Departament de Teoria del Senyal i Comunicacions
UNIVERSITAT POLITÈCNICA DE CATALUNYA

N-gram model improvement for Statistical Machine Translation

(PhD Thesis Proposal)

Maxim Khalilov
khalilov@gps.tsc.upc.edu

Thesis Advisor: Dr. José Adrián Rodríguez Fonollosa
adrian@gps.tsc.upc.edu

TALP Research Center
Department of Signal Theory and Communications
Universitat Politècnica de Catalunya
Campus Nord UPC
C/ Jordi Girona 1-3, Building D5
08034 Barcelona (Spain)

February, 2007

Contents

1 Introduction

The PhD proposal is dedicated to statistical machine translation (SMT), which is one of the methods of machine translation (MT). MT is a field of computational linguistics investigating translation of texts from one human language to another. The proposal is organized as follows: in Chapter 1 the statement of the translation problem, PhD objectives and some historical details of MT are outlined; in Chapter 2 a review of the literature, as well as backgrounds of SMT are presented; Chapter 3 concludes the proposal by describing research methodology and a working plan.

1.1 Motivation

Our world is going through a globalization period, which means an increasing of interaction and intertwining between different language communities. Information globalization extends to all corners of the world, and doubtless multilinguality should be a strategic issue for all the companies willing to play an important role in the upcoming society of information.

Present-day informational world can be easily characterized by broad accessibility to a great number of information resources coming from all over the world and presented in various languages. With a lack of fast and qualitative translation, one confronts the language barrier hampering further information penetration into the multilingual society. It would be utopian to believe that at the modern point of informational society evolution, MT could completely substitute human-made translation. However, the majority of the work made by professional translators are routine and non-literary translations, which are not of a great cultural value and certainly this field tends to be the most attractive for MT application.

Nowadays there is MT software available, however there is a long way to go while the MT systems achieve high-quality translation. Moreover, these imperfect MT systems can and are being used by millions of users to translate web-pages and routine everyday documents where translation quality is not crucial and the main goal is to give to the user an idea of the content.

The Kaija Poysti's statement ”you can always buy in your own language, but you must sell in your customer's language” [1] is coming to be more and more true these days. The new conception of social communications must include engaging customers of commercial companies or users of any informational source regardless of geography and cultural expectations. Although English is becoming the universal second language, users in general still feel more comfortable dealing in their own native language.

The European Union (EU) is crucially attractive for MT, since it is an institution having high demand in translation (after the 1 of January 2007 there are 23 official EU working languages) and sufficient funding potential to support scientific research in the MT sphere.

1.2 Machine Translation

The fundamental of MT was laid in 1955 when Warren Weaver published his very important paper [1], which is considered as a starting point of the modern MT. This work was based on information theory and inspired by successes of code breaking during the Second World War [2].

[1] Globalization Blog: http://gilbane.com/globalization/

Further enthusiasm faded after the ALPAC report in 1966 [3] was published, which demonstrated that real progress was very poor and ten years of research gave results incommensurable with the spent funds.

MT is one of the most complicated tasks of Natural Language Processing, besides that there is not only one perfect translation of the source, there are a lot of factors influencing translation task, as well as a great number of hardly formalizable (or even not formalizable at all) dependencies.

From the 1950s and till the earlier 1990s research in the field of MT was sufficiently restricted apart from some specifically motivated works like Russian-English translation stimulated by the US and Soviet governments because of the Cold War or METEO English-French Translation System [4], developed at the Universitè de Montrèal to translate weather forecasts.

MT got a new impulse in the 1990s when first SMT systems were developed. Their appearance was the result of the tremendous progress made in computer technology and software engineering over the last years, that time MT began to be an application of personal computers and workstations in contrast with mainframes. About that time, IBM commenced to develop one of the first full-scale SMT unlike previous approaches to MT, SMT performs translations generated on the basis of statistical models based on the data derived from the analysis of *bilingual text corpora* (the collection of text and their reference translations), later expanded with additional morphologic, syntactic and semantic information, complementing fundamental models. The following works were stimulated by growing availability of parallel corpora allowing extracting valuable information for the given language pair.

There are several methodologies of MT classification [5], the most popular one is based on the level of linguistic analysis it performs. The MT Pyramid suggested by [6] specifies a way of processing comparable to that used by the human translator and is presented in Figure 1.

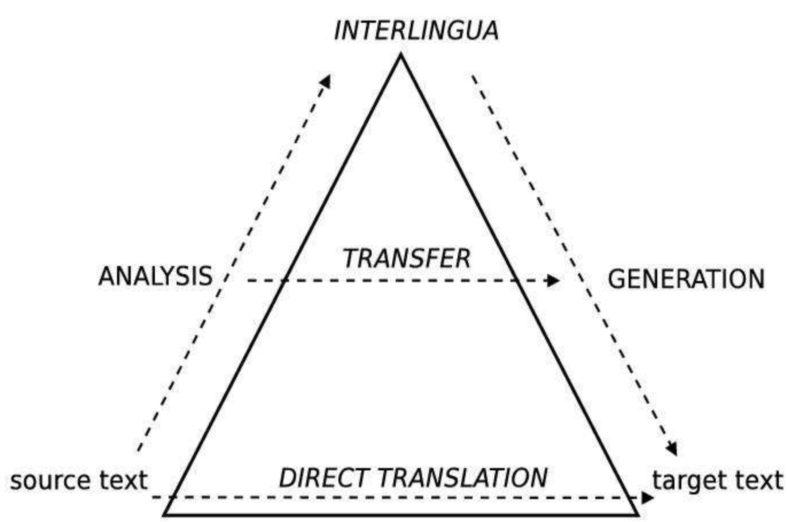

Figure 1: *Machine translation pyramid.*

"Direct" translation approach represents translation without performing any linguistic analysis at all, it becomes more complicated going up to the top of the pyramid. "Transfer" system realize deeper level of intermediary representation, usually on the morphosyntactic level. On the top of the pyramid semantic analysis of the source language text is provided, this information is used further to generate target language text.

Considering system design criteria, MT systems can be divided as shown in the Figure 2.

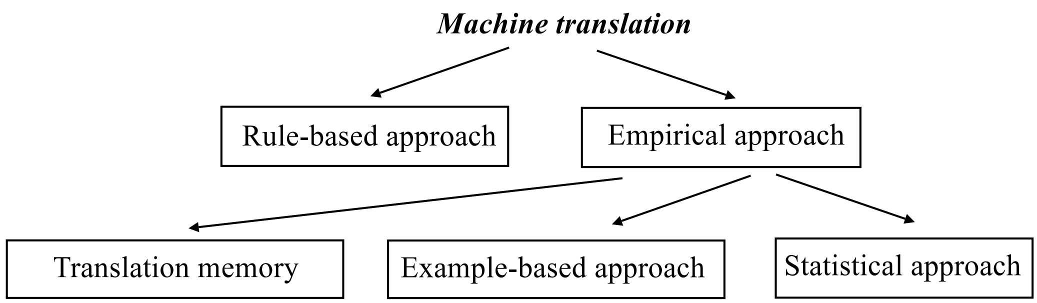

Figure 2: *Machine translation pyramid.*

Rule-based systems are accepted as a classical approach to MT. Translation systems based on this approach use the set of linguistic rules (normally made by human experts), specifically describing the translation process.

