PhD Dissertation

Exploiting Lexical Information and Discriminative Alignment Training in Statistical Machine Translation

Patrik Lambert

Supervisor

Dr. Rafael E. Banchs

Department of Signal Theory and Communications

Tutor

Dr. Núria Castell Ariño

Artificial Intelligence PhD Program

Software Department

Universitat Politècnica de Catalunya

Barcelona, February 2008

L'amitié se nourrit de communication.
Michel de Montaigne

Abstract

The thesis work mainly focused on three aspects of statistical machine translation: the use of lexical information like basic lexical models and multi-word expressions, minimum error training strategies and word alignment models. These aspects were addressed within the n-gram-based machine translation framework. In this approach, the joint translation probability is modelled via a log-linear combination of a bilingual n-gram model and additional feature functions.

First, a thorough study of word alignment evaluation is carried out. We stress the impact on the scores of the way alignment test data are build and we give guidelines for manual alignment.

The n-gram-based machine translation system is then described. After this, we evaluate the impact on alignment quality of linguistic classifications like lemmatising, stemming or verb classification. Although these transformations have a large positive impact on word alignment, we report that this improvement has no effect on translation quality. We also examine the impact on word alignment quality and translation accuracy of grouping data-inferred multi-word expressions before alignment.

Another objective of this thesis was the improvement of minimum error training strategies. Two research lines were considered: the choice of the metric used as objective function and the improvement of the optimisation algorithm itself. In the first research line, parameters were successfully tuned with respect to the Queen score of the Qarla framework, a framework which combines different metrics with a stable and robust criterion. In the second line, the Simultaneous Perturbation Stochastic Approximation algorithm and the downhill simplex method were compared for this parameter optimisation task.

Finally, we propose a novel framework for discriminative training of alignment models with automated translation metrics as maximisation criterion. In order to evaluate

this framework, we implemented an alignment system based on discriminative models adapted to the n-gram-based translation system, and we observed a clear improvement of automated translation scores on a small corpora. We extended this framework to large corpora, tuning the alignment system parameters on a small part of the corpus, and using them to align the whole corpus. The obtained parameters were able to produce at least as good machine translation systems as with standard word alignment tools, but in a more flexible way and with less computational resource requirements.

Resumen

El trabajo de la tesis se ha enmarcado en tres aspectos de la traducción automática estadística: el uso de información léxica como modelos léxicos básicos o expresiones multi-palabras, estrategias de entrenamiento por minimización del error, y finalmente modelos de alineado a nivel de palabra. Estos aspectos han sido estudiados en el marco del sistema de traducción basado en n-gramas. Según este enfoque, la probabilidad de traducción conjunta se modela a través de una combinación log-lineal compuesta de un modelo de n-gramas bilingüe así como de funciones características adicionales.

En primer lugar, se ha estudiado en profundidad el tema de la evaluación del alineado. In particular, se destaca el impacto en los resultados de la manera con la que la referencia de alineado está construida, y se dan pautas para alinear manualmente un corpus. A continuación, se describe el sistema de traducción automática basado en n-gramas. Después de esta descripción, se evalúa el impacto sobre la calidad del alineado de clasificaciones lingüísticas como lematización, stemming o clasificación de verbos. Aunque estas transformaciones tienen un impacto importante sobre el alineado a nivel de palabras, esta mejora no se repercuta a nivel de calidad de traducción. Se examina también el impacto sobre la calidad del alineado y sobre la exactitud de la traducción de agrupar expresiones multi-palabras antes del alineado.

Otro objetivo de esta tesis era la mejora de estrategias de entrenamiento por minimización del error. Dos lineas de investigación se consideraron: la elección de la métrica utilizada como función objetivo y la mejora del propio algoritmo de optimización. En la primera línea de investigación, los parámetros fueron optimizados satisfactoriamente en función del valor de la medida Queen del marco de evaluación Qarla. Este marco combina diferentes métricas con un criterio estable y robusto. En la secunda línea de investigación, el algoritmo SPSA (aproximación estocástica por perturbación simultanea) se comparó al método del simplex.

Por último, se propone un nuevo marco para el entrenamiento discriminativo de modelos de alineado con métricas automáticas de traducción como criterio de maximización. Para poder evaluar este marco, se ha implementado un sistema de alineado basado en modelos discriminativos adaptados al sistema de traducción por n-gramas, y se pudo observar una mejora de los resultados de métricas automáticas con corpora pequeños. Se ha extendido este marco para corpora grandes, entrenando los parámetros del sistema de alineado con una parte pequeña del corpus y utilizándolos para alinear el corpus entero. Los parámetros obtenidos generaron sistemas de traducción tan buenos como los que se pueden obtener con herramientas estándares de alineado, pero de una manera mucho más flexible y con mucho menos recursos computacionales.

Agradecimientos

Sin la ayuda y la amistad que recibí durante estos años, este trabajo ni habría acabado tan pronto ni habría ido tan lejos. Por eso, quiero agradecer a todos los que hayan contribuido con su granito de arena y si me olvido a alguien, es que me hago mayor...

Quiero agradecer a todo el grupo de Procesamiento del Lenguaje Natural de LSI por la buena compañía y los buenos momentos, de ciencia o no. En particular, gracias a Horacio por su excelente curso de introducción al procesamiento del lenguaje natural, que me ha convencido de seguir adelante en este área ; gracias a Núria por haberme ofrecido la posibilidad de dedicarme a tiempo completo al doctorado, y por toda la ayuda que me aportó en esa etapa ; gràcies als Lluïsos (Padró i Màrquez) per la seva eficàcia i serietat, els seus principis i manera de fer ciència, que van ser i seran un model per a mi.

Quiero agradecer también a José Mariño por haberme acogido en el grupo de Veu de TSC, permitiéndome concentrarme en mi trabajo doctoral sin preocupación hasta el final. Le agradezco también por todo su apoyo, consejos y conversaciones muy valiosos. Mil gracias a Rafael Banchs por su ayuda, sus consejos, paciencia y optimismo a lo largo de la tesis. Gracias también al resto del grupo de traducción automática de TSC (Adrián, Adrià, Josep María, Marta, Maxim) porque es muy agradable y enriquecedor trabajar con vosotros. Gràcies al Josep Maria per les discussions tan productives sobre la cerca del BIA i altres temes.

Merci à mes parents pour leur appui et leur aide précieuse tout au long de ma thèse, malgré la difficulté des circonstances.

Finalmente quero agradecer de todo coração a Ana Beatriz por me ter ajudado e dado ânimo o tempo tudo nesse sem fim de avaliações internacionais de tradução, “deadlines”, viagens para congressos. Você pode acreditar que esta tese é também sua.

Contents

Chapter 1

Introduction

This work deals with machine translation (MT), and, more precisely, with various aspects of statistical machine translation (SMT). A first step to understand the context of this project is to see the challenges faced by the MT task.

Most of the difficulties in natural language processing arise from the ambiguity of natural language at various levels. For example, in a sentence like "time flies", processing is made difficult by morphological ambiguity (both words can be a verb or a noun). In the sentence "hand me those glasses on the table", difficulty arises from lexical ambiguity ("glasses" could be containers for drinks or a device to improve sight) and in "I saw the man on the hill with a telescope", difficulty arises from prepositional phrase attachment ambiguity ("with a telescope" is attached to "saw" or to "the man" ?).

In addition to this, MT must face the difficulties arising from the differences between languages. An overview of this topic is presented in the book of Jurafsky and Martin [Jur00]. The major source of translation problems are structural differences, lexical differences, differences in the mappings of syntax to semantics, and idioms or specific constructions.

Syntactically, languages are most different in the basic word order of verbs (V), subjects (S), and objects (O) in declarative clauses. For example, Spanish, English and Mandarin are similar in being SVO languages, meaning that the verb tends to come between subject and object in basic declarative clauses. Hindi and Japanese, by contrast, are SOV languages while Irish, Classical Arabic and Biblical Hebrew are VSO languages. Automatic translation between languages which do not share the same word-order type is confronted with an additional reordering complexity. Other structural differences include word order between constituents (such as the noun-adjective order inversion between English and Spanish) or verb syntax. Morphologically, languages are not equally affected by

case, gender, person and tense. In particular, it is difficult to translate from a language with few inflexions, like English, to a highly inflected language, which has got various word forms for the same English equivalent. For example, translating the English definite article into Spanish, where it has four forms ("el, los, la, las"), requires deciding whether the original is masculine or feminine, singular or plural. These are just examples of systematic morphological and syntactic differences between classes of languages. There are also many constructions specific to just about each language, as is the way of expressing dates, which appear in a variety of formats (YYMMDD in Japanese, MM-DD-YY in American English, DD-MM-YY in British English, etc.) and even with different calenders. In languages so close as Catalan and Spanish, time of day is expressed in a totally distinct manner.

At the lexical level, a major difference between languages is the structure of the conceptual space. As a consequence, sometimes there is no simple mapping between a word in the source language and a word or phrase in the target language, but one-to-many or many-to-many mappings. An example is depicted in the book of Hutchins and Somers [Hut92]: the English "leg" has various translations in Spanish— "pierna" (human), "pata" (animal, table), "pie" (chair) and "etapa" (of a journey). "pie" also translates to "foot" which also translates to "pata" (bird foot), which also translates to "paw". As stated earlier, it is harder to translate from a less specific into a more specific language. Another problem is the presence of lexical gaps, where no target word or phrase (other than a detailed explanation) can express the meaning of a source word. This is the case for the French word "gratiner" (meaning "to cook with a cheese coating") which has no corresponding word in English.

Another type of differences is referred to as Syntax-to-Semantics differences. They correspond to phenomena such as changes in verb and its arguments (like in "John likes the film", "la película le gusta a John") or passive constructions.

Because of all these difficulties, the problem of automatically producing a high-quality translation of an arbitrary text from one language to another is too hard. But simpler translation tasks can be achieved with current computational models. Existing MT systems are able to deal with tasks for which a rough translation is sufficient, or tasks where the MT output is post-edited by a human expert, or tasks limited to small sublanguage domains in which fully automatic high quality translation is achievable (like the Météo system, built to translate Canada's weather bulletins between English and French [Che77]).

Research in MT actually began in the late 1940s, introducing a phase of excitement and promise. Since predicted quality was not achieved (with the computational resources

of the time, MT was slower, much less accurate and twice as expensive as human translation), disillusion times set in the mid 1960s, leading to dramatic cuts in funding and research. Interest in the field was revived in the late 1970s, with the improvement of computational resources, the perspective of using Artificial Intelligence techniques, and when existing systems (such as SYSTRAN [Tom76]) proved to have commercial possibilities. The development of personal computers helped to extend MT usage and sales increased in the 1990s. The 1990s saw also the appearance of statistical systems (like CANDIDE at IBM [Bro90]), enabled by the availability of large parallel corpora.

The traditional use of MT is the translation of technical documents for multinational companies or institutions. MT produces a first draft which is post-edited by human translators. In large scale and/or rapid translation of highly repetitive material, the costs of MT plus essential human preparation and revision can be significantly less than the cost of full manual translation. Still, post-editing can be expensive and a successful way of reducing editing cost is the pre-editing of source texts (typically with a so-called controlled, "regularized" language) in order to minimise incorrect MT output. Translation agencies and smaller companies prefer to work with translation tools, which assist manual translation, than with MT. These tools include creation, use and personalisation of translation memories, multilingual word processing, terminology recognition, retrieval and management.

Recently, the Internet has produced a rapidly growing demand for real-time on-line translation. Most users need a fast look-up of foreign-language information, and the translation quality is not essential. The ruled-based approaches, developed for well-written scientific and technical documents, are not very adequate to this task, which includes translation of spontaneous and less structured text like e-mail and chat messages. This has increased the demand for improving data-driven systems, which are more robust since they are not governed by conventional syntax and semantics and thus always provide some output. At the same time, statistical methods have been applied to the implementation of spoken language translation systems, another task in which the input is often not grammatical (in addition to ungrammaticallity due to the spontaneousness of spoken language, errors are introduced by automatic speech recognisers). Nowadays, SMT performance is comparable to rule-based MT.

1.1 Machine Translation Approaches

In this section, a brief overview of main MT approaches is presented. A more detailed description is given by Hutchins and Somers [Hut92].

1.1.1 Interlingua

The interlingua idea is based on the mapping of the input into a language independent representation of its meaning. This meaning is then expressed in the target language. This idea is symbolised in the "MT pyramid" diagram (figure 1.1). The interlingua represents all sentences that have equivalent meanings in the same way, regardless of the language they happen to be in. This very nice theory is nevertheless hard to implement. The first problem is to devise an adequate interlingua. If the representation is too simple, too many possible translations are admitted. If the representation is complex enough, analysis and generation will be extremely difficult since the representation is inevitably distant from the characteristic features of source and target texts. An exhaustive analysis of the semantics of the domain must be performed, which requires a vast amount of knowledge resources (ontologies, grammatical rules and especially world knowledge). For this reason some interlingua-based systems are also called knowledge-based systems. An additional difficulty is that the system must be fully disambiguated at all times, while in more direct approaches this is not always necessary. For all these reasons, current practical interlingua-based systems are restricted to small domains for which assembling the necessary knowledge is achievable. Ultra [Far91] and Mikrokosmos [Bea95] are examples of interlingua-based systems.

Finally, because the interlingua approach is independent from the language pair, it is particularly attractive in multilingual systems. For each language added to the system, only two modules must be added (one for analysis and one for generation) and only monolingual knowledge is required to develop each module.

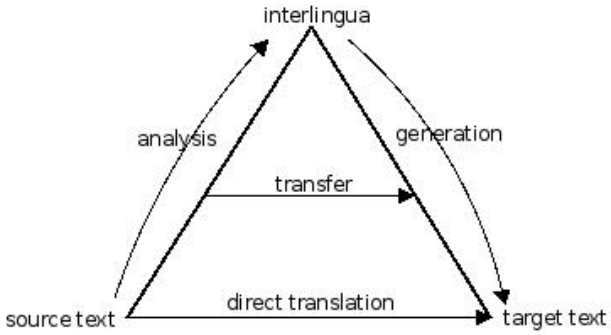


Figure 1.1: Machine translation "pyramid" diagram.

1.1.2 Transfer-based

The transfer strategy is to transform the structure of the input to make it conform to the rules of the target language. This requires knowledge of the differences between the

two languages. The input must be analysed to obtain a representation of its structure (for example a parse tree). Then the input representation is transfered to a target language representation, using transfer rules. Finally the output is generated from the target language representation. As depicted in figure 1.1, the transfer takes place at some level of analysis, ranging from limited syntax to some form of semantic representation. The transfer rules are a large set of syntactic and lexical transformation operations that map from one language structure to another. For example, the rule which inverts in a noun phrase (NP) adjective and noun order in translation from English into French may be the following:

$$[NP \rightarrow Adjective_1\ Noun_2] \Rightarrow [NP \rightarrow Noun_2\ Adjective_1] \tag{1.1}$$

By comparing to the interlingua approach, the complexity of the analysis and generation modules in a transfer system is more reduced, because the intermediate representations involved are still language-dependent abstractions [Hut92]. The level of abstraction of the transfer representation can also be adapted to the particular language pairs. For these reasons, transfer systems normally overcome in multilingual tasks the drawback of having to develop a distinct module for each translation direction.

An example of a practical transfer-based system is METAL [Ben85], introduced in the commercial market by Siemens in 1989.

1.1.3 Direct Translation

1.1.3.1 Rule-based

The strategy adopted by rule-based translation is that the system should do as little work as possible. Typically these are ad-hoc systems built with only one language pair in mind, and that perform simple (but reliable) operations adapted to the specificities of that language pair. In contrast to the approaches seen in §1.1.1 and §1.1.2, all the processing involved to analyse a specific issue (like prepositions) is handled in one stage, including analysis, transfer, and generation aspects. This usually makes each individual problem more tractable. Although their performance is also conditioned to the success of the analysis, they are somewhat more robust than the transfer or interlingua based systems, because they do not deal with complex structures and representations. On the other hand, their quality is more dependent on the bilingual dictionary (and the larger the dictionary the better). The successive stages of a typical Japanese to English direct translation system could be the following [Jur00]: morphological analysis, lexical transfer of content words, work relating to prepositions, Subject-Verb-Object rearrangements, miscellany, and morphological generation.

