from aligned bilingual corpora by considering relative frequencies, and second, the noisy channel approach has been expanded to a more general maximum entropy approach in which a log-linear combination of multiple feature functions is implemented (Och and Ney 2002).

As an extension of the machine translation problem, technological advances in the fields of automatic speech recognition (ASR) and text to speech synthesis (TTS) made it possible to envision the challenge of spoken language translation (SLT) (Kay, Gawron, and Norvig 1992). According to this, SMT has also been approached from a finite-state point of view as the most natural way of integrating ASR and SMT (Riccardi, Pieraccini, and Bocchieri 1996; Vidal 1997; Knight and Al-Onaizan 1998; Bangalore and Riccardi 2000). In this SMT approach, translation models are implemented by means of finite-state transducers for which transition probabilities are learned from bilingual data. As opposed to phrase-based translation models, which consider probabilities between target and source units referred to as phrases, finite-state translation models rely on probabilities among sequences of bilingual units, which are defined by the transitions of the transducer.

The translation system described in this article implements a translation model that has been derived from the finite-state perspective—more specifically, from the work of Casacuberta (2001) and Casacuberta and Vidal (2004). However, whereas in this earlier work the translation model is implemented by using a finite-state transducer, in the system presented here the translation model is implemented by using *n*-grams. In this way, the proposed translation system can take full advantage of the smoothing and consistency provided by standard back-off *n*-gram models. The translation model presented here actually constitutes a language model of a sort of "bilanguage" composed of bilingual units, which will be referred to as **tuples** (de Gispert and Mariño 2002). An alternative approach, which relies on bilingual-unit unigram probabilities, was developed by Tillmann and Xia (2003); in contrast, the approach presented here considers bilingual-unit *n*-gram probabilities. In addition to the tuple *n*-gram translation model, the translation system presented here implements four specific feature functions that are log-linearly combined along with the translation model for performing the decoding (Mariño et al. 2005).

This article is intended to provide a detailed description of the *n*-gram-based translation system, as well as to demonstrate the system performance in a wide-domain, large-vocabulary translation task. The article is structured as follows. First, Section 2 presents a complete description of the *n*-gram-based translation model. Then, Section 3 describes in detail the additional feature functions that, along with the translation model, compose the *n*-gram-based SMT system implemented. Section 4 describes the European Parliament Plenary Session (EPPS) data, as well as the most relevant details about the translation tasks considered. Section 5 presents and discusses the translation experiments and their results. Finally, Section 6 presents some conclusions and intended further work.

2. The Tuple *N*-gram Model

This section describes in detail the tuple *n*-gram translation model, which constitutes the core model implemented by the *n*-gram-based SMT system. First, the bilingual unit definition and model computation are presented in Section 2.1. Then, some important refinements to the basic translation model are provided and discussed in Section 2.2. Finally, Section 2.3 discusses issues related to *n*-gram-based decoding.

2.1 Tuple Extraction and Model Computation

As already mentioned, the translation model implemented by the described SMT system is based on bilingual *n*-grams. This model actually constitutes a language model of a particular bilanguage composed of bilingual units that are referred to as tuples. In this way, the translation model probabilities at the sentence level are approximated by using *n*-grams of tuples, such as described by the following equation:

$$p(T,S) \approx \prod_{k=1}^{K} p((t,s)_k|(t,s)_{k-1},(t,s)_{k-2},\ldots,(t,s)_{k-n+1}) \tag{1}$$

where t refers to target, s to source, and $(t,s)_k$ to the kth tuple of a given bilingual sentence pair. It is important to note that since both languages are linked up in tuples, the context information provided by this translation model is bilingual.

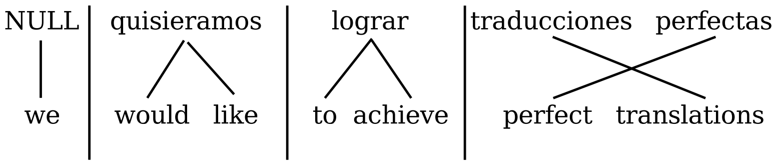
Tuples are extracted from a word-to-word aligned corpus in such a way that a unique segmentation of the bilingual corpus is achieved. Although in principle any Viterbi alignment should allow for tuple extraction, the resulting tuple vocabulary depends highly on the particular alignment set considered, and this impacts the translation results. According to our experience, the best performance is achieved when the union of the source-to-target and target-to-source alignment sets (IBM models; Brown et al. [1993]) is used for tuple extraction (some experimental results regarding this issue are presented in Section 4.2.2). Additionally, the use of the union can also be justified from a theoretical point of view by considering that the union set typically exhibits higher recall values than do other alignment sets such as the intersection and source-to-target.

In this way, as opposed to other implementations, where one-to-one (Bangalore and Riccardi 2000) or one-to-many (Casacuberta and Vidal 2004) alignments are used, tuples are extracted from many-to-many alignments. This implementation produces a monotonic segmentation of bilingual sentence pairs, which allows for simultaneously capturing contextual and reordering information into the bilingual translation unit structures. This segmentation also allows for estimating the *n*-gram probabilities appearing in (1). In order to guarantee a unique segmentation of the corpus, tuple extraction is performed according to the following constraints (Crego, Mariño, and de Gispert 2004):

- a monotonic segmentation of each bilingual sentence pair is produced,
- no word inside the tuple is aligned to words outside the tuple, and
- no smaller tuples can be extracted without violating the previous constraints.

Notice that, according to this, tuples can be formally defined as the set of shortest phrases that provides a monotonic segmentation of the bilingual corpus. Figure 1 presents a simple example illustrating the unique tuple segmentation for a given pair of sentences, as well as the complete phrase set.

The first important observation from Figure 1 is related to the possible occurrence of tuples containing unaligned elements on the target side. This is the case for tuple 1. Tuples of this kind should be handled in an alternative way for the system to be able to provide appropriate translations for such unaligned elements. The problem of how



Tuples:
1. NULL : we 2. quisieramos : would like
3. lograr : to achieve 4. traducciones perfectas : perfect translations

Phrases:
1. quisieramos : would like 2. quisieramos : we would like
3. lograr : to achieve 4. traducciones : translations
5. perfectas : perfect 6. quisieramos lograr : would like to achieve
7. quisieramos lograr : we would like to achieve
8. traducciones perfectas : perfect translations
9. lograr traducciones perfectas : to achieve perfect translations
10. quisieramos lograr traducciones perfectas :
would like to achieve perfect translations
11. quisieramos lograr traducciones perfectas :
we would like to achieve perfect translations

Figure 1
Example of tuple extraction. Tuples are extracted from Viterbi alignments in such a way that the set of shortest bilingual units that provide a monotonous segmentation of the bilingual sentence pair is achieved.

to handle this kind of situation, which we refer to as involving **source-nulled** tuples, is discussed in detail in Section 2.2.2.