Empirical approach (EA) to MT, appeared in the beginning of the 1990s, is a new paradigm to challenge and enrich established rule-based MT. This approach is based on the parallel corpora, and the reusing of examples of already existing translations to generate a final translation.

There are three mostly important classes of EA systems: *Example-based MT (EMT)*, *Translation memory* and *Statistical MT*. EMT and Translation memory both deal with finding and matching examples that contribute to the translation on the basis of their similarity with the input sequence. They differ in that in Translation Memory the example extraction stage is carried out by humans, while EMT systems provide it automatically.

The essence of the *SMT* method is to generate translations using statistical models, which parameters are estimated on the basis of bilingual text corpora.

Academic research in the area of MT has mainly been declined in the area of written language MT, while a single area of SMT application lies next to *automatic speech recognition (ASR)* field, namely in the task of speech translation. In contrast with the area of written language MT, speech transcription is characterized by special peculiarities typical for recognition tasks, like noise, ingrammatical input, spontaneous speech phenomena etc. There is much interest in exploring new techniques in SMT and ASR integrating as it is indicated in [7], [8] and [9]. The final target of the SMT/ASR integrating approach is to develop a speaker-independent real-time translation system (probably, domain-oriented).

1.3 Objectives of the thesis

The main goal of the thesis consists in the development of a new architecture of the *n-gram* translation system in order to improve the translation accuracy. More comprehensively, the pursuing objectives of the work can be divided in three groups:

- *Target language modeling.* The task of language modeling is fundamental to speech and optical character recognition and extremely important for a lot of NLP applications. Normally, *the language model (LM)* is implemented using *n-grams*, this approach was inspired by the speech recognition fields, where n-grams have been successfully applied. Intuitively, there are several possibilities of LM modification with the aim of translation quality improvement (smoothing techniques, cut-off threshold selection, class-based LMs, specific target language modeling techniques as "skip" models, etc.). Special attention is going to be focused on the factored modeling methods [10], which are mainly planing to be applied to bilingual translation modeling. Moreover, in the framework of the thesis some special techniques are going to be considered, mainly dedicated to LM adaptation to specific translation tasks (ASR output or verbatim text translation in contrast to the final text edition (FTE) task).

- *Translation modeling.* N-gram (tuples) *translation model (TM)* is a relatively new and insufficiently investigated approach to SMT. TM is implemented using n-grams in the same manner as LM but handling bilingual units. Hence, various techniques successfully applied to the LM promise to be used in the TM estimation.

  This part of the work is mainly dedicated to the following aspects of the n-gram model improvement, namely:

  - TM factorization (as presented in [10] for language modeling for the phrase-based systems), i.e. a representation of tuples as feature vectors and thus allowing for additional information utilization in a unified or principled framework. This method have already proved beneficial in ASR and in the language modeling for SMT.
  - Another emphasis of the work lies in introducing hierarchical models based on bilingual synchronous grammars into the translation system. The idea lies in synchronous context-free grammar (SynCFG) rules generation, extracting them from the phrase or n-gram translation tables. As it was shown in [11], the grammar rules can be learned from a bilingual corpus without any syntactic information, or as in [12], where parse trees are used.
  - Apart from the existing methods, some other novel techniques of TM improvement are going to be proposed and duly investigated to cope with the ambiguity in the translation of source language phrases.

- *Syntactically motivated word-reordering technique.* In SMT, the use of reordering strategies allows an important improvement in translation accuracy, specially when the translation between language pairs with high disparity in word order. Pure classical approach to the SMT does not take into account syntactical information of the analyzing corpora. But, it seems logical to try to incorporate these additional sources of knowledge into the translation system, especially for highly-inflected languages (like English). There are a several successful attempts to solve the problem of the introducing of linguistic knowledge in the SMT, more details can be found in [13].

  In the framework of the thesis, we plan to introduce lexical information (by means of part-of-speech tags (POS)) and syntactical information (by syntactic trees (parsing/dependency trees)) to improve word reordering. For this purpose, I am planing to follow the approach presented in [14], where a preprocessing step determines a weighted reordering graph that is then incorporated into the final decoding step.

2 State-of-the-art

The idea of SMT lies in the translation of a source sentence f (traditionally referred to French) into a sentence in the target language e (English). The problem is formulated in terms of source and target languages and is defined as arg max operation, as described by the following equation (1):

$$\hat{e}_1^I = \arg\max_{e_1^I} \left\{ p(e_1^I \mid f_1^J) \right\} \tag{1}$$

where I and J represent the number of words of the sentences in target and source languages. Hence, translation problem can be reformulated as a selecting translation with highest probability among the set of target sentences.

This technique of parallel corpus processing was used to decrypt the sign on the Rosetta stone, the basalt stone found by French army in Egypt in 1799. It contained the same text in three languages, two of them were not known (ancient Egyptian dialects). Thanks to some words appeared in the well-known ancient Greek, the unknown languages were decrypted [15].

Modern SMT originates from the work realized in IBM in the earlies 1990s, basically this research was inspired by experiments made on the speech recognition field. First SMT systems were based on the so-called noisy-channel (or source-channel) approach [16]. Due to this approach, equation 1 can be decomposed according to the Bayes rule as follows:

$$\hat{e}_1^I = \arg\max_{e_1^I} \left\{ p(f_1^J \mid e_1^I) \cdot p(e_1^I) \right\} \tag{2}$$

Thus, the problem of finding conditional probability grows into the *argmax* operation over the product of two models:

- $P(f \mid e)$ refers to *translation model (TM)* probability
- $P(e)$ to target *language model (LM)* probability.