A problem of rule-based direct systems is that they hit a ceiling at which they become so complex that the addition of any rule (or word) causes as much degradation as enhancement. To reduce the complexity of the rule system, some aspects of the transfer approach can be introduced. The original versions of the Systran system [Tom76], in operation since the 1970s, used a direct approach. Many modifications have transformed it in a rather transfer-based system.

1.1.3.2 Example-based

The basic principle of example-based systems is to find in the input sentence fragments for which the translation is known [Way05]. The closest match of each fragment is used to construct the output translation. Characteristics of a specific example-based method are the matching criteria, the length of the fragments, the generalisation of the stored translation examples, the level of linguistic knowledge used, etc.

1.1.3.3 Statistical

Historically, SMT originates from the application of methods which were very successful in speech recognition. In contrast to the interlingua, transfer and rule-based direct methods, the focus is on the result and not on the process. In its basic form, the result of translation is modelled as the maximum of some function which represents the importance of faithfulness and fluency. This approach of translation was first described by Brown *et al.*[Bro90; Bro93], in terms of the noisy channel model. In this model, the input sentence $\mathbf{s}$ to be translated is considered to be a distorted version of some target language sentence $\mathbf{t}$ (in this view the distortion due to noise has produced a language change). The task of the translation decoder is, given the distorted sentence $\mathbf{s}$, to find the sentence $\hat{\mathbf{t}}$ which has the best probability to have been converted into $\mathbf{s}$:

$$\hat{\mathbf{t}} = \arg\max_{\mathbf{t}} Pr(\mathbf{t} \mid \mathbf{s}). \tag{1.2}$$

By using Bayes' theorem, we write:

$$Pr(\mathbf{t} \mid \mathbf{s}) = \frac{Pr(\mathbf{t})Pr(\mathbf{s} \mid \mathbf{t})}{Pr(\mathbf{s})} \tag{1.3}$$

Since the denominator is independent of $\mathbf{t}$, the best translation is given by the following equation:

$$\hat{\mathbf{t}} = \arg\max_{\mathbf{t}} Pr(\mathbf{t})Pr(\mathbf{s} \mid \mathbf{t}) \tag{1.4}$$

To implement equation 1.4, three tasks must be performed: quantify fluency ($Pr(\mathbf{t})$), quantify faithfulness ($Pr(\mathbf{s} \mid \mathbf{t})$) and find an algorithm which maximises the product of these two functions.

As will be seen in §2.2.1, the target language model $Pr(\mathbf{t})$ can be implemented with word n-grams. The translation model $Pr(\mathbf{s} \mid \mathbf{t})$ is typically based on fragments of sentences, called translation units, for which the translation probabilities are estimated after introducing an alignment between them or between the sentence words (see §2.4). With respect to this aspect, the difference with example-based methods is that the ranking between fragments is done with probabilities instead of matching measures. The main SMT methods will be described in more details in §2.1.

1.1.4 Discussion

Up to this point, we have presented various MT approaches, and have compared them with respect to the eventual intermediate representation they use (see figure 1.1). First it should be pointed out that practical systems typically combine aspects from several of these approaches. And other types of models are investigated in MT research.

In this subsection we also want to compare the two general approaches: based on rules and based on statistical information.

With data-driven methods, translation information is acquired automatically from a parallel corpus (text along with its translation). Their main advantage is that they eliminate the need for the considerable human effort required for rule-based approaches. Once the system is implemented, the only effort to adapt it to a new language pair is to collect an adequate parallel corpus.

By contrast, an advantage of rule-based systems is that it is relatively easy to design a rule to solve a very specific problem occurring in translation (say, a particular case of some rule). In principle, this is not feasible with a statistical model. However, as mentioned previously, practical rule-based systems are so complex and contain so many rules that adding a rule to solve an error in a sentence often causes unexpected errors in other sentences. While the rule-based approach offers more flexibility to design more specific rules, it makes it complicated to disambiguate between various possible translations because there is no natural way to assign scores to the various rules. On the contrary, the essence of statistical models is to assign probabilities to each event. Finally, as said before, two main advantages of statistical systems are their robustness and their adaptability to any language pair (provided a bilingual corpus is available).

1.2 Contributions of this Thesis

The main contributions of this thesis include:

- A thorough study of word alignment evaluation and of the impact of the manual reference on the alignment error rate (AER). We highlight the ambiguity and difficulty of the word alignment evaluation task. We also show that the ratio between Sure and Possible links in the reference has got a critical influence on the AER scores: low S/P link ratios favour high precision computed alignments, whereas high S/P link ratios favour high recall computed alignments. This observation changes the interpretation of many recent results on word alignment techniques. In many recent papers, the proposed alignment system is evaluated against the same manual reference, the alignment error rate against this reference becoming in some way a measure of the state of the art in the field. This is a freely available reference which contains a low S/P links ratio and thus favours high precision alignments. Thus the claim of improvement of the word alignment in these papers is to be understood as merely a improvement of the word alignment precision. The quality of these systems should be evaluated against various types of references. Other contributions of this thesis are guidelines for manual alignment and the release of freely available word alignment test data.

- The exploration of the idea of exploiting multi-word expressions in word alignment and statistical machine translation. Data-inferred multi-word expressions were grouped so as to be considered as a unique token during alignment training. Although the approach obtained encouraging results in a small and clean corpus, it did not yield any improvement with "real-life" data. Nevertheless, according to a detail error analysis, grouping multi-word expressions before training helped for their correct translation in test when the considered multi-words were fixed expressions, *i.e.* expressions that cannot be translated word to word. On the contrary, grouping multi-word sequences which could in fact be translated word to word or which were erroneously extracted didn't help, introducing unnecessary rigidity and data sparseness in the models. Thus, this study suggests that clean multi-word expression lists would be useful to improve machine translation quality.

- A novel approach for SMT parameter adjustment, based on the IQmt toolkit by Jesús Gímenez *et al.*, which implements the Qarla framework (of E. Amigó *et al.*). Instead of relying on a single evaluation metric, or in an ad-hoc linear combination of metrics, this method works over metric combinations with maximum descriptive power, aiming to maximise the Human Likeness of the automatic translations.

- A study of the experimental error introduced by the minimum error training strategy. We propose to perform the adjustment of the system coefficients with the Simultaneous Perturbation Stochastic Approximation (SPSA) algorithm and present experiments in which the variability of the results is reduced with respect to using the downhill simplex method.

- A novel framework for discriminative training of alignment models with automated translation metrics as maximisation criterion.

  A key issue of real life MT systems (like commercial systems) is the ability for the user to customise the tool, for example through personalised dictionaries or translation memories. For an SMT system, this would mean having the possibility to constantly augment and improve the training corpus. For example, one can imagine that the user may correct the system output and add the corrected sentences to the corpus. The main obstacle to achieving this, though, is the required previous word alignment stage. A vast majority of SMT systems use generative statistical alignment models which are very costly and need to consider the whole corpus. Thus, to align just one more sentence, the whole training corpus has to be aligned again. Recently, several discriminative alignment system were proposed, which consist in a log-linear combination of various models. Once the scaling factors of the models have been trained, these systems have the ability to align separately an update of the corpus. However, scaling factor tuning is performed optimising an alignment quality metric measured on a set of manually annotated data. Annotating data can be a limitation, but the main problem is the lack of correlation between alignment quality metrics and machine translation quality.

  In this thesis, we propose and evaluate a framework which solves this problem by training directly the alignment system scaling factors in function of machine translation metrics. In addition, in this approach no link labels at the word level are needed. We report results which show that on small corpora, the translation systems trained with our approach perform better than the ones trained from standard generative models. We successfully applied this framework to large corpora, tuning the alignment parameters on a small subset of the training corpus and using them to align the whole corpus. The alignment systems built from the resulting alignments were as good as the system built with alignments produced by standard generative alignment models. In this work, we also improve basic characteristics of discriminative alignment systems, like the word association model and the search strategy.

  To conclude, we think that thanks to the proposed framework, it is easier to use an SMT system for real life applications.

Chapter 2

State of the art

This chapter gives a survey of the most salient statistical machine translation and word alignment approaches being followed since the first SMT models, and which are most relevant to our research work.

Since the original SMT systems proposed by Brown *et al.*, which use single words as translation units, within the noisy channel approach, two main aspects of the systems have evolved. First, word-based translation units have been replaced by units based of word sequences (phrases). 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. This evolution has also occurred in the systems which address the translation problem from other perspectives: the finite-state perspective or the sentence structure (syntax) perspective.

In Section §2.1, various translation model approaches are reviewed: the word-based IBM models, the introduction of phrase-based models, the finite-state perspective and finally, we briefly mention some work in syntax-based SMT.

Section §2.2 deals with the other tasks of the SMT process: language modelling and search. With the extension of the noisy channel model to a maximum entropy approach (§2.2.3), tasks such as minimum error training or re-scoring have appeared. Finally, we discuss in Section 2.3 some evaluation issues.

Word alignment, which has become a key process in the extraction of translation units, is reviewed in Section §2.4.

2.1 Statistical Machine Translation Approaches

2.1.1 IBM Translation Models

In the early nineties, Brown *et al.* [Bro90; Bro93] proposed an approach of machine translation based on the noisy channel (§1.1.3.3). In this work, a "hidden" alignment $\mathbf{a}$ was introduced in the translation model to describe the connections between words:

$$Pr(\mathbf{s} \mid \mathbf{t}) = \sum_{\mathbf{a}} Pr(\mathbf{s}, \mathbf{a} \mid \mathbf{t}) \tag{2.1}$$

Without loss of generality, $Pr(\mathbf{s}, \mathbf{a} \mid \mathbf{t})$ can be written in the following way[1]:

$$Pr(\mathbf{s}, \mathbf{a} \mid \mathbf{t}) = Pr(J \mid \mathbf{t}) \prod_{j=1}^{J} Pr(a_j \mid a_1^{j-1}, s_1^{j-1}, J, \mathbf{t}) Pr(s_j \mid a_1^{j}, s_1^{j-1}, J, \mathbf{t}) \tag{2.2}$$

The interpretation of this equation is the following. To generate a source sentence and an associated alignment from the target sentence, we can first choose the length J of the source sentence given what we know about the target sentence. Then we can choose where to link the first source position given the target sentence and the length of the source sentence. Then we can choose the identity of the first source word given the target sentence, the length of the source sentence and the target word linked to the first source position, and so on. In the IBM models formulation, the alignment is a mapping from source positions j to target language positions $i = a_j$. Alignment $a_j = 0$ with an "empty" (Null) word e_0 is assigned to the source words that are not aligned to any target words.

The set of model parameters, or probabilities, is obtained by maximising the likelihood of parallel data with the EM algorithm [Dem77]. In order to handle this huge amount of parameters, the EM algorithm is used with increasingly complex models.

In Model 1, it is assumed that $Pr(J \mid \mathbf{t})$ is a constant ϵ which is independent of J and $\mathbf{t}$; that $Pr(a_j \mid a_1^{j-1}, s_1^{j-1}, J, \mathbf{t})$ depends only on the length of the target sentence I, and therefore must be $(I+1)^{-1}$; and that $Pr(s_j \mid a_1^{j}, s_1^{j-1}, J, \mathbf{t})$ depends only on s_j and t_{a_j} [2]:

$$p_{IBM1}(\mathbf{s} \mid \mathbf{t}) = \frac{\epsilon}{(I+1)^J} \prod_{j=1}^{J} \sum_{i=0}^{I} p(s_j | t_i) \tag{2.3}$$

Thus, Model 1 is a simple word translation model, which only makes use of co-occurrence of word pairs. Model 2 adds local dependencies by introducing position parameters to the

[1] x_y^z denotes the sequence $x_y \dots x_z$. s_j stands for the source word at position j and a_j is the corresponding alignment parameter.

[2] In contrast to general probability distributions with (nearly) no specific assumptions, for which we use the symbol Pr(.), model-based probability distributions will be denoted with the generic symbol p(.).

translation model. In model 3, *fertility* parameters are introduced. Fertility parameters represent the probability of words to be aligned to a certain number of corresponding words. Model 4 includes additional dependencies on the previous alignment and on the word classes of surrounding words in order to handle word groups, which tend to stick together. Model 3 and 4 have deficiency problems, which means that parts of the probability distribution are reserved for impossible events. Model 5 gets rid of the deficiency problems. Vogel *et al.* [Vog96] proposed a improvement of model 2, called HMM model, in which the position parameters depend on the previous word. The HMM model predicts the distance between subsequent source language positions, whereas Model 4 predicts the distance between subsequent target language positions. For Model 1 and 2, a global maximum of the likelihood distributions can be found. However, because of the fertility parameters, only a local maximum can be achieved for Models 3, 4 and 5. Knight [Kni99] wrote a very enlightening informal tutorial on IBM models.

2.1.2 Phrase-based Translation Models

The main problem with word-based models is that they fail to capture local context. Thus novel approaches needed to express their models according to longer units, typically sequences of consecutive words (or *phrases*).

In Wang and Waibel [Wan98], the IBM models are modified to add phrase based parameters to the set of word based parameters. However, if phrases are taken into account at the alignment and reordering level, source words are individually translated into target words.

The first approach using longer translation units was presented by Och *et al.* [Och99b] and named Alignment Templates. In this approach the translation unit is a triple composed of: a source sequence of word classes, a target sequence of word classes, and a set of internal alignment links between word classes inside the borders of the template. Word classes from source and target language can be automatically estimated from monolingual or bilingual data as in [Och99a]. The generative process underlying the Alignment Template approach is drawn in Figure 2.1. As it can be seen, source words (each of them belonging to a word class) are grouped into phrases $\tilde{f}_k$, and for each phrase an alignment template is applied, originating a set of target phrases. Then, these phrases are ordered according to the phrase alignment model, and finally target words are produced. More details on this approach can be found in [Och04b], where significant improvements over word-based approaches are reported.

Marcu and Wong [Mar02] proposed a joint-probability phrase-based model in which

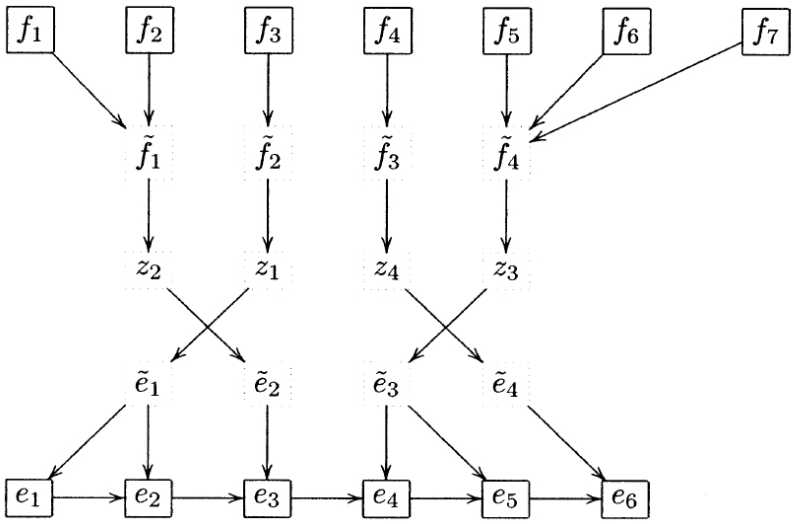


Figure 2.1: Illustration of the translation process of the Alignment Template approach

both word and phrase translation and alignment probabilities are learned from a set of parallel sentences. However, this model is only tractable up to an equivalent of IBM model 3, due to severe computational limitations.

2.1.2.1 Phrase-based Statistical Machine Translation

Zens *et al.* [Zen02] introduced a simpler phrase-based translation model than the alignment templates one, in which internal alignment information is not included in the translation unit and in which no classes are used. Thus the translation unit is simply a pair of (contiguous) word sequences. To use these bilingual phrases, the segmentation B of the sentence pair $(\mathbf{s}, \mathbf{t})$ into K phrases is introduced. A one-to-one phrase alignment is assumed, giving the following model:

$$p(\mathbf{s} \mid \mathbf{t}) = \alpha(\mathbf{t}) \sum_{B} \prod_{k=1}^{K} p(\tilde{s}_k \mid \tilde{t}_k) \tag{2.4}$$

In Equation 2.4, it is also assumed that all segmentations have the same probability $\alpha(\mathbf{t})$.