Also, as observed from Figure 1, the total number of tuples is significantly lower than the total number of phrases, and, in most of the cases, longer phrases can be constructed by considering tuple n-grams, which is the case for phrases 2, 6, 7, 9, 10, and 11. However, phrases 4 and 5 cannot be generated from tuples. In general, the tuple representation is not able to provide translations for individual words that appear tied to other words unless they occur alone in some other tuple. This problem, which we refer to as embedded words, is discussed in detail in Section 2.2.1.

Another important observation from Figure 1 is that each tuple length is implicitly defined by the word links in the alignment. As opposed to phrase-extraction procedures, for which a maximum phrase length should be defined to avoid a vocabulary explosion, tuple extraction procedures do not have any control over tuple lengths. According to this, the tuple approach will strongly benefit from the structural similarity between the languages under consideration. Then, for close language pairs, tuples are expected to successfully handle those short reordering patterns that are included in the tuple structure, as in the case of "traducciones perfectas : perfect translations" presented in Figure 1. On the other hand, in the case of distant pairs of languages, for which a large number of long tuples are expected to occur, the approach will more easily fail to provide a good translation model due to tuple sparseness.

2.2 Translation Model Refinements

The basic n-gram translation model, as defined in the previous section, exhibits some important limitations that can be easily overcome by incorporating specific changes in either the tuple vocabulary or the *n*-gram model. This section describes such limitations and provides a detailed description of the implemented refinements.

2.2.1 Embedded Words. The first issue regarding the *n*-gram translation model is related to the already mentioned problem of embedded words, which refers to the fact that the tuple representation is not able to provide translations for individual words all the time. Embedded words can become a serious drawback when they occur in relatively significant numbers in the tuple vocabulary.

Consider for example the word *translations* in Figure 1. As seen from the figure, this word appears embedded into tuple "traducciones perfectas : perfect translations." If a similar situation is encountered for all other occurrences of that word in the training corpus, then no translation probability for an independent occurrence of that word will exist. A more relevant example would be the case of the embedded word *perfect* since this adjective always moves relative to the noun it is modifying. In this case, providing the translation system with a word-to-word translation probability for "perfectas : perfect" only guarantees that the decoder will have a translation option for an isolated occurrence of such words but does not guarantee anything about word order. So, certainly, any adjective–noun combination including the word *perfect*, which has not been seen during the training stage, will be translated in the wrong order. Accordingly, the problem resulting from embedded words can be partially solved by incorporating a bilingual dictionary able to provide word-to-word translation when required by the translation system. A more complete treatment for this problem must consider the implementation of a word-reordering strategy for the proposed SMT approach (as will be discussed in Section 6, this constitutes one of the main concerns for our further research).

In our *n*-gram-based SMT implementation, the following strategy for handling embedded words is considered. First, one-word tuples for each detected embedded word are extracted from the training data and their corresponding word-to-word translation probabilities are computed by using relative frequencies. Then, the tuple *n*-gram model is enhanced by including all embedded-word tuples as unigrams into the model. Since a high-precision alignment set is desirable for extracting such one-word tuples and estimating their probabilities, the intersection of both alignments, source to target and target-to-source, is used instead of the union.

In the particular case of the EPPS tasks considered in this work, embedded words do not constitute a real problem because of the great amount of training material and the reduced size of the test data set (see Section 4.1 for a detailed description of the EPPS data set). On the contrary, in other translation tasks with less available training material, the embedded-word handling strategy described above has been very useful (de Gispert, Mariño, and Crego 2004).

2.2.2 Tuples with Empty Source Sides. The second important issue regarding the *n*-gram translation model is related to tuples with empty source sides, hereinafter referred to as **source-nulled tuples**. In the tuple *n*-gram model implementation, it frequently happens that some target words linked to NULL end up producing tuples with NULL source sides. Consider, for example, the first tuple of the example presented in Figure 1. In this example, "NULL : we" is a source-nulled tuple if Spanish is considered to be the source language. Notice that tuples of this kind cannot be allowed since no NULL is expected to occur in a translation input.

The classical solution to this problem in the finite-state transducer framework is the inclusion of **epsilon arcs** (Knight and Al-Onaizan 1998; Bangalore and Riccardi

2000). However, epsilon arcs significantly increase decoding complexity. In our *n*-gram system implementation, this problem is easily solved by preprocessing the union set of alignments before extracting tuples, in such a way that any target word that is linked to NULL is attached to either its preceding word or its following word. In this way, no target word remains linked to NULL, and source-nulled tuples will not occur during tuple extraction.

Some different strategies for handling target words aligned to NULL have been considered. In the simplest strategy, which will be referred to as the **attach-to-right strategy**, target words aligned to NULL are always attached to their following word. This simple strategy happens to provide better results, for English-to-Spanish and Spanish-to-English translations, than the opposite one (attachment to the previous word), and also better than a more sophisticated strategy that considers bigram probabilities for deciding whether a given word should be attached to the following or to the previous one.

Notice that in the particular cases of Spanish and English, the attach-to-right strategy can be justified heuristically. Indeed, when translating from Spanish to English, most of the source-nulled tuples result from omitted verbal subjects, which is a very common situation in Spanish. This is the case for the first tuple in Figure 1. Suppose, for instance, that the attach-to-right strategy is used in Figure 1; in such a case, the tuple "quisiéramos : would like" will be replaced by the new tuple "quisiéramos : we would like," which actually makes a better translation unit, at least from a grammatical point of view. Similarly, some common situations can be identified for translations in the English-to-Spanish direction, such as omitted determiners (e.g., "I want information about European countries : quiero información sobre los_países Europeos"). Again, the attach-to-right strategy for the unaligned Spanish determiner *los* seems to be the best one.

Experimental results comparing the attach-to-right strategy to an additional strategy based on a statistical translation lexicon are provided in Section 5.1.3.

2.2.3 Tuple Vocabulary Pruning. The third and last issue regarding the *n*-gram translation model is related to the computational costs resulting from the tuple vocabulary size during decoding. The idea behind this refinement is to reduce both computation time and storage requirements without degrading translation performance. In our *n*-gram-based SMT system implementation, the tuple vocabulary is pruned by using histogram counts. This pruning is performed by keeping the *N* most frequent tuples with common source sides.

Notice that such a pruning, because it is performed before computing tuple *n*-gram probabilities, has a direct impact on the translation model probabilities and then on the overall system performance. For this reason, the pruning parameter *N* is critical for efficient usage of the translation system. While a low value of *N* will significantly decrease translation quality, on the other hand, a large value of *N* will provide the same translation quality than a more adequate *N*, but with a significant increment in computational costs. The optimal value for this parameter depends on data and should be adjusted empirically for each considered translation task.

2.3 *N*-gram-based Decoding

Decoding for the *n*-gram-based translation model is slightly different from phrase-based decoding. For this reason, a specific decoding tool had to be implemented. This section briefly describes MARIE, the *n*-gram based search engine developed for our SMT system (Crego, Mariño, and de Gispert 2005a).