Source-channel approach scheme can be seen in the Figure 3:

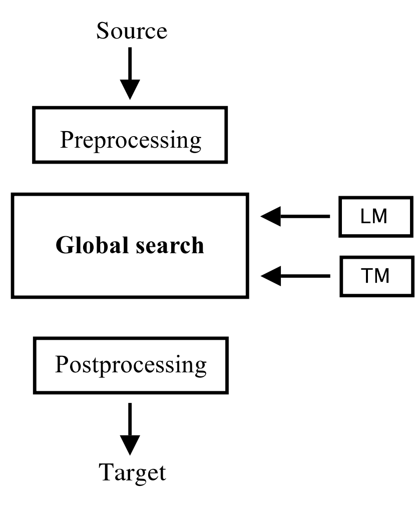

Figure 3: *Source-channel approach.*

Recently, the source-channel model has been supplemented by the maximum entropy approach [17], implementing the posterior probability $p(e|f)$ definition as a log-linear combination of the set of feature functions [18] and generalizing source-channel approach ([19] and [9]). This approach was proposed in [19] for a natural language understanding task. It allows simplifying feature models combinations under the translation hypothesis determination, as described below 3:

$$\hat{e}_1^I = \arg\max_{e_1^I} \left\{ \sum_{m=1}^{M} \lambda_m h_m(e_1^I, f_1^J) \right\} \tag{3}$$

where the feature functions h_m refer to the system models, namely bilingual TM, target LM and additional feature models; the set of λ_m refers to the weights corresponding to these models. Graphical representation of the translation scheme according to the log-linear approach can be found in the Figure 4:

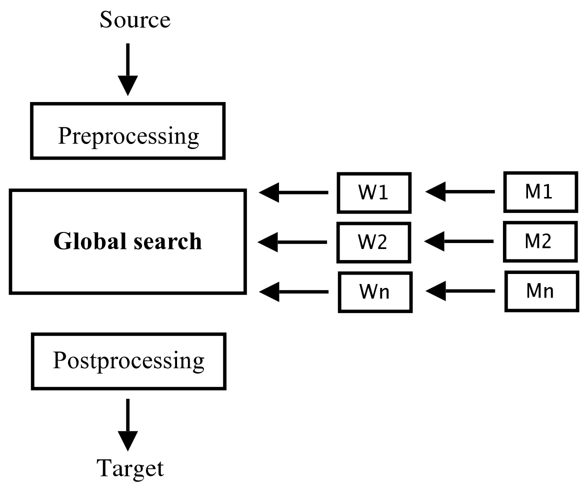

Figure 4: *Log-linear approach, W refers to weight coefficients corresponding to the models M.*

TM is a core feature of the translation system, it is based on the statistical information extracted from the parallel corpus. Training material for TM and LM can vary, in the real modern translation systems frequently supplementary monolingual tracks are using for language modeling. The weight coefficients normally are to be optimized to provide largest value of scoring function.

The features augmenting posterior probability expression include :

- *Language models*, which computes according to the following equation 4:

$$P_{LM}(t_k) \approx \prod_{n=1}^{K} P(w_n | w_{n-N+1}, ..., w_{n-1}) \tag{4}$$

  where t_k refers to the partial translation hypothesis, w_n to the n^{th} word in this partially translated sentence.

  If several language models are used trying to take advantage of different inputs, linear or log-linear combination of the language models can be composed.

- *Sentence length models*, also called *word penalties.* This feature was implemented to compensate system's striving for short output sentences. This phenomenon is due to the target language model. Technically, the penalization depends on the total number of words contained in the partial translation hypothesis, and can be found as follows 5:

$$P_{WP}(t_k) = exp(\text{number of words in } t_k) \tag{5}$$

- *Source-to-target and Target-to-source lexical models*

  This model uses word-to-word IBM model 1 probabilities [20] to estimate lexical weights of each tuple, as follows 6:

$$P_{IBM1}((e,f)_n) = \frac{1}{(I+1)^J} \prod_{j=1}^{J} \sum_{i=1}^{I} p(e_n^i | f_n^i) \tag{6}$$

  where f_n^i and e_n^i are the *j-th* and *i-th* words in the source and target parts of the tuple $(e,f)_n$, being J and I the corresponding total number of words in each side of it. Backward model is calculated for the opposite direction.

- *Other features*, like regarding information on manual lexicon entries, grammatical features, introducing linguistic knowledge as POS target LM [21] or additional statistical model [22].

Currently, most of the SMT systems use this kind of combination (see [23], [24], [25] or [26] as examples).

The MARIE decoder developed in the UPC will be used as search engine for the translation system, the details can be found in [27]. The decoder implements a beam-search algorithm with pruning capabilities.

Optimization of the weight coefficients of the scoring function is based on the *downhill simplex* optimization method [28]. Given development set and references, the log-linear combination of the weights is adjusted to maximize score function (see eq. 3) according to the highest BLEU score or log-linear combination of BLEU and NIST scores.

2.1 Alignment

The majority of the state-of-the-art systems is based on the approach proposed in the [29], it consists in using refined statistical models in the translation process. The idea is to estimate a translation model with the word alignment as a hidden variable, as shown in (4):

$$p(f_1^J \mid e_1^I) = \sum_{a_1^J} p(f_1^J, a_1^J \mid e_1^I) \tag{7}$$

where f_1^J and e_1^I refer to the source and target languages respectively and a_1^J refers to the hidden alignment, describing mapping from source position j to target position a_j. Note that alignment a_j can take on a zero value, i.e. $a_j = 0$ with the NULL word to account for the source word that is not aligned to any target word.

Following IBM approach, translation probability can be estimated using the set of following fundamental models:

- $n(\phi|e)$ or *fertility model.* It accounts for the probability that a target word e_i generates ϕ_i words in the source sentence, i.e. here a number of source words suggested which are generated by each target word.

- $t(f|e)$ or *lexicon model.* It models the probability to produce a source word f_j given a target word e_i, i.e. here strict dependencies suggested between source and target words.

- $d(\pi|\tau, \phi, e)$ or *distortion model.* Which tries to explain the phenomenon of placing a source word in position j given that the target word is placed in position i in the target sentence (also used with inverted dependencies, and known as Alignment model), i.e. here reordering of the set of the source words suggested, which better complies with target language.

Different combinations of these models are known as IBM models, the Expectation-Maximization algorithm is used to train models parameters.After the maximization phase is complete, the word alignment parameters are adjusted to maximize posterior prediction of the model.

IBM Model 1 assigns a uniform distribution to alignment probability, Model 2 introduces a zero-order dependency with the position in the source (NULL alignment). Both of them do not include fertility parameters so that the likelihood distributions are guaranteed to achieve a global maximum.

The ideas presented in the [30] and [31] encouraged creation of the Model 2 modification, that introduced first-order dependencies in alignment probabilities, the so-called homogeneous HMM alignment model. Model 3 introduces fertility and Model 4 and 5 introduce more detailed dependencies in the alignment model to allow for jumps, so that all of them must be numerically approximated and not even a local maximum can be guaranteed. More detailed information about IBM 1-5 Models can be found in [32].

Nowadays, word alignments based on IBM and HMM models, for which a systematic performance comparison can be found in [33] and in [32], are considered to be state-of-the-art. In the majority of the currently presenting translation systems, the implementation by freely-available GIZA++ package [34], is used [23, 35, 36].

Due to the model definition of alignment as a function from positions in the target sentence to positions in the source sentence, the result is strictly asymmetric, generating one-to-many word alignments. Usually, this is tackled by performing the alignment from source to target and from target to source, and symmetrizing via the union of links through the intersection or other refined methods [37].

For a detailed description of IBM models and their training from data, see [29] and [33]. In [38] and [32] a more detailed and clarifying tutorials on IBM models and translation process in general can be found.