The set of bilingual phrases (BP) is extracted from word alignment and represents the set of contiguous word sequence pairs consistent with the word alignment matrix [Och99b; Zen02]. This means that all words within the target language are only aligned to the words of the source language and vice versa. At least one word of the target language phrase has to be aligned with at least one word of the source language phrase. Finally, the algorithm takes into account possibly unaligned words at the boundaries of the target or source language phrases.

The phrase translation probabilities are calculated by relative frequencies:

$$p(\tilde{s}_k \mid \tilde{t}_k) = \frac{N(\tilde{s}_k, \tilde{t}_k)}{N(\tilde{t}_k)} \tag{2.5}$$

Some alternative methods to extract phrases and learn phrase translation tables have been proposed [Til03b; Ven03] and compared in Koehn *et al.*[Koe03]. According to Koehn *et al.*, best results are obtained with the method described above.

Efficient phrase extraction and estimation can be performed with Thot, an open-source tool developed by Ortiz [Ort05].

With the introduction of phrase-based SMT, state-of-art SMT quality has experimented an important improvement and many current systems follow this approach. The availability of open-source decoders like Pharaoh or Moses has contributed in its popularity (see §2.2.2). Another issue, of particular relevance to this thesis, is that with phrase-based SMT, alignment parameters disappeared from the model. The best alignment (called *Viterbi* alignment) is used to extract the translation model probabilities. Thus, word alignment became a stand-alone training stage which can be performed independently. Word alignment techniques will be surveyed in §2.4.

2.1.3 Tuple-based Translation Model

An alternative to the noisy channel approach formalised by Equation 1.2 is to consider a joint model, as in the following equation:

$$\hat{\mathbf{t}} = \arg\max_{\mathbf{t}} Pr(\mathbf{s}, \mathbf{t}), \tag{2.6}$$

which can be decomposed[3] in

$$\arg\max_{\mathbf{t}} \{ \prod_{k=1}^{K} Pr((\tilde{s}, \tilde{t})_k | (\tilde{s}, \tilde{t})_1, ..., (\tilde{s}, \tilde{t})_{k-1}) \}. \tag{2.7}$$

where $(\tilde{s}, \tilde{t})_k$ is the k-*th* bilingual unit, or *tuple*, of a given tuple sequence which generates monotonically both the training source and target sentences. Under this approach, the translation of a given source unit $(\tilde{s}, \tilde{t})_k$ is conditioned by a previous bilingual context $(\tilde{s}, \tilde{t})_1, ..., (\tilde{s}, \tilde{t})_{k-1}$, which in practice must be limited in length.

This approach was first implemented by means of finite-state transducers (FST) for which transition probabilities are learnt from training data [Vid97; Cas01; Cas04]. This

[3] see more details in §2.2.1

implementation allows an elegant integration of speech recognition and machine translation into a global search for speech-to-speech translation. The FST-based approach is monotonous in that its model is based on the sequential order of tuples during training. Therefore, in principle it is more appropriate for pairs of languages with relatively similar word order schemes. However, Bangalore [Ban01] introduced reordering by adding a reordering FST following the lexical translation FST.

De Gispert and Mariño [Gis02] also adopted this approach, but relying on a different segmentation of the sentence pair as for example in the work of Casacuberta. The translation system used in this thesis is a re-implementation of their system with large-vocabulary language modelling tools which includes the most advanced smoothing strategies [Sto02] (see Chapter 4).

An alternative implementation of Equation 2.6, developed by Tillmann and Xia [Til03a], relies on bilingual-unit unigram probabilities.

2.1.4 Syntax-based SMT

Nowadays phrase-based SMT is by far the most popular SMT approach. However this approach is limited by the fact that it makes no or little direct use of syntactic information, although intuitively, it seems evident that syntactic information is very valuable to model the target language structure and systematic differences in word order. In particular, the knowledge of the target language syntax allows a human translator to build a translation in total independence from the source language structure. A number of researchers [Als96; Wu97; Yam01; Gil03; Mel04; Gra04; Gal04] have proposed models where the translation process involves syntactic representations of the source and/or target languages.

However, until recently [Mar06a], syntax-based SMT did not achieve state-of-the-art results and syntactic information was considered by many researchers as not useful for SMT. In the last years, however, syntax-based SMT started to be widely considered as a promising field.

2.2 Statistical Machine Translation Process

In the noisy channel approach (see §1.1.3.3), translation output is obtained by performing three tasks: 1) estimating a translation model, 2) estimating a language model and 3) maximising the product of both (which is called search or decoding). Various translation models were reviewed in Section §2.1. This section is first dedicated to the two other tasks of the translation process within the noisy channel approach. Then an extension

of this approach in the maximum entropy framework is discussed, as well as the related coefficient optimisation and re-scoring tasks.

2.2.1 Language Model

A language model should give an idea of the correctness of a given sentence and, in our case, of the fluency of the translation output. Thus, the model must assign a probability $Pr(\mathbf{t})$ to each translation hypothesis $\mathbf{t}$ of the target language. A simple option for this model consists in dividing the sentences in fragments, small enough to be frequent (and thus appear in the corpus), but large enough to contain some language information, and calculate the probability for each fragment. A sentence which contains many good fragments placed in good order should have a high probability.

For computers the easiest way to cut a sentence in fragments is to consider sub-strings. A sub-string of n words is called n-gram. A general way to decompose $Pr(\mathbf{t})$ is the following:

$$\begin{aligned} Pr(t_1 t_2 \dots t_I) &= Pr(t_I \,|\, t_1 t_2 \dots t_{I-1}) Pr(t_1 t_2 \dots t_{I-1}) \\ &= \prod_{i=1}^{I} Pr(t_i \,|\, t_1 t_2 \dots t_{i-1}) \end{aligned} \tag{2.8}$$

where $t_1^I \equiv t_1 t_2 \dots t_I$ refers to the target sentence and t_i to the i^{th} word in it. The target language model generally consists of a n-gram model, in which the probability of a translation hypothesis is approximated by the product of word n-gram probabilities. For $n = 3$:

$$p_{LM}(t_1^I) \approx \prod_{i=1}^{I} p(t_i | t_{i-2} t_{i-1}) \tag{2.9}$$

Parameters of the model are estimated by maximum likelihood, from relative frequencies in a bilingual corpus. Keeping the $n = 3$ example:

$$p(t_i | t_{i-2} t_{i-1}) = \frac{N(t_{i-2} t_{i-1} t_i)}{N(t_{i-2} t_{i-1})} \tag{2.10}$$

where $N(t_{i-2} t_{i-1} t_i)$ is the number of times that the word sequence $t_{i-2} t_{i-1} t_i$ appears in the corpus, and $p(t_i | t_{i-2} t_{i-1})$ is the probability that the word sequence $t_{i-2} t_{i-1}$ appears behind word t_i.

With this method, n-grams not seen in the corpus will have a probability of zero and will void the whole sentence probability. To avoid this, probability distribution is smoothed such that a very small but positive value is assigned to zero probabilities. A

refined way of solving this problem is with a linear interpolation. For a trigram model, the smoothed probability would be:

$$p(t_i|t_{i-2}t_{i-1}) = \lambda_0 + \lambda_1 p(t_i) + \lambda_2 p(t_i \,|\, t_{i-1}) + \lambda_3 p(t_i \,|\, t_{i-2}t_{i-1}) \tag{2.11}$$

The λ_n parameters can be learnt to find the best weight combination, however the idea is to have $\lambda_3 \gg \lambda_2 \gg \lambda_1 \gg \lambda_0$. For a description of more refined smoothing methods and estimation of n-gram models see for instance the book by Manning and Schütze [Man99]. A successful smoothing method for the kind of language models considered in this thesis is the one proposed by Kneser and Ney [Kne95].

2.2.2 Search

Decoding is also a crucial part in the SMT process. Without a reliable and efficient decoding algorithm, the system may miss the best translation of an input sentence even if it is perfectly predicted by the models.

In word-based SMT systems, search was performed following different approaches including optimal A* search [Och01], integer programming [Ger01], greedy search algorithms [Wan98]. García Varea [GV03] wrote a detailed study on word-based search algorithms. An important issue of these decoders is the computational complexity introduced by reordering (changes in word order) when single-word are considered instead of longer units.

In phrase-based decoders, short-distance reorderings between source and target sentences are already captured within the translation units, which alleviates the reordering problem [Til00; Och04b; Koe04]. Pharaoh [Koe04], an efficient and freely available beam search phrase-based decoder was very successful and contributed in making SMT more accessible and more popular. Recently, Pharaoh has been replaced/upgraded by Moses [Koe07], which is also a phrase-based decoder implementing a beam search, allowing to input a word lattice with confusion networks and using a factored representation of the raw words (surface forms, lemma, part-of-speech, morphology, word classes, *etc.*). Nowadays, many SMT systems employ a phrase-based beam search decoder because of the good performance results achieved (in terms of accuracy and efficiency).

Crego [Cre05b] implemented an n-gram-based decoder called MARIE, which is also freely available. It is the decoder we used in this thesis work.

2.2.3 Direct Maximum Entropy Translation Approach

The best estimated translation as given by equation 1.4 is optimal if the true probability distributions $Pr(\mathbf{t})$ and $Pr(\mathbf{s} \mid \mathbf{t})$ are used. However, the available models and training methods only provide poor approximations of the true probability distributions. In these conditions, a different combination of language model and translation model might yield better results. Another limitation of the source-channel approach is the difficulty to extend the baseline model, composed of a language model and a translation model, to other features. Applying the framework introduced by Papineni [Pap98] for a natural language understanding task, Och and Ney [Och02] proposed to model directly the posterior probability $Pr(\mathbf{t} \mid \mathbf{s})$ in the maximum entropy framework [Ber96]. In this framework, the corresponding translation hypothesis $\mathbf{t}$, for a given source sentence $\mathbf{s}$, is defined by the target sentence that maximises a log-linear combination of feature functions $h_i(\mathbf{s}, \mathbf{t})$, as described in the following equation:

$$\hat{\mathbf{t}} = \arg\max_{\mathbf{t}} \sum_{m} \lambda_m h_m(\mathbf{s}, \mathbf{t}) \tag{2.12}$$

where the λ_m's constitute the weighting coefficients of the log-linear combination and the feature function $h_m(\mathbf{s}, \mathbf{t})$ corresponds to a logarithmic scaling of the m^{th}-model probabilities.

Note that the source-channel approach (equation 1.4) is a special case of this framework. Namely, it is the case for which:

$$\begin{aligned} h_1(\mathbf{s}, \mathbf{t}) &= \log p(\mathbf{t}) \\ h_2(\mathbf{s}, \mathbf{t}) &= \log p(\mathbf{s} \mid \mathbf{t}) \\ \lambda_1 &= \lambda_2 = 1 \end{aligned} \tag{2.13}$$

Och and Ney [Och02] showed that, using these feature functions (i.e. the same models as in the source-channel approach), but optimising the feature weights λ_1 and λ_2, translation quality is significantly improved.

2.2.4 Minimum Error Training

As seen in §2.2.3, translation quality can be improved by adjusting the weight of each feature function in the log-linear combination. This can be effectively performed by minimising translation error over a development corpus for which manually translated references are available [Och03a].

This minimisation problem in multiple dimensions is difficult because of three main characteristics of the objective function. Firstly, it has no analytic representation, so the

gradient cannot be calculated. Secondly, it has many local minima. Finally, its evaluation has a significant computational cost (depending on the scheme, it implies translating the development corpus or re-ranking an n-best list for this corpus, and calculating some translation error measure). Gradient may be approximated, but this is costly since it requires typically as many function evaluations as the number of scaling factors. Thus, algorithms based on derivatives are discarded. Algorithms which require many objective functions evaluations, such as simulated annealing or genetic algorithms, are also discarded. In these conditions, the algorithms that can effectively be used cannot find a global minimum. Two popular alternatives[4] are Powell's method [Pow64; Pre02; Och03a] and the downhill simplex method [Nel65; Pre02; Cet05].

Nowadays, all SMT systems use a log-linear combination of feature models, optimized according to a certain automatic measure on the development data.

2.2.5 Re-scoring

In [She04] a discriminative re-scoring (or re-ranking) strategy is introduced for improving SMT performance. This technique works as follows:

- First, a baseline system generates n-best candidate hypotheses
- Then, a set of features which can potentially discriminate between good and bad hypotheses are computed for each candidate
- Finally, these features are weighted in order to produce a new candidate ranking

The advantage is that, given the candidate sentence, features can be computed globally, enabling rapid experimentation with complex feature functions. This approach is followed in [Och03b] and [Och04a] to evaluate the benefits of a huge number of morphological and shallow-syntax feature functions to re-rank candidates from a standard phrase-based system, with little success.

[4] In recent experiments at the 2006 John Hopkins University Summer Workshop on SMT, both methods achieved similar performance [Ber06]

2.3 Translation Evaluation

2.3.1 Automatic Metrics

There are many good ways to translate the same sentence, thus it is difficult to define objective criteria for translation evaluation. So far, no automatic translation evaluation measure has been generally accepted, so various measures are typically used instead. Some commonly used measures are:

WER (word error rate) or mWER (multi-reference word error rate) The WER is the minimum number of substitution, insertion and deletion operations that have to be performed to convert the generated sentence into the reference target sentence. For the mWER, a whole set of reference translations is used. In this case, for each translation hypothesis, the edit distance to the most similar sentence is calculated.

PER (position-independent word error rate) or mPER (multi-reference position-independent word error rate) These are the position independent versions of WER and mWER.

BLEU score This score measures the precision of unigrams, bigrams, trigrams, and four-grams with respect to a whole set of reference translations, and with a penalty for too short sentences [Pap01]. BLEU measures accuracy, thus larger BLEU scores are better. Recently, an interesting discussion with counterexamples of human correlation was presented in [CB06].

NIST score NIST evaluation metric, introduced in [Dod02], is based on the BLEU matrix, but with some alterations. Whereas BLEU simply calculates n-gram precision considering of equal importance each n-gram, NIST calculates how informative a particular n-gram is, and the rarer a correct n-gram is, the more weight it will be given. NIST also differs from BLEU in its calculation of the brevity penalty, and small variations in translation length do not impact the overall score as much.

METEOR score includes a word stemming process of the hypothesis and references to extend unigram matches (see [Ban05b])

2.3.2 Human Evaluation Metrics

Human evaluation metrics require a certain degree of human intervention in order to obtain the quality score. This is a very costly evaluation strategy that seldom can be

conducted. However, thanks to international evaluation campaigns, these measures are also used in order to compare different systems.

Usually, the tendency has been to evaluate *adequacy* and *fluency* (or other relevant aspects of translation) according to a 1 to 5 quality scale. Fluency indicates how natural the hypothesis sounds to a native speaker of the target language, usually with these possible scores: 5 for Flawless, 4 for Good, 3 for Non-native, 2 for Disfluent and 1 for Incomprehensible. On the other hand, Adequacy is assessed after the fluency judgement is done, and the evaluator is presented with a certain reference translation and has to judge how much of the information from the original translation is expressed in the translation by selecting one of the following grades: 5 for all of the information, 4 for most of the information, 3 for much of the information, 2 for little information, and 1 for none of it[5].

However, this strategy is not easy to apply for annotators and other approaches have shown to allow better inter-annotator agreement [CB07]. Thus recently other evaluation schemes have been introduced. One of them is the ranking of sentences. Annotators have to rank up to 5 sentences, ties being allowed.

Another trend is to manually post-edit the references with information from the test hypothesis translations, so that differences between translation and reference account only for errors and the final score is not influenced by the effects of synonymy. The human targeted reference is obtained by editing the output with two main constraints, namely that the resultant references preserves the meaning and is fluent. In this case, we refer to the measures as their human-targeted variants, such as HBLEU, HMETEOR or HTER as in [Sno05].

2.4 Word Alignment

In many current SMT systems word alignment is used to segment the sentence pair in a set of translation units. Word aligned corpora are also useful in a variety of other fields including automatic extraction of bilingual lexica and terminology [Sma96; Rib01], word sense disambiguation [Dia02] or grammar induction [Kuh04]. Word aligned corpora can also help for transferring language tools such to new languages [DY01]. Actually, the relevance of word alignment has been corresponded by several previous works on the matter, including shared tasks in the frame of HLT-NAACL 2003 and ACL 2005 Workshops on Building and Using Parallel Texts [Mih03; Mar05]. This section surveys

[5]These grades are just orientative, and may vary depending on the task.

the techniques that employ statistical or heuristic methods to obtain a word alignment of a bilingual sentence-aligned corpus.