MARIE implements a beam-search strategy based on dynamic programming. The decoding is performed monotonically and is guided by the source. During decoding, partial-translation hypotheses are arranged into different stacks according to the total number of source words they cover. In this way, a given hypothesis only competes with those hypotheses that provide the same source-word coverage. At every translation step, stacks are pruned to keep decoding tractable. MARIE allows for two different pruning methods:

- Threshold pruning: for which all partial-translation hypotheses scoring below a predetermined threshold value are eliminated.
- Histogram pruning: for which the maximum number of partial-translation hypotheses to be considered is limited to the *K*-best ranked ones.

Additionally, MARIE allows for hypothesis recombination, which provides a more efficient search. In the implemented algorithm, partial-translation hypotheses are recombined if they coincide exactly in both the present tuple and the tuple trigram history.

MARIE also allows for considering additional feature functions during decoding. All these models are taken into account simultaneously, along with the *n*-gram translation model. In our SMT system implementation, four additional feature functions are considered. These functions are described in detail in Section 3.2.

3. Feature Functions for the *N*-gram-based SMT System

This section describes in detail some feature functions that are implemented along with the *n*-gram translation model for the complete translation system. First, in subsection 3.1, the log-linear combination framework and the implemented optimization procedure are discussed. Then, four specific feature functions that constitute our SMT system are detailed in Section 3.2.

3.1 Log-linear Combination Framework

As mentioned in the Introduction, in recent translation systems the noisy channel approach has been replaced by a more general approach, which is founded on the principles of maximum entropy (Berger, Della Pietra, and Della Pietra 1996). In this approach, the corresponding translation for a given source language sentence S is defined by the target language sentence that maximizes a log-linear combination of multiple feature functions $h_i(S,T)$ (Och and Ney 2002), such as described by the following equation:

$$\underset{T}{\operatorname{argmax}} \sum_{m} \lambda_m h_m(S,T) \tag{2}$$

where λ_m represents the coefficient of the mth feature function $h_m(S,T)$, which actually corresponds to a log-scaled version of the mth-model probabilities. Optimal values for the λ_m coefficients are estimated via an optimization procedure by using a development data set.

3.2 Translation System Features

In addition to the tuple n-gram translation model, our n-gram-based SMT system implements four feature functions: a target-language model, a word-bonus model, and two lexicon models. These system features are described next.

3.2.1 Target-language Model. This feature provides information about the target language structure and fluency. It favors those partial-translation hypotheses that are more likely to constitute correctly structured target sentences over those that are not. The model is implemented by using a word n-gram model of the target language, which is computed according to the following expression:

$$h_{TL}(T, S) = h_{TL}(T) = \log \prod_{k=1}^{K} p(w_k | w_{k-1}, w_{k-2}, \ldots, w_{k-n+1}) \tag{3}$$

where w_k refers to the kth word in the considered partial-translation hypothesis. Notice that this model only depends on the target side of the data, and can in fact be trained by including additional information from other available monolingual corpora.

3.2.2 Word-bonus Model. This feature introduces a bonus that depends on the partial-translation hypothesis length. This is done to compensate for the system preference for short translations over large ones. The model is implemented through a bonus factor that directly depends on the total number of words contained in the partial-translation hypothesis, and it is computed as follows:

$$h_{WP}(T, S) = h_{WP}(T) = M \tag{4}$$

where M is the number of words contained in the partial-translation hypothesis.

3.2.3 Source-to-Target Lexicon Model. This feature actually constitutes a complementary translation model. This model provides, for a given tuple, a translation probability estimate between its source and target sides. This feature is implemented by using the IBM-1 lexical parameters (Brown et al. 1993; Och et al. 2004). Accordingly, the source-to-target lexicon probability is computed for each tuple according to the following equation:

$$h_{LF}(T, S) = \log \frac{1}{(I+1)^J} \prod_{j=1}^{J} \sum_{i=0}^{I} q(t_j^n | s_i^n) \tag{5}$$

where s_i^n and t_j^n are the ith and jth words in the source and target sides of tuple $(t, s)_n$, with I and J the corresponding total number of words in each side. In the equation, $q(.)$ refers to IBM-1 lexical parameters, which are estimated from alignments computed in the source-to-target direction.

3.2.4 Target-to-Source Lexicon Model. Similar to the previous feature, this feature function constitutes a complementary translation model too. It is computed in exactly the same way the previous model is, with the only difference that IBM-1 lexical parameters are estimated from alignments computed in the target-to-source direction instead.

4. EPPS Translation Task

This section describes in detail the most relevant issues about the translation tasks considered. Section 4.1 describes the EPPS data set that is used, and Section 4.2 presents the overall implementation details in regard to preprocessing, training, and optimization.

4.1 Corpus Description

The EPPS data set is composed of the official plenary session transcriptions of the European Parliament, which are currently available in eleven different languages (Koehn 2002). However, in the case of the results presented here, we have used the Spanish and English versions of the EPPS data that have been prepared by RWTH Aachen University in the context of the European Project TC-STAR. The training, development, and test data used include session transcriptions from April 1996 until September 2004, from October 21 until October 28, 2004, and from November 15 until November 18, 2004, respectively.

Table 1 presents the basic statistics for the training, development, and test data sets for each considered language. More specifically, the statistics shown in Table 1 are the number of sentences, the number of words, the vocabulary size (or number of distinct words), the average sentence length in number of words, and the number of available translation references.

As seen from Table 1, although the total number of words in the training set is very similar for both languages, vocabulary sizes are substantially different. Indeed, the Spanish vocabulary is approximately 60% larger than the English vocabulary. This can be explained by the more inflected nature of Spanish, which is particularly evident in the case of nouns, adjectives, and verbs, which may have many different forms depending on gender, number, tense, and mode. As will be seen from results presented in Section 5, this difference in vocabulary size has important consequences in translation quality for the English-to-Spanish direction.

Regarding the development data set, only 1, 008 sentences were considered. Notice from Table 1 that in this case, the Spanish vocabulary is 20% larger than the English

Table 1
Basic statistics for the training, development, and test data sets (M and k stand for millions and thousands, respectively; Lmean refers to the average sentence length in number of words, and Ref. to the number of available translation references).

Set	Language	Sentences	Words	Vocabulary	Lmean	Ref.
Train	English	1.22 M	33.4 M	105 k	23.7	1
	Spanish	1.22 M	34.8 M	169 k	28.4	1
Dev.	English	1008	26.0 k	3.2 k	25.8	3
	Spanish	1008	25.7 k	3.9 k	25.5	3
Test	English	1094	26.8 k	3.9 k	24.5	2
	Spanish	840	22.7 k	4.0 k	27.0	2

vocabulary. Another important issue regarding the development data set is the number of unseen words, that is, those words present in the development data that are not present in the training data. In this case, 35 words (0.13%) out of the total number of words in the English development set did not occur in the training data. From these 35 words, only 30 corresponded to different words. Similarly, 61 words (0.24%) out of the total number of words in the Spanish development set were not in the training data. In this case, 57 different words occurred.