The most wide-spread criteria of the word alignment evaluation is AER (Alignment Error Rate), proposed in [37]. Given a manual gold standard alignment with the criterion of Sure and Possible links, Recall, Precision and AER measures are defined. However, in [39] it was shown that the AER measure does not always correlate with MT accuracy, but it does with F-measure value, because of possibility to penalize precision and recall components. Recently, new measures of word alignment quality appeared [40], showing quite promising results.

Other alignment models have been presented based on word cooccurrences [41] and link probabilities, as introduced in [42], with promising results as shown in [43]. However, they generally assume a one-to-one constraint that does not account for many translation phenomena. In [44] contextual information is added to the IBM models in the framework of maximum entropy, with small but consistent improvements in Alignment Error Rate (AER). In [45] experiments combining corpus cooccurrences and linguistic knowledge are shown.

Doubtless state-of-the-art alignment systems can be significantly improved, last time several interesting works, mainly inspired by the work [46], were dedicated to further improvement of statistical word alignment and, subsequently, SMT accuracy.

2.2 Current SMT approaches

A great advance in terms of translation system accuracy was done moving from the first systems based on the noisy-channel approach model realizing word-to-word translation, described in the previous subsection, to the systems working at the phrase level, where a phrase is defined as a sequence of consecutive words in one language (with or without linguistic motivation). The main difference from the initial word-based approach is found on the translation modeling, where the word context is introduced by the use of phrases.

Currently many systems follow a phrase-based approach dealing with aligned bilingual corpora [36, 47] and implementing translation of the bilingual unit [48], [49], [26], [25], [50], [23].

2.2.1 Phrase-based

The translation accuracy has been significantly improved by switching to the phrase-based approach, which basic idea consists in extracting monolingual units from the source, translating them and composing target sentence from the phrases translations. This approach was firstly presented in [35] and named *Alignment Templates*. The translation process consists in grouping source words into phrases, alignment template application for each phrase and, finally, translation generating underlain by phrase alignment model, allowed for word classes.

A simplification of this approach is a so-called *phrase-based SMT* firstly presented in [47]. The simplification consists in the handling with words in spite of word classes and ignoring internal alignment information and assuming one-to-one phrase alignments, as shown in the 8:

$$Pr(f_1^J|e_1^I) = \alpha(e_1^I) \cdot \sum_B Pr(\tilde{f}_k \mid \tilde{e}_k) \tag{8}$$

where bilingual phrases are defined as any pair of source and target phrases that have consecutive words and are consistent with the word alignment matrix. For details on this criterion, see [35] or [47]. A smoothed relative frequency is used in [51].

The phrase translation probabilities are commonly estimated by relative frequency over all bilingual phrases in the corpus for both translation directions:

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

$$Pr(f|e) = \frac{N(f,e)}{N(f)} \tag{10}$$

where $N(f, e)$ refers to the number of times the phrase f is translated by e; $N(f)$ and $N(e)$ to the number of times the phrase in the source or target language correspondingly appears in the training corpus.

Bilingual phrases are defined as any pair of source and target phrases that have consecutive words and are consistent with the word alignment matrix. For details on this criterion, see [35] or [47]. A smoothed relative frequency is used in [51].

Other phrase-based models estimate the joint-probability $p(f, e)$ [52], this model is only tractable up to an equivalent of IBM model 3, due to severe computational limitations. Furthermore, when comparing this approach to the simple phrase generation from word alignments and a syntax-based phrase generation [53], the approach from word alignment achieves best results as shown in [36]. An alternative to compute phrase translation probabilities is to use IBM model 1 lexical probabilities of the words inside the phrase pair, as presented in [54].

Lately, a hierarchical phrase-based translation model (phrases that contain phrases) has been presented [11] and [55]. It differs from previous phrase-based models as it induces a hierarchy of phrases in form of a synchronous Context Free Grammar (CFG). The hierarchy does not have any linguistic motivation, it is induced from a parallel text without linguistic annotations. The improvement is achieved by a better modeling of reordering through the CFG. The decoder used consists of a CKY parser with beam search and a postprocessor for mapping source to target derivations.

Pure Alignment Template approach was further supplemented by introducing log-linear combination of model, more details can be found in [24].

One of the late improvements of the phrase-based approach is so-called *phrase-based back-off* technique, proposed in [56]. The core of the method is a back-off model which translates unseen word forms into a foreign language by hierarchical morphological abstractions at the word and phrase levels.

2.2.2 Ngram-based

In parallel to the phrase-based approach to the SMT, the n-gram based approach has appeared, operating with bilingual units extracted from the aligned bilingual corpus, referred to as tuples [57]. It regards translation as a stochastic process maximizing the joint probability $p(f, e)$, leading to a decomposition based on bilingual Ngrams, typically implemented by means of a Finite-State Transducer [8, 9, 57–60].

$$\hat{e}_1^I = \arg\max_{e_1^I} \{p(f_1^J, e_1^I)\} = \cdots = \tag{11}$$

$$\arg\max_{e_1^I} \{\prod_{n=1}^{N} p((f, e)_n | (f, e)_{n-x+1}, \cdots, (f, e)_{n-1})\} \tag{12}$$

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

Ngram-based (or Finite-state-based) Translation Systems model the translation directly as a composition of bilingual units, called tuples and composed of one or more words in the source side and zero, one or more words in the target side. These units are extracted in training using typically the Viterbi word-to-word alignment.

The main difference between phrase-based and ngram-based approaches lies in the distinct representation of the bilingual units defined by word alignment [61]. The tuples induce a unique segmentation of the pair of sentences [62]. 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 model can be seen here as a language model, where the language is composed by tuples. Figure 5 shows an example of tuples extraction from a bilingual sentence pair. Another clear-cut distinction between the phrase-based and tuples-based systems is distinct representation of the LM: for phrase-based system it is an integrated part, whereas in ngram-based systems LM is using as an additional feature following log-linear approach.

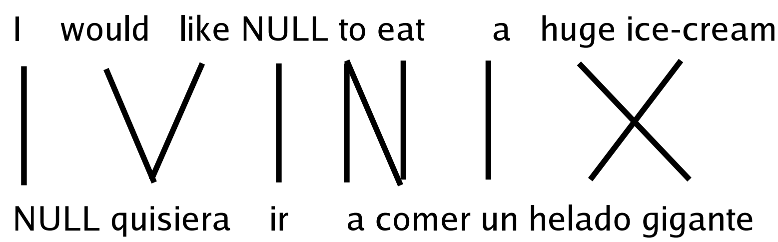

Figure 5: Tuples resulting from a bilingual sentence pair. The segmentation is unique, defined by the word-to-word alignments.

Another implementation with cascaded finite-state transducers combining both statistically learnt transducers and hand-crafted rules can be found in [63].

Tuples-based approach is basically considered as monotonous in that its model is based on the sequential order of tuples during training. Therefore, it is more appropriate for pairs of languages with relatively similar word order schemes. Currently, many research efforts are being made towards adapting the Ngram-based approach to be used for language pairs with different word order([64–66]).