Given a bilingual sentence pair, we refer to any defined set of links between lexical units that are translation of each other, as a word alignment of the sentence pair. Lexical units can be single words or groups of words (phrases). This definition does not specify a unique set of links, since the correspondence between lexical units is subjective and thus ambiguous.

The alignment between two sentences can be quite complicated. It can include word or phrase re-orderings, omissions, insertions and word-to-phrase alignments. Therefore, a very general alignment representation is needed. In particular, an ideal representation should allow various (possibly non-contiguous) source words to be aligned with various (possibly non-contiguous) target words, because sometimes it is not possible to align words of a phrase pair. For instance, in the English-Spanish phrase pair (two o'clock, las dos), if "two" should be aligned with "dos", it doesn't make sense to align "o'clock" with "las". It's actually the entire phrase "two o'clock" that should be aligned with "las dos".

A general definition of an alignment between two word strings is as follows: an alignment is a subset of the Cartesian product of the word positions, i.e. an alignment $\mathcal{A}$ is defined as:

$$\mathcal{A} \subseteq \{(j,i) : j = 1,\ldots,J; i = 1,\ldots,I\} \tag{2.14}$$

if the source and target sentences have got respectively I and J words. Some words are not aligned to any other word. The definition can be modified to take this into account explicitly by adding links of the type $(j,0)$ or $(0,i)$ ("NULL" links). This is equivalent to adding a "NULL" word at position 0 of the source and target sentences. In practice, the representation of equation 2.14 is hard to implement because it induces too many alignments to be considered. There are IJ possible connections between words. Hence, there are 2^{IJ} possible alignments. A typical way to restrict the problem, proposed by Brown *et al.*[Bro93] (see §2.1.1), is to assign each source word to *exactly one* target word. In this way, the number of possible alignments is limited to $(J+1)^I$.

Before comparing various systems, we will see in §2.4.1 the evaluation measures that allow a quantitative comparison of the alignments that are produced.

Two main approaches have been used to calculate basic word correspondence probabilities, referred to as *association approach* and *estimation approach* in Tiedemann [Tie03]. Och and Ney [Och03c] call the corresponding models respectively *heuristic models* and *statistical alignment models*. Actually both approaches use cooccurrence counts between lexical units. In the association approach, alignment probabilities are inferred by association measures, which are based on the statistical analysis of cooccurrence data. In the

estimation approach, alignment parameters result from statistical estimation theory and are adjusted such that the likelihood on the parallel training corpus is maximal. Association measures will be introduced in §2.4.2. Then all types of systems will be reviewed in §2.4.3.

2.4.1 Alignment Scoring Metrics

The metrics described here have been typically used in the recent literature and were the evaluation measures of the HLT-NAACL 2003 workshop [Mih03], and of the ACL 2005 Workshop on "Building and Using Parallel Texts" [Mar05]. Automatically computed alignments[6] are compared to a manually aligned reference corpus (gold standard) and scored with respect to precision (P), recall (R), F-measure (F) and Alignment Error Rate (AER), as defined in equations 2.15 to 2.17. The precision is defined as the proportion of computed links that are present in the reference. The recall is the proportion of reference links that were computed. The F-Measure is a way of combining both metrics [VR79]. The AER was introduced by Och and Ney [Och00b] in order to take into account the ambiguity of the manual alignment task, and involves unambiguous links (called S or Sure) and ambiguous links (called P or Possible). If there is a P link between two words in the reference, a computed link between these words is acceptable, but not compulsory. On the contrary, if there would be an S link between these words in the reference, a computed link would be compulsory. Given an alignment $\mathcal{A}$, and a gold standard alignment $\mathcal{G}$, we can define sets $\mathcal{A}_\mathcal{S}$, $\mathcal{A}_\mathcal{P}$ and $\mathcal{G}_\mathcal{S}$, $\mathcal{G}_\mathcal{P}$, corresponding to the sets of Sure and Possible links of each alignment. The set of Possible links is defined as the union of S and P links, or equivalently $\mathcal{A}_\mathcal{S} \subseteq \mathcal{A}_\mathcal{P}$ and $\mathcal{G}_\mathcal{S} \subseteq \mathcal{G}_\mathcal{P}$, or $\mathcal{A}_\mathcal{P} = \mathcal{A}$ and $\mathcal{G}_\mathcal{P} = \mathcal{G}$. To avoid possible confusions, the $_\mathcal{P}$ subscript will be omitted unless it is necessary, and "P links" will only refer to links labelled with P. The measures which are defined are the following:

$$P_S = \frac{|\mathcal{A}_\mathcal{S} \cap \mathcal{G}_\mathcal{S}|}{|\mathcal{A}_\mathcal{S}|}, \quad R_S = \frac{|\mathcal{A}_\mathcal{S} \cap \mathcal{G}_\mathcal{S}|}{|\mathcal{G}_\mathcal{S}|}, \quad F_S = \frac{2P_S R_S}{P_S + R_S} \tag{2.15}$$

$$P_P = \frac{|\mathcal{A} \cap \mathcal{G}|}{|\mathcal{A}|}, \quad R_P = \frac{|\mathcal{A} \cap \mathcal{G}|}{|\mathcal{G}|}, \quad F_P = \frac{2P_P R_P}{P_P + R_P} \tag{2.16}$$

$$AER = 1 - \frac{|\mathcal{A} \cap \mathcal{G}_\mathcal{S}| + |\mathcal{A} \cap \mathcal{G}|}{|\mathcal{A}| + |\mathcal{G}_\mathcal{S}|} \tag{2.17}$$

If only one type of links is considered in the alignment reference, 2.17 becomes:

$$AER_1 = 1 - \frac{2PR}{P + R} = 1 - F \tag{2.18}$$

[6]Hereinafter, computed alignments will refer to alignments to be evaluated (which are automatically computed in most cases), as opposed to reference alignments.

	t	$\neg t$	
s	$O_{11} = C(s,t)$	$O_{12} = C(s,\neg t)$	$= C(s)$
$\neg s$	$O_{21} = C(\neg s,t)$	$O_{22} = C(\neg s,\neg t)$	$= C(\neg s)$
	$= C(t)$	$= C(\neg t)$	$= N$

Table 2.1: Contingency table of observed frequencies for words s and t.

	t	$\neg t$
s	$E_{11} = \frac{C(s)C(t)}{N}$	$E_{12} = \frac{C(s)C(\neg t)}{N}$
$\neg s$	$E_{21} = \frac{C(\neg s)C(t)}{N}$	$E_{22} = \frac{C(\neg s)C(\neg t)}{N}$

Table 2.2: Contingency table of expected frequencies under the point null hypothesis of independence for words s and t.

2.4.2 Association Measures

Association measures are based on the statistical analysis of cooccurrence frequency data, widely used in the study of collocations in monolingual corpora. A thorough explanatory survey of this area can be found in Evert's PhD thesis [Eve05].

Given a (source,target) word pair (s,t), the parallel corpus token pairs are classified into the cells of a *contingency table* like the one depicted in Table 2.1, depending on whether the source token is equal to s and the target token is equal to t. In this table, $C(s,t) \ldots C(\neg s,\neg t)$ represent the observed joint frequencies of the different combinations of s and t occurring or not, and the row and column totals $C(s) \ldots C(\neg t)$ represent the observed frequencies of s and t occurring or not. The four cells of the contingency table add up to the total number of pair tokens, called the sample size N. Notice that there are two reasonable ways of counting cooccurrences in a sentence pair:

$$C_{sp}(s,t) = min[C_{sp}(s), C_{sp}(t)] \tag{2.19}$$

or

$$C_{sp}(s,t) = \begin{cases} 1 & \text{if } C_{sp}(s) > 0 \text{ and } C_{sp}(t) > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{2.20}$$

In Equations 2.19 and 2.20, the *sp* subscript indicates that frequencies are counted in the considered sentence pair.

Intuitively, a larger value of $C(s,t)$ indicates stronger association, while larger values of $C(s)$ and $C(t)$ indicate weaker association. The Dice coefficient is an association measure based on this intuition:

$$Dice = \frac{2C(s,t)}{C(s) + C(t)} \tag{2.21}$$

However the group of association measures most widely used in word alignment test if two words co-occur significantly more often than it would be expected if they would co-occur purely by chance (*i.e.* if they would be statistically independent). In terms of statistical analysis, a *null hypothesis* H_0 is defined (in this case, that the two words are independent), and these association measures are said to use the amount of evidence against the null hypothesis of independence as an association score. In the measures considered in this thesis, the null hypothesis is reduced to the so-called *point null hypothesis*, in which the probabilities expected for the null hypothesis are given by maximum-likelihood estimates. The contingency table for expected frequencies under the point null hypothesis is represented in Table 2.2. Under the null hypothesis of independence, the joint probability is given by the product of marginal probabilities: $p_{H_0}(x, y) = p(x)p(y)$, and maximum-likelihood estimates are: $p_{ML}(x, y) = C(x, y)/N$, $p_{ML}(x) = C(x)/N$, for x, y in $\{s, \neg s, t, \neg t\}$.

The probability (or likelihood) of the observed frequencies under the null hypothesis indicate if an outcome is usual or unusual. However, likelihood values do not quantify the amount of evidence against the null hypothesis, because they concern a single outcome. A quantity used to measure this amount is the *p-value*, which is the total probability of all possible outcomes that are at least as unusual as the observed contingency table. The p-value is computed by statistical hypothesis tests. When it falls below a certain threshold, the sample is said to provide *significant* evidence against the null hypothesis. Exact hypothesis tests compute the exact p-value. However these tests present two problems [Eve05]: "*(i) their numerical complexity and (ii) the difficulty of defining more "extreme" outcomes.*". These problems are addressed by *asymptotic hypotheses tests*, in which a test statistic is used, which distribution under the null hypothesis converges to a known limiting distribution for $N \to \infty$. This limiting distribution is used to derive an approximated p-value. For convenience, the test statistic itself is often used as an association score. The main test statistics used in the alignment literature are introduced below. Since they are based on the point null hypothesis, they can be understood as a comparison of the contingency table of expected frequencies (Table 2.2) and observed frequencies (Table 2.1).

Fung and McKeown [Fun94] use a t-score of word distribution vectors to generate a bilingual lexicon in a corpus divided in N segments:

$$t = \frac{O_{11} - E_{11}}{\sqrt{O_{11}}} \tag{2.22}$$

Gale and Church [Gal91] use ϕ^2, a χ^2-like statistic ($\phi^2 = \chi^2/N$) to identify word correspondences. Pearson's chi-squared test (χ^2) can be understood as a kind of mean

squared error:

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{2.23}$$

Since column and row sums are equal for expected and observed contingency tables, the χ^2 score can be expressed in a simplified form (see Appendix A).

Dunning [Dun93] showed that the log-likelihood ratio (LLR) was a better statistics to account for rare events occurring in large samples. The χ^2 score indeed overestimates their significance. The likelihood ratio test is based on the ratio between the maximum likelihood of the observed data under the null hypothesis and its unconstrained maximum likelihood (without making any assumptions about the parameters). The LLR association measure was used by Melamed for automatically constructing translation lexicon [Mel00] and by Moore [Moo05b] to build a word type association model in a word alignment system. The standard form of LLR is given by Equation 2.24 (see Appendix A).

$$LLR = 2 \sum_{i,j} O_{ij} \log \frac{O_{ij}}{E_{ij}} \tag{2.24}$$

Notice that χ^2 and LLR scores are two sided measures, which means that they don't distinguish between positive and negative association. Positive association indicates that the elements of a (s, t) pair cooccur more often than if they were independent, and negative association that they occur less often. For both measures, a one-sided test can be obtained by changing the sign of the score when $p(s, t) < p(s)p(t)$.

2.4.3 Word Alignment Approaches

The currently most widely used approach, which is also considered as the state-of-the-art in the field, is the approach based on IBM and HMM statistical translation models (see §2.1.1). After the model parameters have been trained, the Viterbi alignment for each sentence pair is obtained by maximising Equation 2.2. A systematic performance comparison in terms of AER of these models can be found in [Och03c], where the authors also advocate a positive correlation between AER improvement and translation quality. In general, very important ingredients of a good model seem to be a first-order dependence of word positions and a fertility model. It is also important to bootstrap the refined models with good initial parameters (HMM instead of model 2). The alignment parameters of HMM, Model 4 and Model 5 include a dependence on the word classes of the surrounding words. These word classes can be learnt automatically from bilingual corpora using clustering techniques [Och99a]. Statistical alignment models are typically trained by means of GIZA++ [Och00a], a freely-available implementation.

Toutanova *et al.* [Tou02] extended the IBM models introducing an additional lexicon probability for POS tags. García Varea *et al.* [GV02] added contextual information to the IBM models in the framework of maximum entropy, with small but consistent improvements in AER.

An inconvenient of IBM and HMM models is that only one-to-many alignments are allowed, since the alignment is restricted to a mapping from a source position to a target position. This problem can be tackled by performing source-target and target-source alignments, and symmetrising via the union of links, the intersection or other refined methods [Och03c; Koe].

Another popular approach to the word alignment problem is based on association scores. For each sentence pair, a matrix including the association scores between every word at every position can be calculated.

Melamed [Mel00] presented a method to obtain a word alignment from the association score matrix, the *competitive linking algorithm*. In a first step, the highest ranking word pair (l_i, r_j) is aligned. Then, the corresponding row and column are removed from the association score matrix. This procedure is iteratively repeated until every source or target language word is aligned. The resulting alignment contains only one-to-one links. A first refinement of the co-occurrence score function is Melamed's explicit noise model [Mel00]. First, word pairs are scored with an association metric and linked. Then, they are re-scored considering the ratio of the number of times a word pair is linked compared to the number of times it co-occurs.

In Ahrenberg *et al.*[Ahr98] the t-score is used as a basis, but the word-to-word approach is extended by considering also multi-word candidates. Some morphological information is also used (expressions with different suffixes might be treated equally), and weights are distributed depending on positions.

In Tiedemann [Tie03] the association score is obtained by the combination of various clues. An alignment clue C_i is defined as a weighted association A between two lexical units (words or phrases). The association A can be the Dice coefficient or a string similarity measure. Other clues can be estimated from word aligned training data. The features are based on POS labels, chunk labels and relative word position. The various clues are combined using the following addition rule:

$$\begin{aligned} C_{all}(s,t) &= P(a_{all}) = P(a_1 \cup a_2 \cup \ldots \cup a_n) \\ P(a_1 \cup a_2) &= P(a_1) + P(a_2) - P(a_1)P(a_2) \end{aligned} \tag{2.25}$$

The final alignment is obtained with a type of competitive linking algorithm that takes into account overlapping lexical units. This alignment can in turn be used to improve

the feature-based clues. Thus various features have been combined to obtain a refined association matrix.

Another model using various type of features is the probability model of Cherry and Lin [Che03]. An original aspect of this model is that it involves directly the conditional probability of the links given the sentence pair and the previous links:

$$P(A|E,F) = \prod_{t=1}^{T} P(l_t|E,F,l_1^{t-1}), \tag{2.26}$$

where T is the total number of links, A the alignment, E and F are the two aligned sentences, l_t is an individual link between two words. In this model, the probability of each link given its context is given by the product of two terms. One is the probability $Pr(l_t \,|\, e_{i_t}, f_{j_t})$ of a link given the co-occurrence count of its words. The other term is a product of feature contributions. The probability tables are calculated from a word aligned parallel corpus. The initial alignment is obtained with the ϕ^2 association score and a competitive linking algorithm restricted to one-to-one links.

In 2005, following an idea already hinted by Callison-burch *et al.* [CB04], several independent works demonstrated that discriminatively trained models can equal or surpass the alignment accuracy of the standard models, if the usual unlabeled bilingual training corpus is supplemented with human-annotated word alignments for only a small subset of the training data [Liu05; Itt05; Fra05; Moo05b]. Therefore, recent research efforts seem to be shifting towards a log-linear combination of feature models, estimated on a small word-aligned development set [Che06b; Fra06; Moo06; Blu06]. Notice that some systems include association score models as feature and some other include statistical alignment models as feature.