Notice also in Table 1 that a different test set was used for each translation direction, and although a different number of sentences is considered in each case, vocabulary sizes are almost equivalent. Regarding unseen words, in this case, 112 words (0.42%) out of the total number of words in the English test set did not occur in the training data. From these 112 words, only 81 corresponded to different words. Similarly, 46 words (0.20%) out of the total number of words in the Spanish test were not in the training data. In this case, 40 different words occurred.

4.2 Preprocessing, Training, and System Optimization

This section presents the overall implementation details in regard to preprocessing, training, and optimization of the translation system. Two languages, English and Spanish, and both translation directions between them are considered for several different system configurations.

4.2.1 Preprocessing and Alignment. The training data are preprocessed by using standard tools for tokenizing and filtering. In the filtering stage, some sentence pairs are removed from the training data to allow for a better performance of the alignment tool. Sentence pairs are removed according to the following two criteria:

- Fertility filtering: removes sentence pairs with a word ratio larger than a predefined threshold value.
- Length filtering: removes sentence pairs with at least one sentence of more than 100 words in length. This helps to maintain bounded alignment computational times.

After preprocessing, word-to-word alignments are performed in both directions, source-to-target and target-to-source. In our system implementation, GIZA++ (Och and Ney 2000) is used for computing the alignments. A total of five iterations for models IBM-1 and HMM, and three iterations for models IBM-3 and IBM-4, are performed. Then, the obtained alignment sets are used for computing the intersection and the union of alignments from which tuples and embedded-word tuples are extracted, respectively.

4.2.2 Tuple Extraction and Pruning. A tuple set for each translation direction is extracted from the union set of alignments while avoiding source-nulled tuples by using the procedure described in Section 2.2.2. Then, the resulting tuple vocabularies are pruned according to the procedure described in Section 2.2.3. In the case of the EPPS data under consideration, pruning parameter values of $N = 20$ and $N = 30$ are used for Spanish-to-English and English-to-Spanish, respectively.

In order to better justify such alignment set and pruning parameter selections, Tables 2 and 3 present model sizes and translation accuracies for the tuple n-gram model

Table 2
Tuple vocabulary sizes and their corresponding number of n-grams (in millions), and translation accuracy when tuples are extracted from different alignment sets. Notice that BLEU measurements in this table correspond to translations computed by using the tuple n-gram model alone.

Direction	Alignment set	Tuple voc.	Bigrams	Trigrams	BLEU
ES → EN	Source-to-target	1.920	6.426	2.353	0.4424
	union	2.040	6.009	1.798	0.4745
	refined	2.111	6.851	2.398	0.4594
EN → ES	Source-to-target	1.813	6.263	2.268	0.4152
	union	2.023	6.092	1.747	0.4276
	refined	2.081	6.920	2.323	0.4193

when tuples are extracted from different alignment sets and when different pruning parameters are used, respectively. Translation accuracy is measured in terms of the BLEU score (Papineni et al. 2002), which is computed here for translations generated by using the tuple n-gram model alone, in the case of Table 2, and by using the tuple n-gram model along with the additional four feature functions described in Section 3.2, in the case of Table 3. Both translation directions, Spanish to English (ES → EN) and English to Spanish (EN → ES), are considered in each table.

In the case of Table 2, model size and translation accuracy are evaluated against the type of alignment set used for extracting tuples. Three different alignment sets are considered: source-to-target, the union of source-to-target and target-to-source, and the "refined" alignment method described by Och and Ney (2003). For the results presented in Table 2, a pruning parameter value of $N = 20$ was used for the Spanish-to-English direction, while a value of $N = 30$ was used for the English-to-Spanish direction.

As can be clearly seen in Table 2, the union alignment set happens to be the most favorable one for extracting tuples in both translation directions since it provides a significantly better translation accuracy, in terms of BLEU score, than the other two alignment sets considered. Notice also in Table 2 that the union set is the one providing the smallest model sizes according to the number of bigrams and trigrams. This might explain the improvement observed in translation accuracy, with respect to the other two cases, in terms of model sparseness.

Table 3
Tuple vocabulary sizes and their corresponding number of n-grams (in millions), and translation accuracy for different pruning values and both translation directions. Notice that BLEU measurements in this table correspond to translations computed by using the tuple n-gram model along with the additional four feature functions described in Section 3.2.

Direction	Pruning	Tuple voc.	Bigrams	Trigrams	BLEU
ES → EN	$N = 30$	2.109	6.233	1.805	0.5440
	$N = 20$	2.040	6.009	1.798	0.5434
	$N = 10$	1.921	5.567	1.759	0.5399
EN → ES	$N = 30$	2.023	6.092	1.747	0.4688
	$N = 20$	1.956	5.840	1.733	0.4671
	$N = 10$	1.843	5.342	1.677	0.4595

In the case of Table 3, model size and translation accuracy are compared for three different pruning conditions: $N = 30$, $N = 20$, and $N = 10$. For all the cases presented in the table, tuples were extracted from the union set of alignments.

Notice in Table 3 how translation accuracy is clearly affected by pruning. In the case of Spanish to English, values of $N = 20$ and $N = 10$, while providing tuple vocabulary reductions of 3.27% and 8.91% with respect to $N = 30$, respectively, produce a translation BLEU score reductions of 0.11% and 0.75%. On the other hand, in the case of English to Spanish, values of $N = 20$ and $N = 10$ provide tuple vocabulary reductions of 3.31% and 8.89% and a translation BLEU score reductions of 0.36% and 1.98% with respect to $N = 30$, respectively. According to these results, a similar tuple vocabulary reduction seems to affect English-to-Spanish translations more than it affects Spanish-to-English translations. For this reason, we finally adopted $N = 20$ and $N = 30$ as the pruning parameter values for Spanish to English and English to Spanish, respectively.

Another important observation derived from Table 3 is the higher BLEU score values with respect to the ones presented in Table 2. This is because, as mentioned above, the results presented in Table 3 were obtained by considering a full translation system that implements the tuple *n*-gram model along with the additional four feature functions described in Section 3.2. The relative impact of the described feature functions on translation accuracy is studied in detail in Section 5.1.1.

4.2.3 Translation Model and Feature Function Training. After pruning, a tuple *n*-gram model is trained for each translation direction by using the SRI Language Modeling toolkit (Stolcke 2002). The options for Kneser–Ney smoothing (Kneser and Ney 1995) and interpolation of higher and lower *n*-grams are used in these trainings. Then, each tuple *n*-gram translation model is finally enhanced by including the unigram probabilities for the embedded-word tuples such as described in Section 2.2.2.

Similarly, a word *n*-gram target language model is trained for each translation direction by using the SRI Language Modeling toolkit. Again, as in the case of the tuple *n*-gram model, Kneser–Ney smoothing and interpolation of higher and lower *n*-grams are used. Extended target language models might also be obtained by adding additional information from other available monolingual corpora. However, in the translation tasks described here, target language models are estimated by using only the information contained in the target side of the training data set.