2.3 Reordering

Word reordering is one of the most challenging and important problems in SMT. Without reordering capabilities, sentences can be translated correctly only in case when both languages implied in translation have a similiar word order. When translating is between language pairs with high disparity in word order, word reordering is extremely desirable for translation accuracy improvement.

Once permitted, reordering poses exact decoding problem as was shown in [67]; it can be simplified by introducing constraints such as *IBM* [68] or *Local* [65]. Word reordering is crucially important for both phrase-based and ngram-based translation systems. First word reordering attempts were completely linguistically blind, as shown in [64] and [69], later several attempts to apply lexicalized block reordering models were presented in [20] and [70], where each block is associated with an orientation with respect to its predecessor, the probability of a sequence of blocks with particular orientation is used when decoding. Recently, a numerous novel reordering methods have been proposed, mainly based on using of linguistic information somehow incorporated into the distortion model.

In [71] the idea of corpus monotonization with the aim to have similar word order on the source and target sides of the parallel corpus were presented. It can be achieved following a set of patterns, automatically learned using additional linguistic source of knowledge. For example, in [72] syntactical and morphological information in form of POS tags and automatic parser output trees is used to learn reordering patterns, in [73] restriction criteria aiming to at reducing the search graph was presented. In [71] syntactically motivated approach was suggested using 600 rules, based on the parser information for reordering of source data.

In parallel to linguistically motivated reordering techniques, purely statistical methods were developed, like Statistical Machine Reordering technique which is described in [14], the authors used statistical word classes and a word distortion model to incorporate reordering technique into the SMT system on the preprocessing step.

2.4 Evaluation

Research in the MT implies many difficulties, among which performance evaluation is one of the most challenging. So far, scientific community has not accepted unified criteria of MT evaluation. It would be ideal to have a generally accepted criteria of a SMT system automatic evaluation, like it is in the ASR field (word error rate - WER).

Nowadays, there are several automatic measures widely used. Among them *multiple word error rate* (mWER), *position independent word error rate* (PER) [74], *bilingual evaluation understudy* (BLEU) [75], NIST [76] and *String accuracy metrics* [77] are the most commonly used. Evaluation is of crucial importance as MT systems are normally trained to optimize an automatic evaluation measure.

The commonly accepted criteria that defines the quality of the evaluation metric is its level of correlation with human evaluation. Another argument to have automatic measure of MT quality is that the human scoring can be subjective and vary depending on the human or even on his particular point of view on the correct translation. As shown in [78] the above mentioned automatic measures have attained good correlation results at the system level, while the degree of correlation achieved at the sentence level, crucial for an accurate error analysis, is much lower.

In spite of the fact that BLEU measure has been, among others, broadly criticized, the majority of the SMT systems is still oriented to maximize BLEU score and is the geometric mean of the precision of ngrams of various length between a hypothesis and a set of reference translations multiplied by a factor $BP(\cdot)$ [79], as shown in the eq. 13:

$$BLEU = BP(\cdot) \cdot \exp \sum_{n=1}^{N} \frac{\log p_n}{N} \tag{13}$$

Lately, some perfected metrics appeared, namely METEOR [80], ROUGE [81] and WNM [82]. These MT measures consider additional information such as stemming or allowing for WordNet [83]. Parallel approach is to use of syntactic knowledge, as it is shown in [84], where the authors introduce series of syntax-based features based on syntactic tree matching.

Finally, there are several works devoted to design a uniform metric considering information at distinct linguistic levels and permit metric scores combination into a single measure of MT quality. In this case automatic evaluation is considered as the application of similarity metrics between a set of candidate translations and a set of reference translations. The metric realizing this approach is QARLA [85] and ORANGE [86], however, it does not permit metric combination.

2.5 Shared tasks

Due to scientific community's high interest in the SMT field, there are several shared tasks (evaluation campaigns), proposed to the research grouped all over the world to evaluate translation quality of their translation systems on an unseen test set. Evaluation is normally based on the automatic or/and human criteria, robusting evaluation results. A parallel corpus as training data and additional resources are provided. There are several evaluation campaigns, the most widely known being:

- *The NIST MT evaluation.* Annual evaluation, organized by *the National Institute of Standards and Technology.* The evaluation is a part of an ongoing series of evaluations of human language translation technology. This evaluation campaign is characterized by large amount of training material, mainly belonging to the news domain.

- *IWSLT (The International Workshop on Spoken Language Translation)* evaluation. Annual evaluation organized by *the Consortium for Speech Translation Advanced Research (C-STAR)*, focusing on the different aspects of the SMT (IWSLT 2004 evaluation - automatic evaluation metrics for speech-to-speech translation, IWSLT 2005 - automatic translation of the ASR output, IWSLT 2006 - spontaneous speech translation) and mainly supplied with small amount of in-domain training data [87].

- *ACL-WMT evaluation.* Evaluation organized in the framework of the annual ACL workshop and facing, among other challenges, the problem of automatic translating towards languages other than English [88].

3 Work Plan

The main objective of this PhD thesis is the improvement of the n-gram model for SMT, namely I am planing to focus on the three crucial aspects: target language modeling adaptation, n-gram translation model improvement and new word reordering techniques based on the introduction of syntactical knowledge.

The work plan is divided into three tasks: *target language modeling*, *translation model improvement* and *syntactically motivated word reordering*. Below, the tasks are discussed with a time scheduling. 24 months are estimated to complete all the tasks.

3.1 Target language modeling (8 months)

Language modeling is the attempt to capture and exploit regularities exhibited by natural language. The main difference between phrase-based and ngram-based approaches lies in the distinct representation of the bilingual units defined by word alignment [27], another clear-cut distinction is distinct role of the language model: for phrase-based system it is an integrated part, whereas in ngram-based systems language model uses as an additional feature. Regardless of the fact that some attention has been devoted to the language modeling for phrase-based systems, an additional investigation is needed to define LM impact on the tuple-based system.

The common point of view is that SMT systems are robust to non-grammatical input data. The majority of the up-to-date developed text-to-text translation systems address the problem of pre-edited text translation, whereas systems of recognized speech translation or manually corrected were set aside. In the framework of the work, I am considering specific language modeling techniques for FTE (Final Text Edition) system adaptation to spontaneous speech translations by means of regular and specific LMs linear interpolation. The bilingual translation model, as well as the additional target language model, correlate with the morphological nature of the bilingual corpus. However, spontaneous speech language peculiarities such as non-grammatical or bad structured input are not duly reflected in the regular translation systems. The verbatim (slightly manually corrected ASR output) transcription includes spontaneous speech phenomena (hesitations, false-starts, half-words, etc). Hence, the application of bilingual and language models calculated with semantically and grammatically corrected phrases to the spontaneous speech input data may cause translation errors.