Chapter 3

Resources for Alignment and SMT

Because Machine Translation was a new subject for the research group, consequential preliminary work was needed. It consisted in developing linguistic resource and basic software, especially in the area of word alignment. This chapter is structured as follows:

- §3.1 deals with alignment evaluation resources. An automatic evaluation can be performed comparing an alignment to a manual reference. In the recent literature in this domain a consensus on scoring methods seems to appear (see §2.4.1). However there is little about common standards to build reference data, although it has a great impact on the scores, as will be shown in §3.1. This section motivates and details guidelines for word alignment evaluation and manual alignment. It also presents a manual reference we built for Spanish-English data from the European Parliament proceedings, and which is publicly available[1].

- We implemented a freely available[2] Perl toolkit to handle (visualize, evaluate, process) a set of sentences aligned at the word (or phrase) level. The list of available tools is outlined in §3.2.

- In §3.3, we mention other resources which were achieved during the thesis work.

[1] http://gps-tsc.upc.es/veu/LR/
[2] http://gps-tsc.upc.es/veu/personal/lambert/software/AlignmentSet.html

3.1 Guidelines for Word Alignment Evaluation and Manual Alignment

3.1.1 Related Work

In this thesis we focus on the evaluation of full text alignment, as opposed to the method adopted in the ARCADE [Vó0] and PLUG [Ahr00] projects, where the reference corpus was a collection of sample words.

Reference sets of fully aligned sentence pairs include data of the Blinker project [Mel98b] and test data of the word alignment shared task in the HLT-NAACL 2003 Workshop on "Building and Using Parallel Texts: Data Driven Machine Translation and Beyond". In both cases various annotators took part in the annotation task, but the strategy in case of disagreement between annotators varied. In the Blinker project, only one reference alignment was accepted, and the decision was taken by a vote between annotators. For the Romanian-English test data of the HLT-NAACL 2003 Workshop [Mih03], these cases were judged in an arbitration phase with a third annotator. The authors claim that an inter-annotator agreement was reached for all word alignments, so that only one resulting alignment had to be selected. For the French-English test data of the same workshop [Och00c], the resulting alignment was the union of (two) different proposals. To make this possible, two types of links were introduced: explicit ambiguous links, called P links, and unambiguous links, called S links (a more detailed description of S and P links was given in section 2.4.1). First, annotators were presented with the mutual errors and asked to improve their alignments where possible. Two resulting sets were obtained by each annotator: ${\mathcal{G}_{\mathcal{S}}}^i$ (S links) and ${\mathcal{G}_{\mathcal{P}}}^i$ (P links), where $i = 1, 2$. Then, alignments from both annotators were compared automatically and the final set of S links was produced by forming the intersection of Sure links ($\mathcal{G}_{\mathcal{S}} = {\mathcal{G}_{\mathcal{S}}}^1 \cap {\mathcal{G}_{\mathcal{S}}}^2$) and the union of Possible links ($\mathcal{G}_{\mathcal{P}} = {\mathcal{G}_{\mathcal{P}}}^1 \cup {\mathcal{G}_{\mathcal{P}}}^2$).

As to the Blinker strategy, we consider it arbitrarily neglects potentially important alignment information, since the fact that a solution wins the majority of votes does not prove it is correct. The arbitration phase carried out by Mihalcea and Pedersen is crucial to detect inadvertent mistakes. On the other hand, it is reasonable to think that when two different alignments are proposed, in many cases it is because both make sense. So we do not believe it is necessary to try to reach an agreement on only one proposal. Actually, according to our experience, such an agreement was not possible in most cases. Therefore, we concluded that it is important to take into account the ambiguity of the annotation task, as Och and Ney did by the introduction of P links. In Och and Ney's work, each annotator could also mark links that he judged ambiguous with P links, the S links being

the links considered unambiguous. The resulting reference is composed of 77 % of P links. As we will see in section 3.1.3, with such a large proportion of P links, the reference corpus favours high precision (as opposed to high recall) computed alignments[3]. Another issue to bear in mind is that an unnecessary large proportion of P links limits the discriminative power of the AER.

In this section, we propose a refined strategy to build a reference corpus which favours higher recall computed alignments. This is because we noted that our SMT system prefers such alignments. We first give some remarks about the scoring methods used in recent literature (described in §2.4.1). Then we study the influence of the reference corpus on alignment scores (section 3.1.3), so as to determine what type of reference corpus should be build depending on the requirements. In section 3.1.4, we present our general guidelines for building a reference corpus. Finally we describe the manual alignments we made for the Final Edition of the Spanish-English European Parliament Proceedings. The detailed alignment guidelines we used for Spanish-English alignment are provided in appendix B.

3.1.2 Alignment Scoring Methods

3.1.2.1 About the Alignment Error Rate

From the definition of the alignment error rate (Equation 2.17) the following property is derived:

$$\text{If } \mathcal{G}_S \subseteq \mathcal{A} \subseteq \mathcal{G}, \;\; \text{AER} = 0 \tag{3.1}$$

Note also that by definition of a Possible link, including more P links in the reference alignment can only lower the error rate.

3.1.2.2 Unlinked Word Representation

In a sentence pair, there are two ways of representing the fact that a word has got no correspondence in the other sentence. The first one is leaving this word unlinked. The second one is inserting an explicit link between this word and a fictitious word, called NULL word, in the alignment. A link to NULL is also called a NULL link. It should be pointed out that if links to NULL are used, their treatment will be the same as for all the other words, although the NULL word is a particular word: it does not make sense to have a word linked both to other words (with S links) and to NULL, whereas this is possible for normal words.

[3] As already indicated, computed alignments refer to alignments to be evaluated, as opposed to reference alignments.

The scores presented above are greatly affected by the representation of unlinked words. Both computed and reference alignments must have the same alignment mode, which can be one of the following:

- null-align, where each word is enforced to belong to at least one alignment (if a word does not belong to any alignment, a NULL Possible link is assigned by default).
- no-null-align, where all NULL links are removed from both computed and reference alignments.

Explicit Possible NULL links contribute to a lower P_P, R_P and a higher alignment error rate because it adds P links to $\mathcal{A}$ which are not necessarily in the reference. We prefer the "no-null-align" mode, since we consider that worsening the AER with links to an artificial word adds noise in the measure.

3.1.2.3 Link Weights

In the evaluations of Och and Ney [Och00c] and Mihalcea and Pedersen [Mih03], each link contributes with the same weight to the count of the various sets ($|\mathcal{A}_\mathcal{S}|$, $|\mathcal{G}_\mathcal{S}|$, $|\mathcal{A}_\mathcal{S} \cap \mathcal{G}_\mathcal{S}|$, $|\mathcal{A}|$, $|\mathcal{G}|$, $|\mathcal{A} \cap \mathcal{G}|$ and $|\mathcal{A} \cap \mathcal{G}_\mathcal{S}|$). This tends to give more importance to words aligned in groups than to words linked individually. To correct this effect, Melamed [Mel98b] proposes to weight each link. According to Melamed, the weight $w(\mathrm{x}, \mathrm{y})$ of a link between two words x and y should be inversely proportional to the number of links (num_links) in which x and y are involved:

$$w(\mathrm{x}, \mathrm{y}) = \frac{1}{2}\left[\frac{1}{num\_links(\mathrm{x})} + \frac{1}{num\_links(\mathrm{y})}\right] \tag{3.2}$$

We point out that weights must be calculated with respect to the union of computed and reference sets. Otherwise, weights would be different in each set, which would be a problem for calculating the intersection $|\mathcal{A}_\mathcal{S} \cap \mathcal{G}_\mathcal{S}|$, $|\mathcal{A} \cap \mathcal{G}|$ or $|\mathcal{A} \cap \mathcal{G}_\mathcal{S}|$, where common links should have a unique common weight. For measures involving only Sure alignments (P_S and R_S), they may be calculated with respect to the union of the Sure sets: $\mathcal{A}_\mathcal{S} \cup \mathcal{G}_\mathcal{S}$. For measures mixing Sure and Possible alignments (AER, P_P and R_P), weights may be calculated based on the union of the full sets: $\mathcal{A} \cup \mathcal{G}$.

3.1.3 Influence of the Reference Corpus

Alignment scores do not only depend on the metric used but also on some characteristics of the reference corpus.

First of all, results depend on the criteria of the annotators responsible for making the decision of what exactly is translation of what. If an agreement on a set of criteria can be reached, the resulting alignments should gain in consistency. This is especially important if different persons annotate the same corpus. Yet, we believe (and have experienced) that it is almost impossible to reach such an agreement on every detail, because of the ambiguity of the task. Ambiguous cases are naturally solved introducing Possible links in the reference corpus.

However, the proportion of P and S links in the reference corpus influences scores as well. In particular:

1. if $|\{\text{P links}\}| >> |\{\text{S links}\}|$, high precision computed alignments are favoured
2. if $|\{\text{S links}\}| >> |\{\text{P links}\}|$, high recall computed alignments are favoured

Let us consider case 1. Since there are few S links in $\mathcal{G}$, they must be the most probable links and they should be nearly all included in a high precision computed alignment $\mathcal{A}$. So $|\mathcal{A} \cap \mathcal{G}_\mathcal{S}| \approx |\mathcal{G}_\mathcal{S}|$. If there are many P links in the reference, they should include the high precision computed links, so $|\mathcal{A} \cap \mathcal{G}| \approx |\mathcal{A}|$. Thus in this case the AER will be small. Now, if we add links to the computed alignment to have a higher recall, $|\mathcal{A} \cap \mathcal{G}_\mathcal{S}|$ will not change, but $|\mathcal{A} \cap \mathcal{G}|$ may become smaller than $|\mathcal{A}|$ because some of the links we added may not be included in $\mathcal{G}$. So higher precision alignments get higher scores.

In case 2, increasing the number of computed links increases both intersections in the numerator ($|\mathcal{A} \cap \mathcal{G}_\mathcal{S}|$ and $|\mathcal{A} \cap \mathcal{G}|$) and lowers the error rate, thus higher recall alignments get higher scores.

Table 3.1 gives an empirical illustration of the same statement. It shows the AER of high recall and high precision alignments evaluated with reference corpora having a low S/P link ratio (case 1) and a high S/P link ratio (case 2). The first reference corpus [Och00c], composed of 484 French-English sentences from the Hansards corpus, contains 23% of S links and 77% of P links. The other one [Gis04b] , composed of 400 Spanish-English sentences from a translation of the Verbmobil database, contains 82% of S links and 18% of P links. Computed alignments were obtained with IBM translation model 4, using the Giza++ toolkit. Training data are 3000K words for the Hansards, and 200K words for Verbmobil. In this translation model, source to target and target to source Viterbi alignments are distinct. Their intersection has got a high precision and low recall whereas their union has got lower precision and higher recall.

In terms of AER, we see that the intersection is the best alignment if evaluated with the reference with low S/P link ratio, but the worst if evaluated with the reference with

high S/P link ratio.

		P_P (%)	R_S (%)	AER (%)
Verbmobil	Intersection	97.6	71.3	17.5
(S/P ratio=4.6)	Union	90.8	84.5	**12.3**
Hansards	Intersection	98.1	82.8	**9.1**
(S/P ratio=0.3)	Union	85.6	94.1	11.4

Table 3.1: Comparison of AER for high precision (intersection) and high recall (union) alignments based on IBM model 4, evaluated with corpora of different S/P link ratio.

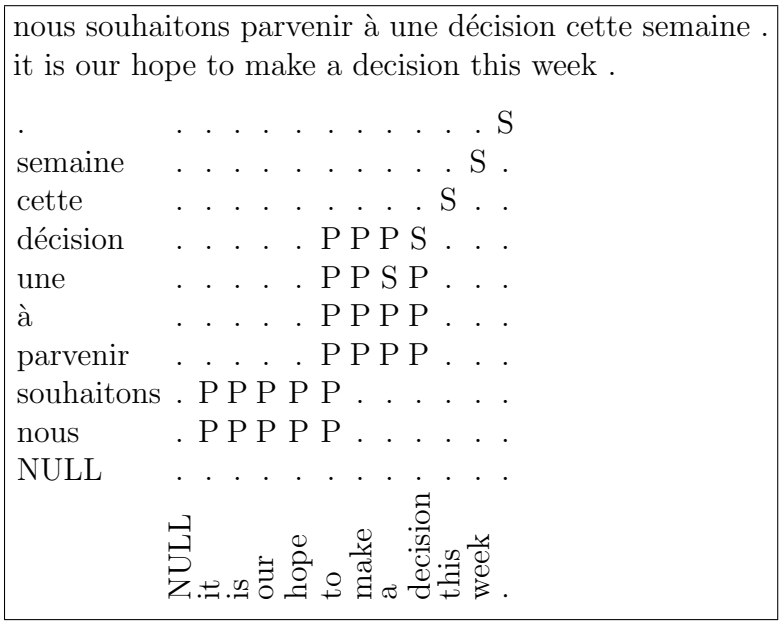


Figure 3.1: Example alignment with few Sure links and many ambiguous links

A related issue is that a reference corpus with many ambiguous links allows many different computed alignments to have the same AER, while being of different quality. Consider for instance sentence pair 76 of the reference corpus of Och and Ney [Och00c], which is displayed in figure 3.1. With such a reference, both alignments of figure 3.2 would get the same AER of zero (as well as all the alignments for which $\mathcal{G}_\mathcal{S} \subseteq \mathcal{A} \subseteq \mathcal{G}$), although the alignment in the lower half of figure 3.2 is much poorer than the other one.

Therefore, ambiguous links in the reference corpus should only be used for those situations representing equally acceptable computed alignments. Even though this may not be possible for all practical cases, all combinations of computed links matching the ambiguous reference links should ideally be considered alignments of equal quality. A second reason for this restricted criterion in the use of ambiguous links, is that the lower the S/P

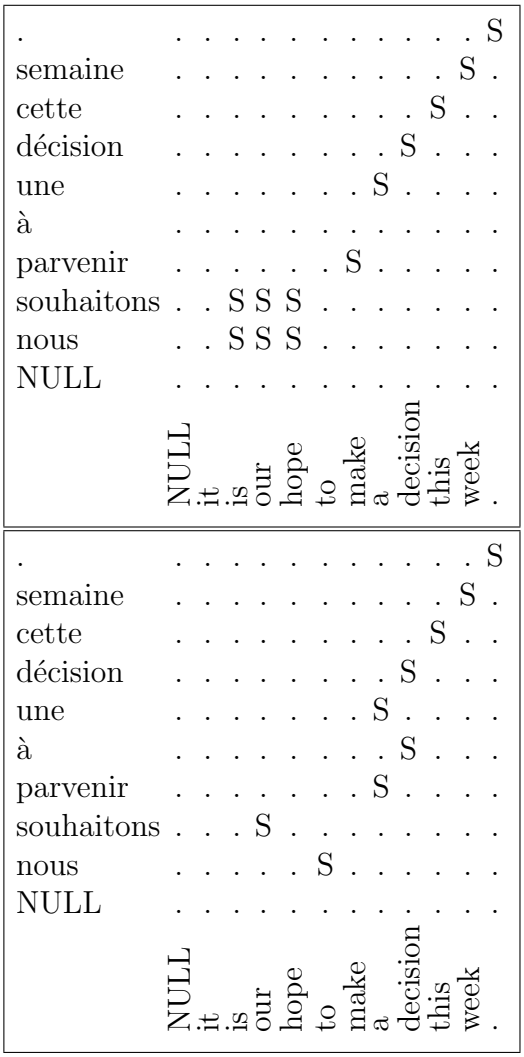


Figure 3.2: Two possible computed alignments with AER=0. Only the upper one is acceptable.

link ratio, the less discriminative power a given reference corpus provides (unless the size of the corpus is increased), as discriminative power results only from unambiguous links.