In our SMT system implementation, trigram models are considered for both the tuple translation model and the target language model. This selection is based on perplexity measurements (over the development data set) obtained for *n*-gram models computed from the EPPS training data by using different *n*-gram sizes. Table 4 presents

Table 4
Perplexity measurements for translation and target language models of different *n*-gram sizes.

Type of model	Language	Bigram	Trigram	4-gram	5-gram
Translation	ES → EN	201.75	161.26	156.88	157.24
Translation	EN → ES	223.94	179.12	174.10	174.49
Language	Spanish	81.98	52.49	48.03	47.54
Language	English	78.91	50.59	46.22	45.59

perplexity values obtained for translation and target language models with different n-gram sizes.

Although our system implements trigram models, the performance of translation systems using different n-gram sized models is also evaluated. These results are presented and discussed in Section 5.1.2.

Finally, the source-to-target and target-to-source lexicon models are computed for each translation direction according to the procedure described in Section 3.2.3. For each considered lexicon model, either the alignment set in the source-to-target direction or the alignment set in the target-to-source direction is used, accordingly.

4.2.4 System Optimization. Once the models are computed, a set of optimal log-linear coefficients is estimated for each translation direction and system configuration via an optimization procedure, which is described as follows. First, a development data set that does not overlap either the training set or the test set is required. Then, translation quality over the development set is maximized by iteratively varying the set of coefficients. In our SMT system implementation, this optimization procedure is performed by using a tool developed in-house, which is based on a simplex method (Press et al. 2002), and the BLEU score (Papineni et al. 2002) is used as a translation quality measurement.

As will be described in the next section, several different system configurations are considered in the experiments. For all these optimizations, the development data described in Table 1 are used. As presented in the table, the development data included three translation references for both English and Spanish, which are used to compute the BLEU score at each iteration of the optimization procedures.

The same decoder settings are used for all system optimizations. These settings are the following:

- decoding is performed monotonically, that is, no reordering capabilities are used,
- decoding is guided by the source sentence to be translated,
- although available in the decoder, threshold pruning is not used, and
- a value of $K = 50$ for during-decoding histogram pruning is used.

5. Translation Experiments and Error Analysis

This section presents all translation experiments performed and a brief error analysis of the obtained results. In order to evaluate the relative contributions of different system elements to the overall performance of the n-gram-based translation system, three different experimental settings are considered. The experiments and their results are described in Section 5.1, and a brief error analysis of results is presented in Section 5.2. Finally, a comparison between n-gram-based SMT and state-of-the-art phrase-based translation systems is presented in Section 5.3.

5.1 Translation Experiments and Results

As already mentioned, three experimental settings are considered. For each setting, the impact on translation quality of a different system parameter is evaluated, namely, feature function, n-gram size, and the source-nulled tuple strategy. Evaluations in all three experimental settings are performed with respect to the same standard system configuration, which is defined in terms of the following parameters:

- Alignment set used for tuple extraction: UNION
- Tuple vocabulary pruning parameter: $N = 20$ for Spanish to English, and $N = 30$ for English to Spanish
- N-gram size used in translation model: 3
- N-gram size used in target language model: 3
- Expanded translation model with embedded-word tuples: YES
- Source-nulled tuple handling strategy: attach-to-right
- Feature functions considered: target language, word-bonus, source-to-target lexicon, and target-to-source lexicon

In the three experimental settings considered, which are presented in the following subsections, a total of seven different system configurations are evaluated in both translation directions, English to Spanish and Spanish to English. Thus, a total of 14 different translation experiments are performed. For each of these cases, the corresponding test set is translated by using the corresponding estimated models and set of optimal coefficients. The same decoder settings (which were previously described in Section 4.2.4) that were used during the optimizations are used for all translation experiments. Translation results are evaluated in terms of mWER and BLEU by using the two references available for each language test set.

5.1.1 Feature Function Contributions. This experiment is designed to evaluate the relative contribution of feature functions to the overall system performance. In this section, four different systems are evaluated. These systems are:

- System A. This constitutes the basic n-gram translation system, which implements the tuple trigram translation model alone, that is, no additional feature function is used.
- System B. This is a target-reinforced system. In this system, the translation model is used along with the target-language and word-bonus models.
- System C. This is a lexicon-reinforced system. In this system, the translation model is used along with the source-to-target and target-to-source lexicon models.
- System D. This constitutes the full system, that is, the translation model is used along with all four additional feature functions. This system corresponds to the standard system configuration that was defined at the beginning of Section 5.1.

Table 5 summarizes the results of this evaluation, in terms of BLEU and mWER, for the four systems considered. As can be seen from the table, both translation directions,

Table 5
Evaluation results for experiments on feature function contribution.

Direction	System	λ_{lm}	λ_{wb}	λ_{s2t}	λ_{t2s}	mWER	BLEU
ES → EN	A	–	–	–	–	39.71	0.4745
	B	0.29	0.31	–	–	39.51	0.4856
	C	–	–	0.77	0.08	35.77	0.5356
	D	0.49	0.30	0.94	0.25	34.94	0.5434
EN → ES	A	–	–	–	–	44.46	0.4276
	B	0.33	0.27	–	–	44.67	0.4367
	C	–	–	0.29	0.15	41.69	0.4482
	D	0.66	0.73	0.32	0.47	40.34	0.4688

Spanish to English and English to Spanish, are considered. Table 5 also presents the optimized log-linear coefficients associated with the features considered in each system configuration (the log-linear weight of the translation model has been omitted from the table because its value is fixed to 1 in all cases).

As can be observed in Table 5, the inclusion of the four feature functions into the translation system definitively produces a significant improvement in translation quality in both translation directions. In particular, it becomes evident that the features with the most impact on translation quality are the lexicon models. The target language model and the word bonus also contribute to improving translation quality, but to a lesser degree.

Also, although it is more evident in the English-to-Spanish direction than in the opposite one, it can be noticed from the presented results that the contribution of target-language and word-bonus models is more relevant when the lexicon models are used (full system). In fact, as seen from the λ_{lm} values in Table 5, when the lexicon models are not included, the target-language model contribution to the overall translation system becomes much less significant. A comparative analysis of the resulting translations suggests that including the lexicon models tends to favor short tuples over long ones, so the target-language model becomes more important for providing target context information when the lexicon models are used. However, more experimentation and research are required for fully understanding this interesting result.

Another important observation, which follows from comparing results between both translation directions, is that in all cases the Spanish-to-English translations are consistently and significantly better than the English-to-Spanish translations. This is clearly due to the more inflected nature of Spanish vocabulary. For example, the single English word *the* can generate any of the four Spanish words *el*, *la*, *los*, and *las*. Similar situations occur with nouns, adjectives, and verbs that may have many different forms in Spanish. This would suggest that the English-to-Spanish translation task is more difficult than the Spanish-to-English task.

5.1.2 Translation and Language *N*-gram Size. This experiment is designed to evaluate the impact of translation- and language-model *n*-gram sizes on overall system performance. In this section, the full system (System D in the previous experiment) is compared with two similar systems for which 4-grams are used for training the translation model and/or the target language model. More specifically, the three systems compared in this experiment are:

- System D, which implements a tuple trigram translation model and a word trigram target language model. This system corresponds to the standard system configuration that was defined at the beginning of Section 5.1.
- System E, which implements a tuple trigram translation model and a word 4-gram target language model.
- System F, which implements a tuple 4-gram translation model and a word 4-gram target language model.