Most of the work of the thesis is concentrated on the improvement of the translation model and the development of new techniques for word reordering. The recently published work on factored language models applied to phrase-based translation systems [10] are a promising approach for bilingual translation modeling, but first some efforts on factored model adaptation to the target LM are planned. By contrast to the phrase-based system LM factorization, this work is tackled to the ngram-based SMT system with the help of Moses [2]. Moses is an open source statistical machine translation system dealing with automatically trained translation models for words which may have a factored representation. Factored modeling can be a good solution to the sparseness data problem, since it opens a new way to combine informational sources in a intelligent manner. We plan to study different morphologically or statistically determined classes as informational sources.

[2]http://www.statmt.org/moses/

3.2 Translation model improvement (16 months)

As it was indicated in the Section 1.3, TM factorization is going to be one of the ways to improve the n-gram translation model. I plan to find the most efficient way to include and use additional lexical or morphological information in the translation process by using a factored n-gram model. Inspired by the work of [10], I consider the TM improvement on the decoding or preprocessing step. For each tuple, the cost function is calculated depending not only on a single stream of temporally preceding tuples, but also on additional parallel streams of features. It probably will help to solve the well-known problem of n-grams *no appearance*, i.e. factored models can help to avoid the situation when a particular n-gram has not been observed in the training data, but it was called in the test set, thus a factored LM can provide a more robust cost function value using the corresponding feature combination.

The second way of n-gram model improvement is hierarchical phrase model incorporation into the SMT system. In [11] the pure statistical hierarchical phrase model was presented, while in [12] statistical method was completed with linguistic information. The grammar rules are generated and, according to them, phrase pairs can be represented as combination of special terminal and non-terminal bilingual phrases, where non-terminal entries represent placeholders of inserting additional phrase pairs. Motivated by this approach, I plan to implement n-gram hierarchical TM based on the n-gram based system in order to have generalized TM driving the decoding process.

Apart of the above mentioned objectives, I plan to attack the problem of unambiguous translation that is crucial for n-gram based systems. For example, the English word *"to bite"* can be translated to Spanish as *"morder"* or as *"picar"* depending on the context of the phrase.

The part of the work dedicated to factored TM implementation is going to be held in parallel with the corresponding LM task (3 months).

3.3 Syntactically motivated word reordering (3 months)

There is an ongoing debate whether the availability of data about the syntax of language is needed for SMT. However, recent researches show that this information used in the rational way can benefit SMT system, especially for highly inflected language. On the other hand, one of the most important problems facing SMT society is the reordering problem in SMT. The research work which is expecting to carry out comprises algorithm and tools development which should combine different approaches to the SMT: statistical, syntactical, and morphological to be applied to the word reordering task. The tool which is planning to be elaborated means to reorder the entrance of the SMT system and can be considered as a preprocessing step, which should lead to the simplification of the translation task.

Our work in word reordering will be a continuation of the statistical machine reordering approach firstly presented in [14] that considers a translation of a source language phrase (S) into a reordered source language phrase (S') which leads to a monotonized word alignment and a improved SMT system performance.

The planned work has 3 fundamental objectives:

1. Development and implementation of an algorithm to integrate new reordering strategies to the entrance of the translator using the syntax transfer approach.

2. Development of a software tool for reordering the source side of the parallel corpus on the preprocessing step. This tool will be based on the mentioned developed algorithm and on the dependency distances minimization algorithm [89].

3. Adapt the developed algorithms to generate a reordering graph that will be used as input of the UPC SMT decoder.

3.4 Experimentation

The experimentation described in the previous sections will be mainly conducted over two parallel corpus. Both are used in the framework of the TC-Star [3] European Union research project, and consists of large size databases:

- The European Parliamentary sessions in Spanish-English (from April 1996 to October 2004). It contains about one million sentences (35 million running words), and allows to experiment on a monotone task with little reordering needs.

- A large variety of corpora from LDC (Linguistic Data Consortium) in Chinese-English. The most important domain includes some newswire texts, the Hong Kong Handsards and documents of the United Nations. It contains about seven million sentences (200 million running words), and allows to experiment on a task with clear needs for reordering.

- The BTEC corpus. This corpus belongs to the tourist domain and contains text from the phrase books for tourists in several languages [90]. Translation to English from four languages is considered: Mandarine Chinese, Japanese, Arabic and Italian. These translations tasks are used in the framework of the IWSLT [4] evaluation.

[3]TC-Star - Technology and Corpora for Speech to Speech Translation. http://www.tc-star.org
[4]http://www.slc.atr.jp/IWSLT2006/

4 Publicatiions

- Costa-jussà, M.R., Crego, J.M., de Gispert, A., Lambert, P., Khalilov, M., Mariño, J.B., Fonollosa, J.A.R., and Banchs, R.
  TALP Phrase-based statistical translation system for European language pairs,
  in Proceedings of the HLT/NAACL 2006 Workshop on Statistical Machine Translation (WMT'06), pp. 142-145. New York City, June 8 and 9, 2006.

- Crego, J.M., de Gispert, A., Lambert, P., Costa-jussà, M.R., Khalilov, M., Banchs, R., Mariño, J.B. and Fonollosa, J.A.R.
  N-gram-based SMT System Enhanced with Reordering Patterns,
  in Proceedings of the HLT/NAACL 2006 Workshop on Statistical Machine Translation (WMT'06), pp. 162-165. New York City, June 8 and 9, 2006.

- Mariño, J.B., Banchs, R.E., Crego, J.M., de Gispert, A., Lambert, P., Fonollosa, J.A.R., Costa-jussà, M.R. and Khalilov, M.
  UPC's Bilingual N-gram Translation System,
  in Proceedings of the TC-STAR Workshop on Speech-to-Speech Translation (TCSTAR'06), pp. 43-48. Barcelona, June 19-21, 2006, ISBN 2-9517408-3-2.

- Costa-jussà, M.R., Crego, J.M., de Gispert, A., Lambert, P., Khalilov, M., Fonollosa, J.A.R., Mariño, J.B. and Banchs, R.E.
  TALP Phrase-Based System and TALP System Combination for the IWSLT 2006,
  in Proceedings of the Int. Workshop on Spoken Language Translation (IWSLT'06), pp.123-129 appear, Kyoto (Japan), November 27 and 28, 2006.

- Crego, J.M., de Gispert, A., Lambert, P., Khalilov, M., Costa-jussà, M.R., Mariño, J.B., Banchs, R.E. and Fonollosa, J.A.R.
  The TALP Ngram-Based SMT System for IWSLT 2006,
  in Proceedings of the Int. Workshop on Spoken Language Translation (IWSLT'06), pp. 116-122, Kyoto (Japan), November 27 and 28 , 2006.

- Khalilov, M. and Fonollosa, J.A.R.
  Language modeling for verbatim translation task,
  in Proceedings of the IV Jornadas en Tecnología del Habla - the IV Biennial Workshop on Speech Technology, pp. 83-87, Zaragoza (Spain), November 8-10, 2006, ISBN 84-96214-82-6.