Contribution of the previous results to the word alignment field

In the recent litterature, numerous authors [Moo05a; Moo05b; Moo06; Tas05; Lia06; Che06b] evaluated their alignments using as unique reference the one of Och and Ney [Och00c]. As seen above, this reference corpus has a notably low S/P ratio and thus favours high precision computed alignments. As a consequence, the claims to improve alignment quality over this reference should be smoothed to a claim of improving alignment precision. In addition, some researchers typically implemented many features

and selected which ones to include in their systems depending of the impact of each feature on the AER. Thus their research was biased towards building systems producing high precision alignments. This is the case of Moore [Moo06]: "After experimenting with many features and combinations of features, we made the final selection based on minimizing training set AER". As another example, Moore [Moo05a] makes the following statement: "the approach does not work very well, because it tends to build clusters too often when it should produce one-to-one alignments. The problem seems to be that translation tends to be nearly one-to-one, especially with closely related languages, and this bias is not reflected in the method so far. To remedy this, we introduce two biases in favor of one-to-one alignments". We don't agree that "translation tends to be nearly one-to-one". It depends on the alignment reference, and thus on the annotator's criteria. In the reference corpus described in Subsection 3.1.4, for example, it is not the case. Thus Moore [Moo05a] explicitely introduced a bias in its system to satisfy a constraint which is actually only an artifact of the reference corpus. The introduction of this bias may seriously affect the ability of the alignment system to generalise when evaluated with another reference corpus.

To conclude, the results shown is this section should be taken into account in order to avoid reference-dependent biases in an alignment system.

3.1.4 General Alignment style

Taking into account the considerations presented in section 3.1.3, we propose some manual alignment guidelines for annotators. We expose the general ideas in this section. Details (some of them only valid for the English-Spanish language pair), are available in the appendix. The annotations that are presented here were performed by using the Alpaco editor [Ped03], which was modified to allow the distinction between ambiguous and unambiguous links.

3.1.4.1 Minimum lexical unit size

In the alignment, the correspondence between two lexical units should involve on both sides as few words as possible but as many words as necessary, with the requirement that the linked words or groups bear the same meaning. According to this, if two single words match, they should be linked together, but if there is no correspondence at a single word level, groups of words should be linked together. Groups of words linked together should be as small as possible, with all unambiguous (i.e. S) links.

This rule is illustrated with the example of figure 3.3. In the pair "In a community"–

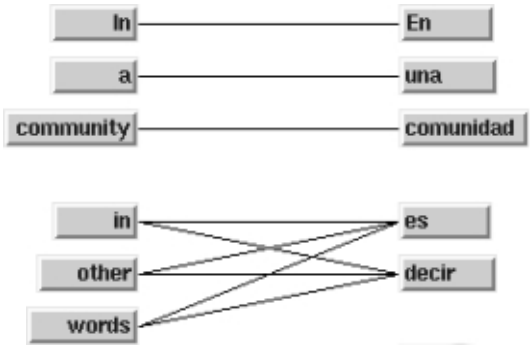


Figure 3.3: Minimum size of lexical units

"En una comunidad", a word-by-word correspondence can be specified, unlike in "in other words"–"es decir", where the whole phrase "in other words" has to be linked to the whole phrase "es decir". Note that the rule requiring to link groups of words together applies even though some parts of the group have a clear correspondence with parts of the other group. This is the case in the pair "let us see"–"a ver", where "see" is clearly the translation of "ver". Since the meaning of the group (the expression "let us see" or "a ver") is distinct from that of the sequence of each word's meaning, "ver" or "see" cannot be separated from the rest of the group. The whole group must be considered as an indivisible lexical unit.

3.1.4.2 Indivisibility rule

"The only valid elements in an alignment are single words and indivisible groups of words."

This means that a word (or phrase) cannot be aligned to only a part of a group. Let us illustrate this idea with the very simple example depicted in figure 3.4. In this example, imagine that "e1" may be translated to "f1 f2" and "e2" to "f2". If you link "e1" to both "f1" and "f2", it defines them as a group "f1 f2", thus "e2" has to be linked to every word of "f1 f2". This rule simply comes from the reciprocity of alignments: if "e1" corresponds to "f1 f2", then "f1 f2" must correspond to "e1", and not to "e1 e2".

Because they can represent two acceptable options of a single link, P links do not need to obey this rule.

3.1.4.3 Absence of correspondence

If some lexical unit does not have any corresponding lexical unit in the other language, it must remain unlinked.

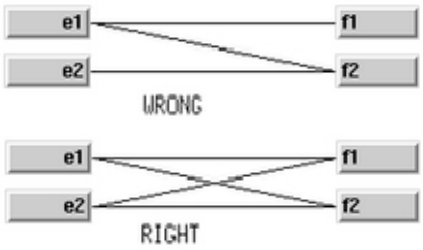


Figure 3.4: Illustration of the indivisibility rule

3.1.4.4 Ambiguity of manual alignments

As stated previously, sometimes various alignments are acceptable and there is no clear reason to choose one of them among others. To deal with this, explicit ambiguous links are introduced in the reference. An ambiguous link is used for a computed link that may or may not exist.

As explained in section 3.1.3, the presence of P links allows various alignment combinations with same AER score. Therefore, P links should ideally be used exclusively when these equally scored combinations are considered equally acceptable, even though in some practical cases this may not be possible.

In the modified Alpaco editor, S links are represented with plain lines and P links with dashed lines. This is illustrated in the example of figure 3.5. This example indicates that both aligning "I" to "quiero" or leaving "I" unlinked are considered acceptable. In the same manner, the three possibilities to align the English word "to" (unlinked, linked to "quiero" or linked to "informar") are considered equally acceptable. Note that if the computed alignment obeys the indivisibility rule, the possibility to align "to" to both "quiero" and "informar" (which would not be as acceptable as the others), is excluded.

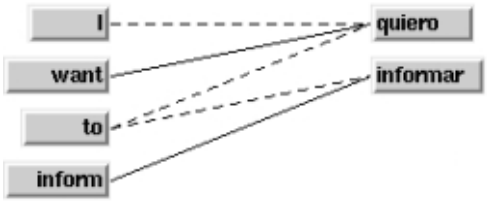


Figure 3.5: Example of ambiguous alignment

3.1.4.5 General Hints

In this subsection we focus on some delicate cases which could be encountered in many language pairs. Difficulties arise when a segment in one language is expressed with a different structure in the other.

This can occur when translation is not literal. In this case the annotator must judge if the source and target lexical units are translation of each other (S-linking them, *i.e.* associating them with a S link), could be considered as translations (P-linking them, *i.e.* associating them with a P link) or cannot be considered so (leaving them unlinked).

This also typically occurs with functional words, especially particles. Case markers, clitics and double clitics, double negations, degrees of comparison for the adjective, prepositions or adverbs of the phrasal verbs, appear in some languages and in others do not. The functional words which are required for grammatical reasons should be S-linked, unless they do not have an exact counterpart in the other language (in which case they should be P-linked or left unlinked, depending on the case). For instance "más grande" in Spanish (literally 'more great') should be S-linked, as a group, to "greater". In contrast, the English particle "to" following verbs like "to want" or "to need" has no counterpart in Spanish. Taking again the example depicted in figure 3.5, "to" could actually be associated with "quiero" or "informar", so it is P-linked to both of them.

Some constructions, where an exact counterpart exists and the correspondence is obvious, are also correct if a word is omitted, in which case a certain ambiguity is introduced. Consider a construction "(a) b"–"x y" where "a" is the translation of "x" and "b" is the translation of "y", but where "a" is omitted. In general, if "a" is redundant, "b" should be aligned to "x" with a P link and to "y" with a S link. If "a" is necessary, "b" should be linked to "y" and "x" remains unlinked. For example, in some languages (like Spanish), the pronominal subject may be omitted, but in some others (such as English), the realisation of the subject is obligatory. In this case, if the pronoun in Spanish-like language is omitted, the pronoun (in English-like language) should be P-linked to the verb in the Spanish-like language. If both pronouns are present, they should be S-linked (see the appendix).

With translations which are not very literal, there may be many anaphoric links, where a pronoun in one language is referring to a semantically equivalent word in the other language, but which is a noun or a proper noun. For example, "Informe Arias Canete"—"His report", where "his" actually represents "Arias Canete". Since "his" is not an acceptable translation of "Arias Canete", in our machine translation task, we want the alignment tools to be penalised if they associate them. So if anaphora is an artifact

of the translator, no link should be made. However, if anaphora is required because the repetition of the noun or pronoun would not be grammatically correct, the equivalents could be P-linked.

3.1.5 Spanish-English Alignment Reference

The bilingual texts have been extracted from the Final Edition of the European Parliament Proceedings, available from the European Parliament's website[4]. We used the version distributed by RWTH Aachen University, within the TC-STAR project[5]. This version has been aligned at the sentence level, tokenised and comprises punctuation marks and true case. For our reference corpus, 500 sentences of at most 100 words have been selected at random. This collection contains 14691 English words and 15458 Spanish words.

First, three annotators aligned manually, each, the first 50 sentence pairs of the reference corpus. Then, alignments were compared manually and cases of disagreement or doubts were discussed to refine the guidelines (without necessarily trying to reach an agreement). With the refined guidelines, the three annotators aligned the rest of the reference corpus. Again, alignments were compared manually to detect inadvertent mistakes. At this stage, differences between annotators alignments were considered distinct valid options. These different options were merged automatically to build the final reference alignment, in the following way. A score of respectively 1, 0 and -1 was given to a S link, a P link and the absence of link. For each possible link, annotator's scores were summed. The merged link was set to respectively Sure or absent when the sum was strictly greater or strictly less than half the number of annotators. In the other cases, it was set to Possible. For example, if the three annotations were S, S and P, sum was equal to 2, which is greater than 1.5, so the final link was set to S.

We also considered taking the majority annotation, or forming the intersection of Sure links ($\mathcal{G}_\mathcal{S} = {\mathcal{G}_\mathcal{S}}^1 \cap {\mathcal{G}_\mathcal{S}}^2 \cap {\mathcal{G}_\mathcal{S}}^3$) and the union of Possible links ($\mathcal{G}_\mathcal{P} = {\mathcal{G}_\mathcal{P}}^1 \cup {\mathcal{G}_\mathcal{P}}^2 \cup {\mathcal{G}_\mathcal{P}}^3$) of annotators 1,2 and 3. The sum method was seen to provide the most coherent reference. It was also the most discriminative method, considering the difference between AER of a specific automatically computed alignment (union of GIZA++ IBM model 4 source-target and target-source alignments) and the highest annotator's AER (see table 3.2).

AER between annotators before merging their respective alignments, as well as with respect to the final reference, are indicated in table 3.2. These scores have been calculated in no-null-align mode and with equal weights for each link (see sections 3.1.2.2 and 3.1.2.3). Scores are different if a link set is taken as reference or computed set, because the AER

[4]http://www.europarl.eu.int
[5]http://www.tc-star.org

metric distinguishes between S and P links in the reference, but not in the computed set.

Computed \ Ref	A_1	A_2	A_3	Reference
A_1	-	8.3	12.3	1.0
A_2	8.6	-	13.6	1.5
A_3	12.4	13.0	-	2.4

Table 3.2: AER between annotators and with respect to the final reference. The alignments of each annotator have been taken as reference to calculate the AER of the other annotators, whose alignments have been considered as computed alignments. Last column displays the AER of each annotator with respect to the final reference.

The final reference contains 33.3% of P links and 66.7% of S links (S/P ratio is 2.0). It is available from the web [6].

In order to facilitate the comparison of results, the complete set was split into a 100 sentences development set and a 400 sentences test set. The test data statistics (including number of sentences, number of words, vocabulary and average sentence length for each language) are reported in Table 3.3.

	sent	words	vocab.	avg len
English	400	11.7 k	2.7 k	29.1
Spanish		12.3 k	3.1 k	30.4

Table 3.3: *Alignment test data statistics.*

3.2 The AlignmentSet Toolkit

The Lingua-AlignmentSet distribution is a Perl Tools Library (and command-line utilities) to handle an Alignment Set, i.e. a set of sentences aligned at the word (or phrase) level. It provides methods to display the links, to apply a function to each alignment of the set, to evaluate the alignments against a reference, and more. One of the objectives of the module is to allow the user to perform all these operations without bothering with the particular physical format of the Alignment Set. Anyway it also provides format conversion methods.

This library is also available from the Natural Language Software Registry[7] of the Association for Computational Linguistics, as well as from the Perl CPAN archive[8].

The available tools are listed below. Nearly all tools have options to indicate the input

[6] http://gps-tsc.upc.es/veu/LR/epps_ensp_alignref.php3
[7] http:registry.dfki.de, section "Written Language-Alignment tools"
[8] http://www.cpan.org/

and output format, the treatment of NULL links (see §3.1.2.2) and the range of input lines to be processed.

- Visualisation tool: displays the aligned sentence pairs as link enumeration or matrix
- Evaluation tool: calculates Precision, Recall, F-measure, AER.
- Processing tool: apply a function to each alignment of the Alignment Set. Implemented functions include (see more with the '-man' option):
  - regexpReplace: substitutes, in a side of the corpus, a string (defined by a regular expression) by another and updates the links accordingly. Note: function based in "algorithm::diff", which in some cases doesn't find the minimal set of links to be changed. To avoid this the "replaceWords" function should be used.
  - replaceWords: substitutes, in a side of the corpus, a string (of words separated by a white space) by another and updates the links accordingly.
  - intersect, getUnion: takes respectively the intersection and union between source-to-target and target-to-source alignments
  - joined2ManyToMany,manyToMany2joined: respectively removes or introduces underscore between links of many-to-many groups in source to target alignment
  - internal functions: like splice or getAlClusters, a function which returns the alignment as clusters of positions aligned together. To be used within the AlignmentSet perl library.
- Corpus adaptation tool: adapts the links of an Alignment Set to a slightly different bilingual corpus. For example, use it if you have a manual alignment reference that you need to adapt to a different version of the corpus (e.g. with different tokenization).
- Corpus reordering tool: places sentence pairs of a secondary corpus at the head of the Alignment Set, in the same order.
- Format conversion tool
- Symmetrisation / Multi-word detection tool: Detects multi-words based on asymmetries between source-target and target-source alignments [Lam04].

The multi-word expression detection method is described in a paper by Lambert and Castell [Lam04].

3.3 Other Contributions

One of the earliest work of the thesis is a participation in the creation of LC-STAR consortium's trilingual corpus [Arr04].

Another contribution is a participation in a preliminary feasibility study for Chinese-Spanish SMT [Ban06]. The main issue is the generation of a training corpus, since no large parallel corpus is currently available for this language pair.

3.4 Conclusions and Further Work

We implemented a toolkit to visualise, evaluate and process a set of word aligned sentence pairs.

We described word alignment scoring metrics and pointed out that evaluation results are very sensitive to the scoring method details, like unlinked words representation or link weighting scheme.

We showed that scores are also sensitive to the type of reference corpus used. From our experience, full text alignment is an ambiguous task and it is difficult and unnecessary to expect various annotators to perform exactly the same alignments. Ambiguous cases can be represented with P links, while unambiguous ones can be represented with S links. However the ratio between S and P links in the reference has got a critical influence on the alignment error rate: low S/P link ratios favour high precision computed alignments, whereas high S/P link ratios favour high recall computed alignments. This result should be taken into account in order to avoid reference-dependent biases in an alignment system.

A related issue, which has got a less critical effect on the scores, is that an excessive amount of P links in the reference reduces unnecessarily the discriminative capacity of the AER.

Taking these remarks into account, we wrote manual alignment guidelines. Based on these guidelines, a freely available manual alignment reference for the Spanish-English European Parliament Corpus was built.

This work highlights the fact that the evaluation scores should not depend so much on the ratio between S and P links in the reference. Further work may include metrics which would take this ratio into account in order to smooth scores accordingly.

Work reported in this chapter is also presented in the following publications:

- [Lam04] Patrik Lambert and Núria Castell. 2004. **Alignment of parallel corpora exploiting asymmetrically aligned phrases**. In Proc. of the LREC 2004 Workshop on the Amazing Utility of Parallel and Comparable Corpora, pp 26-29. Lisbon, Portugal, May 25.

- [Lam05b] Patrik Lambert, Adrià de Gispert, Rafael Banchs and José B. Mariño. 2005. **Guidelines for Word Alignment Evaluation and Manual Alignment**. Language Resources and Evaluation, 39 (4) pp. 267-285. Springer.

- [Arr04] Victoria Arranz, Núria Castell, Josep M. Crego, Jesús Giménez, Adrià de Gispert and Patrik Lambert. 2004. **Bilingual Connections for Trilingual Corpora: An XML Approach**. In Proc. of the LREC 2004, Lisbon, Portugal, May 26-28.

- [Ban06] Rafael E. Banchs, Josep M. Crego, Patrik Lambert, José B. Mariño. 2006. **A Feasibility Study For Chinese-Spanish Statistical Machine Translation**. Proc. of the 5th Int. Symposium on Chinese Spoken Language Processing (ISCSLP), pp. 681-692. Kent Ridge, Singapore, December 13-16.