Table 6 summarizes the results of this evaluation for Systems E, F, and D. Again, both translation directions are considered and the optimized coefficients associated with the four feature functions are also presented for each system configuration.

As can be seen in Table 6, the use of 4-grams for model computation does not provide a clear improvement in translation quality. This is more evident in the English-to-Spanish direction for which System F happens to be the worst ranked one, while System D is the one obtaining the best mWER score and system E is the one obtaining the best BLEU score. On the other hand, in the Spanish-to-English direction, it seems that a little improvement with respect to System D is achieved by using 4-grams. However, it is not clear which system performs the best since System E obtains the best BLEU score while System F obtains the best mWER score.

According to these results, more experimentation and research are required to fully understand the interaction between the *n*-gram sizes of translation and target language models. Notice that in the particular case of the *n*-gram SMT system described here, such an interaction is not evident at all since the *n*-gram-based translation model itself contains some of the target language model information.

5.1.3 Source-nulled Tuple Strategy Comparison. This experiment is designed to evaluate a different strategy for handling source-nulled tuples. In this section, the standard system configuration (System D) presented at the beginning of Section 5.1, which implements the attach-to-right strategy described in Section 2.2.2, is compared with a similar system (referred to as System G) implementing a more complex strategy for handling those tuples with NULL source sides. More specifically, the latter system uses the IBM-1 lexical parameters (Brown et al. 1993) for computing the translation probabilities of two possible new tuples: the one resulting when the null-aligned-word is attached to

Table 6
Evaluation results for experiments on *n*-gram size incidence.

Direction	System	λ_{lm}	λ_{wb}	λ_{s2t}	λ_{t2s}	mWER	BLEU
ES → EN	D	0.49	0.30	0.94	0.25	34.94	0.5434
	E	0.50	0.54	0.66	0.45	34.66	0.5483
	F	0.66	0.50	1.01	0.57	34.59	0.5464
EN → ES	D	0.66	0.73	0.32	0.47	40.34	0.4688
	E	0.57	0.45	0.51	0.26	40.55	0.4714
	F	1.24	1.07	0.99	0.57	40.91	0.4688

the previous word and the one resulting when it is attached to the following one. Then, the attachment direction is selected according to the tuple with the highest translation probability.

Table 7 summarizes the results of evaluation Systems D and G. Again, both translation directions are considered and the optimized coefficients associated with the four feature functions are also presented for each system configuration.

As can be seen in Table 7, consistently better results are obtained in both translation tasks when using IBM-1 lexicon probabilities to handle tuples with a NULL source side. Even though slight improvements are achieved in both cases, especially with the English-to-Spanish translation task, the results show how the initial attach-to-right strategy is easily improved by making use of some bilingual knowledge.

5.2 Error Analysis

In this last section, we present a brief description of an error analysis performed on some of the outputs provided by the standard system configuration that was described in Section 5.1 (system D). More specifically, a detailed review of 100 translated sentences and their corresponding source sentences, in each direction, was conducted. This analysis was very useful since it allowed us to identify the most common errors and problems related to our *n*-gram based SMT system in each translation direction.

A detailed analysis of all the reviewed translations reveals that most translation problems encountered are typically related to four basic different types of errors:

- Verbal forms: A significant number of wrong verbal tenses and auxiliary forms were detected. This problem turned out to be the most common one, reflecting the difficulty of the current statistical approach to capture the linguistic phenomena that shape head verbs, auxiliary verbs, and pronouns into full verbal forms in each language, especially given the inflected nature of the Spanish language.
- Omitted translations: A large number of translations involving tuples with NULL target sides were detected. Although in some cases these situations corresponded to correct translations, most of the time they resulted in omitted-word errors.
- Reordering problems: The two specific situations that most commonly occurred were problems related to adjective–noun and subject–verb structures.

Table 7
Evaluation results for experiments on strategies for handling source-nulled tuples.

Direction	System	λ_{lm}	λ_{wb}	λ_{s2t}	λ_{t2s}	mWER	BLEU
ES → EN	D	0.49	0.30	0.94	0.25	34.94	0.5434
	G	0.49	0.45	0.78	0.39	34.15	0.5451
EN → ES	D	0.66	0.73	0.32	0.47	40.34	0.4688
	G	0.96	0.93	0.53	0.44	40.12	0.4694

- Concordance problems: Inconsistencies related to gender and number were the most commonly found.

Table 8 presents the relative number of occurrences for each of the four types of errors identified in both translation directions.

Notice in Table 8 that the most common errors in both translation directions are those related to verbal forms. However, it is important to mention that 29.5% of verbal-form errors in the English-to-Spanish direction actually correspond to verbal omissions. Similarly, 12.8% of verbal-form errors in the Spanish-to-English direction are verbal omissions. According to this, if errors due to omitted translations and to omitted verbal forms are considered together, it is evident that errors involving omissions constitute the most important group, especially in the case of English-to-Spanish translations. It is also interesting to note that the Spanish-to-English direction exhibits more omitted-translation errors that are not related to verbal forms than the English-to-Spanish direction.

Also in Table 8, it can be seen that concordance errors affect more than twice as many English-to-Spanish translations as Spanish-to-English ones. This result can be explained by the more inflected nature of Spanish.

Finally, as an illustrative example, three Spanish-to-English translation outputs are presented below. For each presented example, errors have been boldfaced and correct translations are provided in brackets:

Example 1
The policy of the European Union on Cuba **NULL must** *[must not] change.*

Example 2
To achieve these purposes, it is necessary NULL for the governments **to be allocated** *[to allocate], at least, 60,000 million NULL dollars a year ...*

Example 3
In the UK we have **NULL** *[already]* **laws enough** *[enough laws], but we want to encourage NULL other States ...*

5.3 *N*-gram-based SMT Compared with Phrase-Based SMT

The *n*-gram-based translation system here described has been also evaluated and compared to other phrase-based translation systems in the context of the European Project

Table 8
Percentage of occurrence for each type of error in English-to-Spanish and Spanish-to-English translations that were studied.

Type of error	English-to-Spanish	Spanish-to-English
Verbal forms	31.3%	29.9%
Omitted translations	22.0%	26.1%
Reordering problems	15.9%	19.7%
Concordance problems	10.8%	4.6%
Other errors	20.0%	19.7%

TC-STAR. A detailed description of the first evaluation campaign (including the main characteristics of every system) is available through the consortium's Web site as a progress report (Ney et al. 2005).