References

[1] W. Weaver. Translation. In W.N. Locke and A.D. Booth, editors, *Machine Translation of Languages*, pages 15–23. MIT Press, Cambridge, MA, 1955.

[2] C.E. Shannon. Prediction and entropy of printed english. *The Bell System Technical Journal*, 30:50–64, January 1951.

[3] John R. Pierce and John B. Carroll. *Language and Machines: Computers in Translation and Linguistics*. National Academy of Sciences/National Research Council, Washington, DC, USA, 1966.

[4] Philippe Langlais, Simona Gandrabur, Thomas Leplus, and Guy Lapalme. The long-term forecast for weather bulletin translation. *Machine Translation*, 19(1):83–112, March 2005.

[5] Ismael García Varea. *Traducción automática estadística: modelos de traducción basados en máxima entropía y algoritmos de búsqueda*. Tesis doctoral en informática, Departamento de Sistemas Informáticos y Computación, Universidad Politécnica de Valencia, 2003.

[6] B. Vauquois. A survey of formal grammars and algorithms for recognition and transformation in machine translation. *IFIP Congress*, pages 254–260, 1968.

[7] M. Kay, J. Gawron, and P. Norvig. Verbmobil: a translation system for face-to-face dialog. *CSLI*, 1992.

[8] K. Knight and Y. Al-Onaizan. Translation with Finite-State Devices. *Proc. 3rd Conf. Assoc. for for Machine Translation in the Americas*, pages 421–437, October 1998.

[9] E. Vidal. Finite-state speech-to-speech translation. *Proc. of 1997 IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 111–114, 1997.

[10] K. Kirchhoff and M. Yang. Improved language modeling for statistical machine translation. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 125–128, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

[11] D. Chiang. A hierarchical phrase-based model for statistical machine translation. *43rd Annual Meeting of the Association for Computational Linguistics*, pages 263–270, June 2005.

[12] Andreas Zollmann and Ashish Venugopal. Syntax augmented machine translation via chart parsing. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 138–141, New York City, June 2006. Association for Computational Linguistics.

[13] A. de Gispert. *Introducing linguistic knowledge into statistical machine translation*. Doctoral thesis, Departament of Signal Theory and Communications, Polytechnical University of Catalonia, 2007.

[14] M.R. Costa-jussà and J.A.R. Fonollosa. Statistical machine reordering. *Empirical Methods in Natural Language Processing (EMNLP)*, July 2006.

[15] D. Sole, R.and Valbelle. *The Rosetta Stone: The Story of the Decoding of Hieroglyphics.* Four Walls Eight Windows, 2002.

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

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

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

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

[20] F.J. Och, D. Gildea, S. Khudanpur, A. Sarkar, K. Yamada, A. Fraser, S. Kumar, L. Shen, D. Smith, K. Eng, V. Jain, Z. Jin, and D. Radev. A smorgasbord of features for statistical machine translation. *Proc. of the Human Language Technology Conference, HLT-NAACL'2004*, pages 161–168, May 2004.

[21] J. M. Crego, A. de Gispert, P. Lambert, M. Costa-jussà, M.R.and Khalilov, R. Banchs, J.B. Mariño, and J.A.R. Fonollosa. N-gram-based smt system enhanced with reordering patterns. *HLT-NAACL Workshop on Statistical Machine Translation (HLT-NAACL'06/Wkshp)*, June 2006.

[22] F.J. Och. An efficient method for determining bilingual word classes. *9th Conf. of the European Chapter of the Association for Computational Linguistics*, pages 71–76, June 1999.

[23] J.B. Mariño, R Banchs, J.M. Crego, A. de Gispert, P. Lambert, M. R. Costa-jussà, and J.A.R. Fonollosa. Bilingual n–gram statistical machine translation. *Proc. of the MT Summit X*, pages 275–282, September 2005.

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

[25] E. Ettelaie, K. Knight, D. Marcu, D.S. Munteanu, F.J. Och, I. Thayer, and Q. Tipu. The ISI/USC MT system. *Proc. of the Int. Workshop on Spoken Language Translation, IWSLT'04*, pages 59–60, October 2004.

[26] N. Bertoldi, R. Cattoni, M. Cettolo, and M. Federico. The ITC-irst statistical machine translation system for IWSLT-2004. *Proc. of the Int. Workshop on Spoken Language Translation, IWSLT'04*, pages 51–58, October 2004.

[27] J. M. Crego, J. Mariño, and A. Gispert. An ngram-based statistical machine translation decoder. *Proc. of the 9th European Conference on Speech Communication and Technology, Interspeech'05*, September 2005.

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

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

[30] S. Vogel, H. Ney, and C. Tillmann. Hmm-based word alignment in statistical translation. *Proc. of the Int. Conf. on Computational Linguistics, COLING'96*, pages 836–841, August 1996.

[31] I. Dagan and K. Church. Termight: identifying and translating technical terminology. In *Proceedings of the fourth conference on Applied natural language processing*, pages 34–40, San Francisco, CA, USA, 1994. Morgan Kaufmann Publishers Inc.

[32] F. Och. *Statistical Machine Translation: From Single Word Models to Alignment Templates.* Doctoral thesis, RWTH Aachen, 2002.

[33] F.J. Och and H. Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, March 2003.

[34] F.J. Och. Giza++ software. http://www-i6.informatik.rwth-aachen.de/˜och/ software/giza++.html. Technical report, RWTH Aachen University, 2003.

[35] F.J. Och, Ch. Tillmann, and H. Ney. Improved alignment models for statistical machine translation. *Proc. of the Joint Conf. of Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 20–28, June 1999.

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

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

[38] K. Knight. A statistical machine translation workbook. Technical report, Johns Hopkins University, http://www.clsp.jhu.edu/ws99/projects/mt/wkbk.rtf, April 1999.

[39] A. Fraser and D. Marcu. Measuring word alignment quality for statistical machine translation. Technical report, May 2006.

[40] Necip Fazil Ayan and Bonnie J. Dorr. Going beyond aer: an extensive analysis of word alignments and their impact on mt. In *ACL '06: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL*, pages 9–16, Morristown, NJ, USA, 2006. Association for Computational Linguistics.

[41] I.D. Melamed. Models of translational equivalence among words. *Computational Linguistics*, 26(2):221–249, 2000.

[42] C. Cherry and D. Lin. A probability model to improve word alignment. *41st Annual Meeting of the Association for Computational Linguistics*, July 2003.

[43] R. Mihalcea and T. Pedersen. An evaluation exercise for word alignment. In Rada Mihalcea and Ted Pedersen, editors, *HLT-NAACL 2003 Workshop: Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, pages 1–10, Edmonton, Alberta, Canada, May 2003. Association for Computational Linguistics.

[44] I. García Varea, F.J. Och, H. Ney, and F. Casacuberta. Improving word alignment quality using morpho-syntactic information. *Proc. of the 19th Int. Conf. on Computational Linguistics, COLING'02*, pages 1051–1057, August 2002.