Chapter 4

N-gram-based Machine Translation

In this chapter we describe the translation system developed at UPC. In §4.1, the n-gram translation model, which is the core of the system, is presented. In §4.2, we describe the full SMT system, which consists of a log-linear combination of the translation model and four feature functions. In §4.3, translation evaluation results are shown.

4.1 Tuple N-gram Model

The n-gram translation model actually consists in a n-gram language model of bilingual units (which are referred to as tuples), as specified by the following equation:

$$p(\mathbf{s},\mathbf{t}) \approx \prod_{k=1}^{K} p((\tilde{s},\tilde{t})_k|(\tilde{s},\tilde{t})_{k-1},(\tilde{s},\tilde{t})_{k-2},\ldots,(\tilde{s},\tilde{t})_{k-n+1}) \tag{4.1}$$

where $\mathbf{s}$ and $\mathbf{t}$ refer respectively to the source and target sentence and $(\tilde{s},\tilde{t})_k$ to the k^{th} tuple of a given bilingual sentence pair.

Tuples are extracted from a word-to-word aligned corpus. More specifically, word-to-word alignments are performed in both directions, source-to-target and target-to-source, by using GIZA++ [Och03c], an implementation of IBM Model 4, and tuples are extracted from the union set of alignments according to the following constraints [Gis04a]:

- a monotonous segmentation of each bilingual sentence pairs 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.

As a consequence of these constraints, only one segmentation is possible for a given sentence pair. Figure 4.1 presents a simple example illustrating the tuple extraction process.

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 4.1: Example of tuple extraction from an aligned bilingual sentence pair.

Once tuples have been extracted, the tuple vocabulary is pruned by using histogram counts. This pruning is performed by keeping the N-most frequent tuples with same source sides. After pruning, the tuple n-gram model is trained by using the SRI Language Modelling toolkit [Sto02] and the improved back-off smoothing method proposed by Kneser and Ney [Kne95]. Thus, whereas in most earlier tuple-based work (see §2.1.3) the translation model was implemented by using finite-state transducers, in the system described here the translation model is implemented by using a large-vocabulary n-gram language modelling tool which includes the most advanced smoothing strategies.

Two important issues regarding this translation model must be mentioned:

- When tuples are extracted, some words always appear embedded into tuples containing two or more words, so no translation probability for an independent occurrence of such words exists (consider for example the words "perfect" and "translations" contained in tuple t_4 of Figure 4.1). To overcome this problem, the tuple n-gram model is enhanced by incorporating 1-gram translation probabilities for all the embedded words [Gis04a], which are extracted from the intersection set of alignments.
- It occurs very often that some words linked to NULL end up producing tuples with

NULL source sides. This cannot be allowed since no NULL is expected to occur in a translation input. This problem is solved by preprocessing alignments before tuple extraction such that any target word that is linked to NULL is attached to either its precedent or its following word. De Gispert[Gis06] provides a detailed report on this issue.

4.2 Translation System

An n-gram SMT system consisting of the translation model alone can produce good translations without additional target language model since the target language is modeled inside the bilingual n-gram model. However, search is more informed with additional models. The UPC-TSC translation system implements a log-linear combination of feature functions (see §2.2.3). The first feature function is the tuple n-gram model. This section describes each of the four additional feature functions. Also, a description of the customised decoding tool that is used is presented.

4.2.1 Target Language Model

The second feature function implemented is a target language model. This model is trained from the target side of the bilingual corpus by using the SRI Language Modelling toolkit and, again, the Kneser-Ney smoothing method.

An extended target language model might also be obtained by considering additional information from other available monolingual sources. These extended target language models are actually computed by performing a log-linear combination of independently computed target language models. The weights of the log-linear combination are adjusted so that perplexity, with respect to a given development set, is minimised.

4.2.2 Word Bonus Model

The third feature function corresponds to a word bonus model. This function introduces a sentence length bonus in order to compensate the system preference for short output sentences. This bonus depends on the total number of words contained in the partial translation hypothesis, and it is computed as follows:

$$wp(\mathbf{t}) = exp(\text{number of words in } \mathbf{t}), \tag{4.2}$$

where $\mathbf{t}$ refers to the partial translation hypothesis.

4.2.3 Lexicon Models

The fourth and fifth feature functions correspond to a forward and backwards lexicon models. They actually constitute complementary translation models. These models provide lexicon translation probabilities for each tuple based on the word to word IBM model 1 probabilities (see §2.1.1). This type of feature was one of the features tried by Och *et al* [Och04a] which achieved best performance. These lexicon models are computed according to the following equation:

$$p((\tilde{s},\tilde{t})_k) = \frac{1}{(I_k+1)^{J_k}} \prod_{j=1}^{J_k} \sum_{i=0}^{I_k} p_{IBM1}(s_j|t_i) \tag{4.3}$$

where s_j and t_i are the j^{th} and i^{th} words in the source and target sides of tuple $(\tilde{s},\tilde{t})_k$, being J_k and I_k the corresponding total number words in each side of it. The position $i = 0$ in corresponds to a NULL target word (as mentioned earlier, NULL in tuple source sides can not be allowed).

For computing the forward lexicon model, IBM model 1 probabilities from GIZA++ source-to-target alignments are used. For a missing translation word pair, where $p_{IBM1}(s_j|t_i) = 0$ according to the model, the same constant ($p_{IBM1}(s_j|t_i) = 10^{-40}$) as Och *et al* was used as smoothing value.

The implementation of this feature is fairly obvious, but presents some variants. For example, it doesn't always make sense to include the target NULL word in the $\sum_i$, so the following cases were considered:

1. include it if $J_k > 1$
2. include it if $J_k > I_k$
3. include it always
4. never include it

Best results were obtained in the first case.

The backward lexicon model is calculated in the same way as the forward one, with the only difference that the IBM 1 lexical parameters are estimated from alignments computed in the target-to-source direction.

Notice also that Equation 4.3 is the same as the original equation (2.3) with $\epsilon = 1$, although it is applied in a different context. The original equation was designed to give a probability to translation hypotheses of the same source sentence. Thus it is only

normalised in function of the number of target words. On the contrary, the lexicon models give probabilities to tuples of different source and target length. The forward lexicon probability is only normalised in function of target words[1] and thus benefits to tuples with shorter source side. Similarly, the backward lexicon model benefits to tuples with shorter target side. Thus lexicon models contribute in two ways to the system. Firstly, they rank tuples independently from their context. This can be useful to reduce noise, especially when the bilinguel model backs off to lower n-grams. Secondly, they favour shorter tuples. This may result in promoting high-order bilingual n-grams (by catenating short tuples) rather than using long unfrequent tuples which tend to cause a back-off fall to unigram. De Gispert [Gis06] tried to split the model into these two supposed contributions: forward and backward normalised lexicon models and a tuple bonus feature function (propotional to the number of tuples in the hypothesis). The results showed that the combination of normalised lexicon models and tuple bonus achieves a significant improvement over the baseline (the system without these features), as well as over their separate contributions (*i.e.* over the baseline with only the normalised lexical models or with only the tuple bonus). However, obtained results with this combination are still slightly worse than with the lexicon models of Equation 4.3. This is possibly explained by the fact that lexicon models depend on tuple source length and tuple target independently, while the simple tuple bonus depends more globally on source and target tuple length.

Finally, alternative implementations of the lexicon models would consist, for example, in substituting IBM model 1 probabilities by IBM model 2 or HMM model lexical probabilities, or even by a link relative frequency on the word-aligned training data.

4.2.4 *N*-gram Based Decoder

The search engine used for the presented translation system was developed by Crego *et al.* [Cre05b]. This decoder, which takes into account all the five feature functions described above simultaneously, implements a beam-search strategy and allows for three 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 the K-best ranked ones.

[1] In order to normalise the model with respect to source words, the product over source positions j should be introduced inside a $J_k th$ root $\sqrt[J_k]{}$.

Additionally, the decoder 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 present tuple and tuple n-gram history.

In order to achieve a fair comparison between candidate tuples when decoding a sentence, partial hypothesis are grouped in different lists according to their length (all the hypothesis in a list have translated the same words of the source sentence). In this way, probabilities of tuple sequences that are compared at each decoding step are on average of the same order of magnitude.

4.2.5 Scaling Factors Optimisation

We implemented an optimisation tool based on the downhill simplex algorithm [Pre02] (see §2.2.4). The method uses a geometrical figure called a *simplex* consisting, in N dimensions, of $N+1$ points and all their interconnecting line segments, polygonal faces, etc. The starting point is a set of $N+1$ points in parameter space, defining an initial simplex. At each step, the simplex performs geometrical operations (reflexions, contractions and expansions) until a local minimum is reached. The implemented tool has been designed to be easily extended with new optimisation algorithms or optimisation tasks.

Two optimisation schemes are possible. In the first one (depicted in Figure 4.2), the development corpus is translated at each iteration. With 4 parameters (one parameter can remain fixed to 1, the others being scaled accordingly), the algorithm converges after about 50 or more iterations. Thus, in this scheme, in the order of 50 development corpus translations are required.

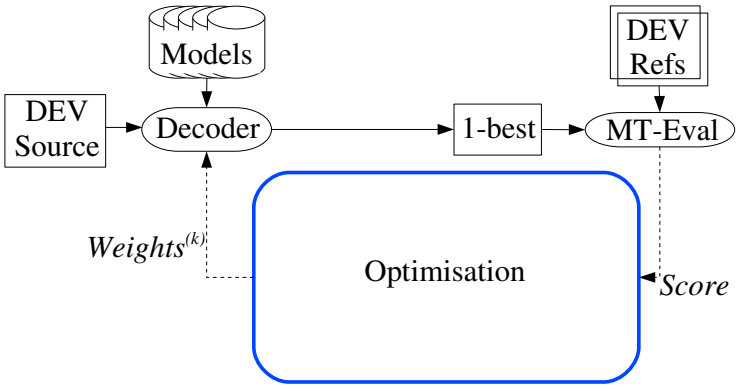


Figure 4.2: Single loop optimisation scheme.

In the second scheme, an N-best list is produced by the decoder. The optimisation algorithm is used to minimise the translation error while rescoring this N-best list. With

the optimal coefficients, a new decoding is performed so as to produced an updated N-best list [Ber06]. This process converges after only 5 to 10 decodings. For each internal optimisation, about 50 iterations are still required, but each iteration is much shorter since it only requires to rescore an N-best list. Optimisation based on re-scoring an N-best list does not always provide adequate parameters to be used in a different context, especially if the N-best list is too homogeneous. This in turn causes the next N-best list (decoded with these not adequate parameters) to be of low quality. In order to reduce the risk of entering in this vicious circle, we always add the best N-best list obtained so far to the current N-best list during re-scoring.

The SMT scaling factors optimisation enables large translation quality gains. Nevertheless, this process is also the origin of a serious experimental issue. As mentioned in §2.2.4, due to the difficulty of the optimisation task, no global minimum can be found in practice. Thus a slight modification of initial parameters can result in an appreciable difference in both the value of the local minimum found and the value of the optimal parameters. This difference is transmitted when these parameters are used to translate a test corpus. When a translation system is compared to a baseline, the difference arising only from the tuning process can be even greater than the difference arising from the two systems differences, leading to insignificant results. In some data sets, some inconsistencies of the tuning method have also been reported [Cre05a]. In some way, this problem also questions the reproducibility of the results. This problem will be further studied in chapter 6.

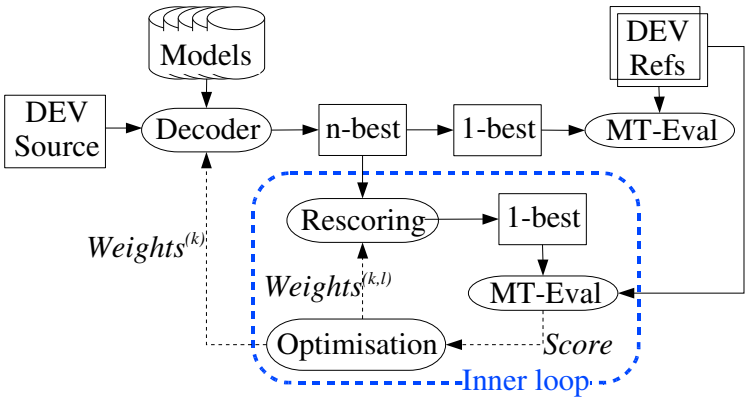


Figure 4.3: Double loop optimisation scheme.

4.3 Translation of the European Parliament Proceedings

This section presents and discusses translation results obtained at the TC-STAR internal evaluation of march 2005. Translations were required for the European Parliament Plenary Sessions (EPPS) tasks in both directions, English to Spanish and Spanish to English. First the EPPS data as well as the training procedure are described. Then some experimental results are presented.

4.3.1 European Parliament Data

The EPPS data set corresponds to the parliamentary session transcriptions of the European Parliament and is currently available at the Parliament's website (`http://www.europarl.eu.int/`). In the case of the results presented here, we have used the version of the EPPS data that was made available by RWTH Aachen University through the TC-STAR consortium. The training and test data used included session transcriptions from April 1996 until September 2004, and from November 15th until November 18th, 2004, respectively. This data set will be referred to as EPPS-04.

Table 4.1, presents some basic statistics of both, training and test, data sets for each considered language: English (eng) and Spanish (spa). More specifically, the statistics presented in Table 4.1 are, the total number of sentences, the total number of words, the vocabulary size (or total number of distinct words) and the average number of words per sentence.

1.a.- Train data set

Lang.	Sentences	Words	Vocab.	Aver.
Eng	1.22 M	33.4 M	105 k	27.3
Spa	1.22 M	34.8 M	169 k	28.4

1.b.- Test data set

Lang.	Sentences	Words	Vocab.	Aver.
Eng	1094	26.8 k	3.9 k	24.5
Spa	840	22.7 k	4.0 k	27.0

Table 4.1: Basic statistics for the considered (EPPS-04) training (a) and test (b) data sets (M and k stand for millions and thousands, respectively).

4.3.2 Preprocessing and Training

The training data was preprocessed by using standard tools for tokenising and filtering. By tokenisation, we refer to separating punctuation marks, classifying numerical expressions into a single token, and in general, simple normalisation strategies tending to reduce vocabulary size without an information loss (ie. which can be reversed if required). In the filtering stage, sentence pairs with a word ratio larger than a threshold (here equal to 3) have been removed , as well as blank lines or sentence pairs with at least one sentence of more than 100 words in length.

Then, a tuple set for each translation direction, Spanish to English and English to Spanish, was extracted from the union set of alignments, and the resulting tuple vocabularies were pruned, as explained in section 4.1. Table 4.2 presents the total tuple vocabulary size, as well as tuple vocabulary sizes for two different pruning values $N = 30$ and $N = 20$. The best trade off between translation quality and computational expenses have been obtained for values of $N = 20$ and $N = 30$ for Spanish to English and English to Spanish, respectively. Notice from Table 4.2, that in both cases the resulting tuple vocabulary sizes are very similar.

Direction	Total	$N = 30$	$N = 20$
Spa → Eng	2.5 M	2.1 M	2.0 M
Eng → Spa	2.5 M	2.0 M	1.9 M

Table 4.2: Tuple vocabulary sizes for different pruning values (M stands for millions).

A tuple 3-gram translation model was trained for each translation direction by using the respective pruned tuple sets. Then, each model was enhanced by including the 1-gram probabilities for the embedded word tuples, which were extracted from the intersection set of alignments. 3-grams were pruned out if they occurred only once[2].

In order to better justify the choice of the union alignment set, Table 4.3 presents model sizes and translation accuracies for the tuple n-gram model when tuples are extracted from different alignment sets. Translation accuracy is measured in terms of the translation BLEU, which is computed here for translations generated by using the tuple n-gram model alone. Both translation directions, Spanish-to-English (ES → EN) and English-to-Spanish (EN → ES), are considered in this table. 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 [Och03c].

[2] A thorough study of the effect of training parameters like n-gram model history length, tuple pruning or n-gram pruning is available in de Gispert's PhD thesis [Gis06]

Direction	Alignment set	Tuple voc.	2-grams	3-grams	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

Table 4.3: Tuple vocabulary sizes and their corresponding number of n-grams (in millions), and translation accuracy when tuples are extracted from different aligment sets. Notice that BLEU measurements in this table correspond to translations computed by using the tuple n-gram model alone.