Table 9 presents the four best BLEU results for the EPPS translation task in the first TC-STAR's evaluation campaign, where the results corresponding to our *n*-gram-based translation system are provided in brackets. A total of six systems were evaluated in this evaluation campaign. The task consisted of two translation directions: English to Spanish and Spanish to English, and three different evaluation conditions: final text edition, verbatim, and ASR output. The final text edition condition corresponds to the official transcripts of the EPPS, so it is actually a written-language translation condition. On the other hand, the other two conditions are spoken-language translation conditions. More specifically, the verbatim condition corresponds to literal transcriptions of parliamentary speeches, which include hesitations, repeated words, and other spontaneous speech effects; and the ASR output condition corresponds to the output of an automatic speech recognition system, so it additionally includes speech-recognition errors.

As can be seen in Table 9, performance of the *n*-gram-based translation system is among the three best systems for the translation directions and conditions considered in the first TC-STAR evaluation campaign.

Another independent comparison of the translation system proposed here with other phrase-based translation systems is available through the results of the second shared task of the ACL 2005 workshop on "Building and using parallel texts: Data-driven machine translation and beyond." In this shared task, which was entitled "Exploiting Parallel Texts for Statistical Machine Translation," our *n*-gram-based translation system was evaluated in four different translation directions: Spanish to English, French to English, German to English, and Finish to English (Banchs et al. 2005). The domain of this task was also the European Parliament; however, the data set considered in this evaluation was different from the one used in TC-STAR's evaluation campaign. The final text edition condition (official transcripts) was the only one considered here. A total of twelve different systems participated in this shared task. Table 10 presents the four best BLEU results for each of the four translation directions considered in the shared task. Again, results corresponding to our *n*-gram-based translation system are provided in brackets.

As can be seen in Table 10, the performance of the *n*-gram-based translation system is among the three best systems for the four translation directions considered in the ACL 2005 workshop shared task. The third system in Table 10 for ES to EN translation

Table 9
The four best BLEU results for the EPPS translation task in TC-STAR's first evaluation campaign. *N*-gram based system results are provided in brackets. All BLEU values presented here have been taken from TC-STAR's SLT Progress Report, available at: http://www.tc-star.org/.

Direction	Condition	First	Second	Third	Fourth
ES → EN	Final text edition	[53.3]	53.1	47.5	46.1
	Verbatim	45.9	44.1	[42.1]	38.1
	ASR output	41.5	39.7	[37.7]	34.7
EN → ES	Final text edition	[46.2]	45.2	38.9	37.6
	Verbatim	42.5	[38.1]	36.8	33.4
	ASR output	38.7	34.3	[33.8]	33.0

Table 10
The four best BLEU results for the four translation directions considered in the shared task "Exploiting Parallel Texts for Statistical Machine Translation" (ACL 2005 workshop on "Building and using parallel texts: Data-driven machine translation and beyond"). *N*-gram-based system results are provided in brackets. All BLEU values presented here have been taken from the shared task's Web site: http://www.statmt.org/wpt05/mt-shared-task/.

Direction	Condition	First	Second	Third	Fourth
FR → EN	Final text edition	30.27	[30.20]	29.53	28.89
ES → EN	Final text edition	30.95	[30.07]	29.84	29.08
DE → EN	Final text edition	24.77	[24.26]	23.21	22.91
FI → EN	Final text edition	22.01	20.95	[20.31]	18.87

deserves some comment. This system is a conventional phrase-based system sharing the same decoder MARIE, IBM features, word bonus, and target-language model as the *n*-gram-based system. The specific characteristics of the phrase-based system are direct and inverse phrase conditional probabilities and phrase penalty. Additional comparisons between an *n*-gram system and a phrase-based system sharing a common decoder and training and test framework can be found in Crego et al. (2005c).

6. Conclusions and Further Work

As can be concluded from the results presented, the tuple *n*-gram translation model, when used along with additional feature functions, provides state-of-the-art translations for the considered translation directions.

Another important result is that the quality of Spanish-to-English translations is significantly and consistently better than those obtained in English-to-Spanish translations. Consequently, significant efforts should be dedicated towards properly exploiting morphological analysis and synthesis methods for improving English-to-Spanish translation quality.

Additionally, four commonly occurring types of translation errors were identified by reviewing a significant number of translated sentence pairs. This analysis has provided us with useful hints for future research and improvement of our SMT system. However, more evaluation and discussion are required in this area in order to fully understand these common translation failures and then implementing appropriate solutions.

All the experiments presented in this work were performed using monotone decoding, and no reordering strategies were implemented. Although this system configuration proved to provide state-of-the-art translations for the tasks presented, this may not hold for tasks involving more distant language pairs for which reordering capabilities must be implemented. Accordingly, along with other results obtained in the present work, we consider that further research on *n*-gram SMT should focus on the following issues:

- Reordering strategies, as well as non-monotonous decoding schemes, for the proposed SMT system must be developed and tested. As mentioned before, reordering problems specifically related to adjective–noun and subject–verb structures occur very often in Spanish-to-English and

English-to-Spanish translations. Preliminary results concerning the use of word class deterministic reordering and POS-tag-based reordering patterns can be found in Costa-jussà, Fonollosa, and Monte (2006) and Crego and Mariño (2006), respectively.

- An effective long-tuple unfolding strategy must be developed to avoid the occurrence of long tuples resulting from long alignment links, which happens to be a common situation when dealing with translations between distant pairs of languages. This problem is closely related to reordering, and some preliminary results have been presented by Crego, Mariño, and de Gispert (2005b).
- The definition of the tuple as a bilingual pair will be revised in order to better handle unaligned words in both the source and the target sides. As mentioned above, a better strategy for dealing with target words aligned to NULL is required. Similarly, a better handling of NULLs in the target side will result in fewer omitted-translation errors.
- The extension of the embedded-word concept to the more general idea of embedded *n*-grams should be evaluated and implemented. Accordingly, a translation probability should be estimated for those groups of words that always occur embedded in tuples. This would guarantee that the decoder will always have a translation option for any given word or word combination previously seen in the training data. Further work is required to determine the relative impact of these embedded *n*-grams on the translation model, and the most appropriate strategy for handling them.
- Linguistic information must be used to cope with the observed morphological problems in the English-to-Spanish translation direction, as well as the more general problem of incorrect verbal form translations. In this regard, ongoing research on linguistic tuples classification is being done in order to improve translation results. Preliminary results on detecting and classifying verb forms have been presented by de Gispert (2005).
- A more detailed error analysis than the one presented in Section 5.2 is required to fully understand the *n*-gram SMT system behavior and the specific causes of each resulting type of error. It would be very useful for improving our translation system performance to clearly identify whether these errors are due to unseen information while training, to modeling problems, or to decoding errors.

Acknowledgments

This work has been partly 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 Department of Education and Science (MEC), the Department of Universities, Research and Information Society (Generalitat de Catalunya), and the Universitat Politècnica de Catalunya.

References

Banchs, Rachel E., Josep Maria Crego, Adrià de Gispert, Patrik Lambert, and José Bernardo Mariño. 2005. Statistical machine translation of Euparl data by using bilingual *n*-grams. In *ACL Workshop on Data-Driven Machine Translation and Beyond*, pages 133–136, Ann Arbor, MI.