[45] A. de Gispert, J. Mariño, and J. M. Crego. Phrase-based alignment combining corpus cooccurrences and linguistic knowledge. *Proc. of the Int. Workshop on Spoken Language Translation, IWSLT'04*, pages 107–114, October 2004.

[46] C. Callison-Burch, D. Talbot, and M. Osborne. Statistical machine translation with word- and sentence-aligned parallel corpora. *ACL04*, 2004.

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

[48] E. Sumita, Y. Akiba, T. Doi, A. Finch, K. Imamura, H. Okuma, M. Paul, M. Shimohata, and T. Watanabe. EBMT, SMT, Hybrid and More: ATR spoken language translation system. *Proc. of the Int. Workshop on Spoken Language Translation, IWSLT'04*, pages 13–20, October 2004.

[49] Y.S. Lee and S. Roukos. IBM spoken language translation system evaluation. *Proc. of the Int. Workshop on Spoken Language Translation, IWSLT'04*, pages 39–46, October 2004.

[50] S. Vogel, S. Hewavitharana, M. Kolss, and A. Waibel. The ISL statistical translation system for spoken language translation. *Proc. of the Int. Workshop on Spoken Language Translation, IWSLT'04*, pages 65–72, October 2004.

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

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

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

[54] S. Vogel, Y. Zhang, F. Huang, A. Tribble, A. Venogupal, B. Zhao, and A. Waibel. The cmu statistical translation system. *Proc. of the MT Summit IX*, September 2003.

[55] Taro Watanabe, Hajime Tsukada, and Hideki Isozaki. Left-to-right target generation for hierarchical phrase-based translation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 777–784, Sydney, Australia, July 2006. Association for Computational Linguistics.

[56] Yang M. and K. Kirchoff. Phrase-based backoff models for statistical machine translation. 2006.

[57] 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.

[58] F. Casacuberta, E. Vidal, and J.M. Vilar. Architectures for speech-to-speech translation using finite-state models. *Proceedings of the Workshop on Speech-to-Speech Translation: Algorithms and Systems*, pages 39–44, July 2002.

[59] David Picó, Jesús Tomás, and Francisco Casacuberta. Giati: A general methodology for finite-state translation using alignments. In *SSPR/SPR*, pages 216–223, 2004.

[60] F. Casacuberta. Finite-state transducers for speech-input translation. *ASRU01*, pages 375–380, December 2001.

[61] J. M. Crego, M. R. Costa-jussà, J. Mariño, and J.A.R. Fonollosa. Ngram-based versus phrase-based statistical machine translation. *IWSLT05*, October 2005.

[62] J. M. Crego, J. Mariño, and A. de Gispert. Finite-state-based and phrase-based statistical machine translation. *Proc. of the 8th Int. Conf. on Spoken Language Processing, ICSLP'04*, pages 37–40, October 2004.

[63] S. Vogel and H. Ney. Translation with cascaded finite state transducers. *38th Annual Meeting of the Association for Computational Linguistics*, pages 23–30, October 2000.

[64] J. M. Crego, J. Mariño, and A. Gispert. Reordered search and tuple unfolding for ngram-based smt. *Proc. of the MT Summit X*, pages 283–89, September 2005.

[65] S. Kanthak, D. Vilar, E. Matusov, R. Zens, and H. Ney. Novel reordering approaches in phrase-based statistical machine translation. *Proceedings of the ACL Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond*, pages 167–174, June 2005.

[66] E. Matusov, S. Kanthak, and H. Ney. Efficient statistical machine translation with constrained reordering. *Proceedings of EAMT 2005 (10th Annual Conference of the European Association for Machine Translation)*, pages 181–188, May 2005.

[67] K. Knight. Decoding complexity in word replacement translation models. *Computational Linguistics*, 26(2):607–615, 1999.

[68] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155, 2002.

[69] D. Wu. A polynomial-time algorithm for statistical machine translation. *34th Annual Meeting of the Association for Computational Linguistics*, pages 152–158, June 1996.

[70] C. Tillmann and T Zhang. A localized prediction model for statistical machine translation. *43rd Annual Meeting of the Association for Computational Linguistics*, pages 557–564, June 2005.

[71] Michael Collins, Philipp Koehn, and Ivona Kucerova. Clause restructuring for statistical machine translation. In *43rd Annual Meeting of the Association for Computational Linguistics*, pages 531–540, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

[72] F. Xia and M. McCord. Improving a statistical mt system with automatically learned rewrite patterns. *Proc. of the 20th Int. Conf. on Computational Linguistics, COLING'04*, pages 508–514, August 22-29 2004.

[73] J. M. Crego and J. Mariño. Reordering experiments for n-gram-based smt. *First International Workshop on Spoken Language Technology (SLT'06)*, December 2006.

[74] H. Ney, S. Nießen, F.J. Och, H. Sawaf, C. Tillmann, and S. Vogel. Algorithms for statistical translation of spoken language. *IEEE Trans. on Speech and Audio Processing*, 8(1):24–36, 2000.

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

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

[77] Palmira Marrafa. Quantitative evaluation of machine translation systems: Sentence level.

[78] M. Eck and C. Hori. Overview of the iwslt 2005 evaluation campaign. *Proc. of the Int. Workshop on Spoken Language Translation, IWSLT'05*, 2005.

[79] F.J. Och. Minimum error rate training for statistical machine translation. *41st Annual Meeting of the Association for Computational Linguistics*, July 2003.

[80] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

[81] Chin-Yew Lin and Franz Josef Och. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, pages 605–612, Barcelona, Spain, July 2004.

[82] Bogdan Babych and Tony Hartley. Extending the bleu mt evaluation method with frequency weightings. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, pages 621–628, Barcelona, Spain, July 2004.

[83] C. Fellbaum. Wordnet. an electronic lexical database. 1998.

[84] Ding Liu and Daniel Gildea. Syntactic features for evaluation of machine translation. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 25–32, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

[85] Jesús Gimenéz and Enrique Amigó. Iqmt: A framework for automatic machine translation evaluation. *Proceedings of the 5th International Conference on Language Resources and Evaluation*, 2006.

[86] Chin-Yew Lin and Franz Josef Och. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *ACL '04: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 605, Morristown, NJ, USA, 2004. Association for Computational Linguistics.

[87] Michael Paul. Overview of the IWSLT 2006 Evaluation Campaign. In *Proc. of the International Workshop on Spoken Language Translation*, pages 1–15, Kyoto, Japan, 2006.

[88] Philipp Koehn and Christof Monz. Manual and automatic evaluation of machine translation between european languages. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 102–121, New York City, June 2006. Association for Computational Linguistics.

[89] S. Zwarts and M. Dras. This phrase-based smt system is out of order: Generalised word reordering in machine translation. In *Proceedings on the Australasian Language Technology Workshop*, December 2006.

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