Bangalore, Srinivas and Giuseppe Riccardi. 2000. Stochastic finite-state models for spoken language machine translation.

In *Proceedings of the Workshop on Embedded Machine Translation Systems*, pages 52–59, Seattle, WA.

Berger, Adam, Stephen Della Pietra, and Vincent Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

Brown, Peter, John Cocke, Stephen Della Pietra, Vincent Della Pietra, Frederick Jelinek, John Lafferty, Robert Mercer, and Paul S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85.

Brown, Peter, Stephen Della Pietra, Vincent Della Pietra, and Robert Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Casacuberta, Francisco. 2001. Finite-state transducers for speech input translation. In *Proceedings IEEE ASRU*, pages 375–380, Madonna di Campiglio, Italy.

Casacuberta, Francisco and Enrique Vidal. 2004. Machine translation with inferred stochastic finite-state transducers. *Computational Linguistics*, 30(2):205–225.

Costa-jussà, Marta Ruiz, José Adrián Rodriguez Fonollosa, and Enric Monte. 2006. Using reordering in statistical machine translation based on alignment block classification. Internal Report. http://gps-tsc.upc.es/veu/personal/mruiz/docs/br06.pdf.

Crego, Josep Maria, José Bernardo Mariño, and Adrià de Gispert. 2004. Finite-state-based and phrase-based statistical machine translation. In *Proceedings of the 8th International Conference on Spoken Language Processing*, pages 37–40, Jeju, Korea.

Crego, Josep Maria, José Bernardo Mariño, and Adrià de Gispert. 2005a. An Ngram-based statistical machine translation decoder. In *INTERSPEECH 2005*, pages 3185–3188, Lisbon, Portugal.

Crego, Josep Maria, José Bernardo Mariño, and Adrià de Gispert. 2005b. Reordered search and tuple unfolding for Ngram-based SMT. *Proceedings of the Tenth Machine Translation Summit*, pages 283–289, Phuket, Thailand.

Crego, Josep Maria, Marta Ruiz Costa-jussà, José Bernardo Mariño, and José Adrián Rodriguez Fonollosa. 2005c. Ngram-based versus phrase-based statistical machine translation. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 177–184, Pittsburgh, PA.

Crego, Josep Maria and José Bernardo Mariño. 2006. Integration of POStag-based source reordering into SMT decoding by an extended search graph. In *Proceedings of the 7th Biennial Conference of the Association for Machine Translation in the Americas*, Boston, MA.

de Gispert, Adrià and José Bernardo Mariño. 2002. Using X-grams for speech-to-speech translation. In *Proceedings of the 7th International Conference on Spoken Language Processing*, pages 1885–1888, Denver, CO.

de Gispert, Adrià, José Bernardo Mariño, and Josep Maria Crego. 2004. TALP: Xgram-based spoken language translation system. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 85–90, Kyoto, Japan.

de Gispert, Adrià. 2005. Phrase linguistic classification and generalization for improving statistical machine translation. In *ACL'05 Student Workshop*, pages 67–72, Ann Arbor, MI.

Hutchins, John. 1986. *Machine Translation: Past, Present and Future*. Ellis Horwood, Chichester, England.

Kay, Martin, Jean Mark Gawron, and Peter Norvig. 1992. *Verbmobil: A Translation System for Face-to-Face Dialog*. CSLI.

Kneser, Reinhard and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 49–52, Detroit, MI.

Knight, Kevin and Yaser Al-Onaizan. 1998. Translation with finite-state devices. In *AI Lecture Notes in Artificial Intelligence*, volume 1529, Springer-Verlag, pages 421–437.

Koehn, Philippe, Franz Joseph Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Meeting of the North American chapter of the ACL*, pages 48–54, Edmonton, Alberta, Canada.

Koehn, Philippe. 2002. Europarl: A multilingual corpus for evaluation of machine translation. Available online at: http://people.csail.mit.edu/people/koehn/publications/europarl/.

Mariño, José Bernardo, Rafael E. Banchs, Josep Maria Crego, Adrià de Gispert, Patrik Lambert, José Adrián Rodriguez Fonollosa, and Marta Ruiz. 2005. Bilingual N-gram statistical machine translation. In *Proceedings of the Tenth Machine Translation Summit*, pages 275–282, Phuket, Thailand.

Ney, Hermann, Volker Steinbiss, Richard Zens, Evgeny Matusov, Jorge González, Young-suk Lee, Salim Roukos, Marcello Federico, Muntsin Kolss, and Rafael Banchs. 2005. SLT progress report. *TC-STAR Deliverable D5,* European Community project no. FP6-506738. Available online at: http://www.tc-star.org/pages/f_documents.htm.

Och, Franz Joseph and Hermann Ney. 2000. Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting of the ACL*, pages 440–447, Hong Kong, China.

Och, Franz Joseph and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the 40th Annual Meeting of the ACL*, pages 295–302, Philadelphia, PA.

Och, Franz Joseph and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Och, Franz Joseph, Daniel Gildea, Sanjeev Khudanpur, Anoop Sarkar, Kenji Yamada, Alexander Fraser, Shankar Kumar, Libin Shen, David Smith, Katharine Eng, Viren Jain, Zhen Jin, and Dragomir Radev. 2004. A smorgasbord of features for statistical machine translation. In *Proceedings of the Human Language Technology Conference NAACL*, pages 161–168, Boston, MA, May.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Conference of the ACL*, pages 311–318, Philadelphia, PA.

Press, William H., Saul Teukolsky, William Vetterling, and Brian P. Flannery. 2002. *Numerical Recipes in C++: The Art of Scientific Computing*, Cambridge University Press.

Riccardi, Giuseppe, Roberto Pieraccini, and Enrico Bocchieri. 1996. Stochastic automata for language modeling. *Computer Speech and Language*, 10(4):265–293.

Shannon, Claude E. 1949. Communication theory of secrecy systems. *Bell System Technical Journal*, 28:656–715.

Shannon, Claude E. 1951. Prediction and entropy of printed English. *Bell System Technical Journal*, 30:50–64.

Shannon, Claude E. and Warren Weaver. 1949. *The Mathematical Theory of Communication*, University of Illinois Press, Urbana, IL.

Stolcke, Andreas 2002. SRLIM: An extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, pages 901–904, Denver, CO.

Tillmann, Christoph and Fei Xia. 2003. A phrase-based unigram model for statistical machine translation. In *Proceedings of HLT-NAACL - Short Papers*, pages 106–108, Edmonton, Alberta, Canada.

Vidal, Enrique. 1997. Finite-state speech-to-speech translation. In *Proceedings of 1997 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 111–114, Munich, Germany.

Weaver, Warren. 1955. Translation. In William Locke and A. Donald Booth, editors, *Machine Translation of Languages: Fourteen Essays*. John Wiley & Sons, New York, pages 15–23.

Zens, Richard, Franz Joseph Och, and Hermann Ney. 2002. Phrase-based statistical machine translation. In *25th German Conference on Artificial Intelligence*, pages 18–32, September. Aachen, Springer Verlag.