As clearly seen from Table 4.3, the union alignment set happens to be the most favorable one for extracting tuples in both translation directions since its provides a significantly better translation accuracy, in terms of translation BLEU, than the other two alignment sets considered. Notice also from Table 4.3, that the union set is the one providing the smallest model sizes according to the total amount of 2-grams and the amount of 3-grams which appeared at least once. This might explain the improvement observed in translation accuracy, with respect to the other two cases, in terms of model sparseness.

To complete the translation system, a target language model and the forward and backwards lexicon models were computed for each translation direction according to the procedures described in subsection 4.2.

Once the models were computed, sets of optimal log-linear coefficients were estimated for each translation direction and system configuration according to the procedure described in subsection 4.2.5. As will be described in the next section, four different system configurations were considered. For all these optimisations, a development data set of five hundred sentences was used. This data corresponded to parliamentary session transcriptions from October 21th until October 28th, 2004, so it was independent from both the training and the test data sets. The development data included three reference translations for both English and Spanish.

Finally, the English and Spanish test data was translated into Spanish and English, respectively, by using the computed models and the estimated optimal coefficients for each system configuration and translation direction. The n-gram based decoding tool described in subsection 4.2.4 was used for generating the translations. The translation results are presented in the next section. All the results presented here were obtained by using the monotonous search modality of the decoder (i.e. without including reordering capabilities) and the decoding was always guided by the source. Threshold pruning was not used, but

histogram pruning was performed. A value of $K = 50$ for histogram pruning happened to provide a good trade off between translation quality and computational expenses for both directions, English to Spanish and Spanish to English.

Finally, regarding the UPC-TSC system efficiency, we can mention that in both translation directions the memory required during decoding was around 1.2 Giga-bytes. Translation times, on the other hand, vary according to the number of sentences and translation direction. However, as illustrative examples we can give the following measurements realised with a 3 GHz processor: Spanish to English, 840 sentences, total translation time = 1300 seconds; and English to Spanish, 1094 sentences, total translation time = 2300 seconds.

4.3.3 Task Results and Discussion

In order to evaluate the translation model performance and the feature contributions to the translation tasks, four different system configurations were considered in each translation direction:

- *Baseline System*: In this case, only the translation model is used (here a tuple 3-gram model).
- *Target-reinforced System*: In this case, the translation model is used along with the target language and word bonus models.
- *Lexicon-reinforced System*: In this case, the translation model is used along with the forward and backwards lexicon models.
- *Full System*: In this case, the translation model is used along with all the four additional feature functions.

Tables 4.4 and 4.5 present the mWER and BLEU values obtained.

System	mWER	BLEU
Baseline	39.55	0.476
Target	39.61	0.485
Lexicon	35.65	0.537
Full	34.91	0.543

Table 4.4: Evaluation results for the Spanish to English translation task.

System	mWER	BLEU
Baseline	44.45	0.428
Target	44.67	0.436
Lexicon	41.69	0.450
Full	40.96	0.466

Table 4.5: Evaluation results for the English to Spanish translation task.

The algorithms used for computing the mWER and BLEU were the official TC-STAR evaluation tools distributed by ELDA (`http://www.elda.org/`). Two reference translations were available for each language test set.

As can be observed from tables 4.4 and 4.5, the inclusion of the four feature functions into the translation system definitively produces an important improvement in translation quality in both translation directions. Particularly, it becomes evident that features with the most impact on translation quality are the lexicon models. The target language model and the word bonus also contributed to improve translation quality, but in less degree.

Another important observation, which follows from comparing results between both translation directions, is that in all the cases Spanish to English translations are consistently and significantly better than English to Spanish translations. This is clearly due to the more inflected nature of Spanish vocabulary. According to this, significant efforts should be dedicated for properly exploiting morphological analysis and synthesis methods for improving English to Spanish translation quality.

Additionally, a detailed review of about 130 translated sentence pairs, in each direction, was performed. This exercise resulted to be very useful since it allowed to identify the most common errors and problems related to the proposed SMT system in each translation direction.

A brief analysis of the reviewed translations revealed that most of translation problems encountered were basically related to the following four different types of errors:

- *Verbal Forms*: In this case, a great amount of wrong verbal tenses and auxiliary forms were detected. This problem happens to be the most common one, and it reflects the noisy nature of the training corpus for which, for example, passive voice translations of active voice statements (and vice-versa) are very common.
- *Omitted Translations*: In this case, a large amount of translations involving tuples with NULL target sides were detected. Although in some cases these situations correspond to a correct translation, most of the time they result in omitted-word errors.

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 4.6: The four best BLEU results for the EPPS translation task of 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/*

- *Reordering Problems*: In this case, the two specific situations that most commonly occurred were problems related to adjective-noun and subject-verb structures.
- *Concordance Problems*: In this case, inconsistencies related to gender and number were the most commonly found.

4.3.4 *N*-gram based SMT compared with phrase based SMT

The *n*-gram based translation system described here has been also evaluated and compared to other phrase-based translation systems in the context of the European Project TC-STAR. A detailed description of the first evaluation campaign is available through the consortium's website as a progress report [Ney05].

Table 4.6 presents the four best translation BLEU results for the EPPS translation task of 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 European Parliament Plenary Sessions, 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 seen from table 4.6, performance of the *n*-gram based translation system is among

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

Table 4.7: 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 website: *http://www.statmt.org/wpt05/mt-shared-task/*

the three best systems for the translation directions and conditions considered in the first TC-STAR's 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 [Ban05a]. 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 4.7 presents the four best translation 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 seen from table 4.7, 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.

4.4 Enhanced N-gram-based SMT

The monotonous, basic n-gram SMT system has been enhanced with reordering strategies [Cre05c; Cre06a] and an additional Part-Of-Speech target n-gram language model feature. Although these enhancements are not part of this thesis work, they were included in the system which participated in the international evaluations enumerated at

the end of this section. A significant part of the thesis work has actually been dedicated to participation in international evaluations.

4.4.1 Constrained Reordered Search

A reordering strategy for the tuple-based SMT system consists in relaxing the tuple extraction constraint forcing tuples to monotonically generate the source and target sentences. Unfolding the tuples is now permitted, meaning that whereas the target-sentence order is still preserved, this restriction does not apply to the source sentence anymore [Cre05c]. Graphically, this is illustrated in Figure 4.4.

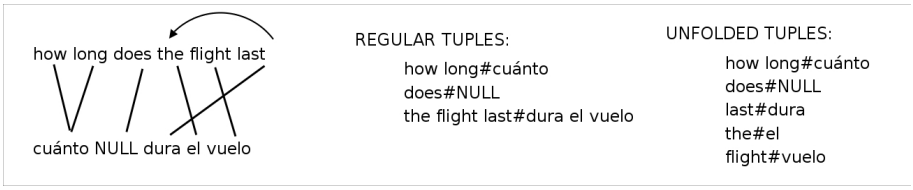


Figure 4.4: Differences between regular and unfolded tuple extraction.

Once this unfolded tuples are extracted, the same translation model is estimated, as well as additional features are. However, during decoding time, the decoder must allow for a reordered search. A distortion model, as described in equation 4.4, is included in the log-linear combination of features.

$$h_{DIST} = \sum_{k=1}^{K} d_k \tag{4.4}$$

where d_k is the distance between the first source word of the K^{th} tuple, and the last source word of the $K - 1^{th}$ tuple plus 1.

In order not to suffer from a computational complexity explosion, two parameters constrain this reordered search space:

- A distortion limit (μ): Any source word (phrase or tuple) is only allowed to be reordered if it does not exceed a distortion limit, measured in words.
- A reordering limit (ι): Any translation path is only allowed to perform ι reordering jumps.

4.4.2 Reordering Patterns

This reordering framework consists in using a set of automatically learnt rewrite rules to extend the monotonic search graph with reordering hypotheses [Cre06a].

A reordering pattern consists in the next rewrite rule:

$$\tau_1, \ldots, \tau_n \Rightarrow j_1, \ldots, j_n$$

where $\tau_1, \ldots, \tau_n$ is a sequence of POS tags (corresponding to a sequence of source words), and $j_1, \ldots, j_n$ indicates by which source word order the target words are monotonically generated.

Patterns are extracted in training from the crossed links found in word alignment or, in other words, found in translation tuples (no word within a tuple can be linked to a word out of it).

Once all rewrite patterns instances have been obtained, we compute a score for each pattern on the basis of relative frequency:

$$p(\tau_1, \ldots, \tau_n \Rightarrow j_1, \ldots, j_n) = \frac{N(\tau_1, \ldots, \tau_n \Rightarrow j_1, \ldots, j_n)}{N(\tau_1, \ldots, \tau_n)}$$

This score is used in training to prune out those patterns not achieving a threshold limit.

Starting from the monotonic graph, each sequence of input POS tags fulfulling a source-side rewrite rule implies the addition of a reordering arc (which encodes the reordering detailed in the target-side of the rule). Figure 4.5 shows how three rewrite rules applied over an input sentence extend the search graph given the reordering patterns that match the source POS tag sequence[3].

In the search, the decoder makes use of the whole set of models to score each reordering hypothesis, mainly driven by the n-gram translation model, as it has been estimated with reordered source words.

4.4.3 Target Part-Of-Speech Language Model

In order to tackle disagreement errors (detected in the error analysis of §4.3.3), a new feature was added to the log-linear combination, corresponding to the 5-gram language

[3]NC,CC and AQ stand respectively for name, conjunction and adjective.

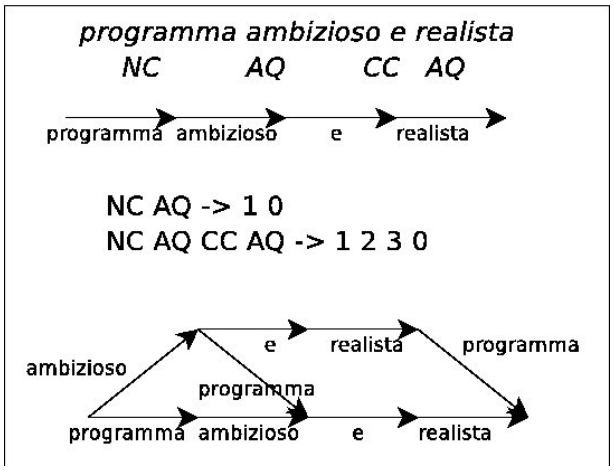


Figure 4.5: Search graph extension.

model of the target POS-tag sequence. This feature does not require POS-tagging of the output sentence, as the POS information is carried within the tuple.

Accordingly, the bilingual unit is redefined in terms of a triplet comprising the source word sequence, the target word sequence, and the Part-Of-Speech sequence representing that target word sequence. For simplicity, we only allow one single POS representation for each target word sequence (*and* given a fixed source word sequence).

Note that the POS information contained in the triplet is not actually used for computing the bilingual translation model probabilities, thus keeping the n-gram model unchanged. This information is only used during decoding, when the target POS language model is computed for each hypothesis and included in the log-linear combination with its own weight.

4.4.4 Participation in International Evaluations

With the enhanced system (and variants of it), we participated in a number of international evaluations:

- the second International Workshop on Spoken Language Translation (IWSLT 2005) [Eck05; Cre05a]
- the second evaluation of the TC-STAR project [Ney06; Mn06]
- the HLT/NAACL 2006 Workshop on Machine Translation [Koe06; Cre06b]

- the third International Workshop on Spoken Language Translation (IWSLT 2006) [Pau06; Cre06c]
- the third evaluation of the TC-STAR project [Mos]
- the fourth International Workshop on Spoken Language Translation (IWSLT 2007) [For07; Lam]

In all these evaluations, our system achieved competitive results. Result details can be checked from the corresponding cited papers.

4.5 Conclusions

As can be concluded from the presented results the tuple n-gram translation model, when used along with additional feature functions, provides state of the art translations for the considered translation directions. The impact of the lexicon models is remarkable, although probably a part of it is due to the lack of normalisation with respect to the length of one side of the tuple, which benefits to smaller tuples.

Another important result is that the quality of Spanish-to-English translations is significantly and consistently better than the one obtained in English-to-Spanish translations.

Additionally, four commonly-occurring types of translation errors were identified by reviewing a significant amount of translated sentence pairs.

A large part of this chapter's research has been co-authored with the members of UPC-TSC SMT team. Although the thesis author has participated in all of the chapter's issues (except when otherwise stated), the main contribution of this thesis work to the SMT system was to extend the basic system composed of the n-gram translation model and a target language model into a state-of-art translation system. To this end, the translation model was combined to other features within the maximum entropy framework proposed by [Och02]. Lexicon models were very successfully introduced as features in the system. The log-linear scaling factors optimisation process was also implemented. The n-gram-based machine translation system is described in the following journal publication:

- [Mar06b] José B. Mariño, Rafael Banchs, Josep M. Crego, Adrià de Gispert, Patrik Lambert, José A. R. Fonollosa, Marta R. Costa-jussà. 2006. **N-gram-based Machine Translation**. Computational Linguistics, 32 (4) pp. 527-549. MIT Press.

Chapter 5

Linguistic Classification and Multi-word Expression Grouping

In this chapter we report two types of experiments in which the training corpus is subject to some processing aiming at improving word alignment. After alignment is modified, the corpus is restored as it was initially. A translation system can then be trained from the original sentence pairs with the modified links. The chapter is organised as follows:

In §5.1, the impact on alignment quality and translation accuracy of linguistic classifications like lemmatising, stemming or verb classification is evaluated and studied. This work was conducted in cooperation with Adrià de Gispert (from UPC in Barcelona), Deepa Gupta (from ITC-irst in Trento), Maja Popović (from RWTH in Aachen). Only the main results are presented here. The complete study is available in a publication by de Gispert *et al.* [Gis] (alignment results) and de Gispert [Gis06] (translation results). The experiments were carried out on the Spanish-English European Parliament task, including an additional small-data corpus containing 1% of the whole Spanish–English parallel corpus.

In §5.2, the impact of grouping multi-word expressions before alignment is examined. Results for alignment and translation quality, for a small corpus (Verbmobil) and a large corpus (EPPS-04, see §4.3.1), are reported. Beside the size, an important difference between both corpora is the style of data: in Verbmobil, sentence pairs are formed by translations which closely mirror the original, realised to build a parallel corpus. In constrast, free translations are common in the EPPS corpus.

5.1 Linguistic Classification to Improve Word Alignment

With the goal of finding out which linguistic features are relevant for improving statistical word alignment, we followed a corpus transformation approach, ie. data was modified using morphosyntactic information before word alignment, as shown in the flow diagram in Figure 5.1.

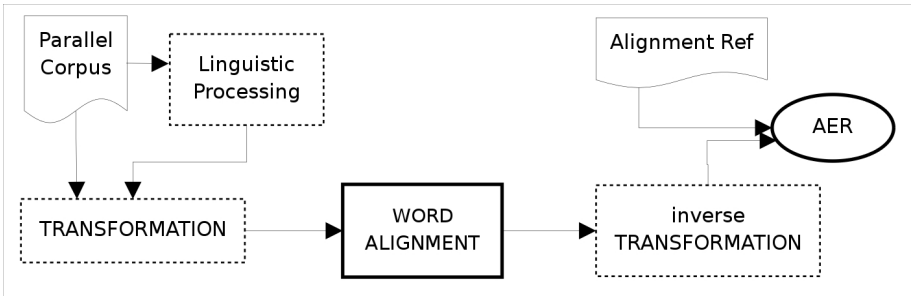


Figure 5.1: Experimental configuration to evaluate impact of using morphological information on word alignment.

Then, the obtained alignment of the transformed parallel corpus was mapped to the original sentence pairs in order to evaluate Alignment Error Rate against a manual reference. The same word alignment algorithm and configuration has been used in all cases, therefore acting as a black-box.

Two basic types of transformations have been considered, namely word classifications and word order modifications. Here we discuss the main results for the first type of transformations. Again, the complete study is available in the paper by de Gispert *et al.* [Gis].

5.1.1 Word Classifications

In general, word classifications aim at reducing data sparseness, by mapping some words to a unique token according to a certain criterion. In our case, criteria are based on the linguistic information provided by state-of-the-art language tools, in the particular case of processing the Spanish and English languages. In this chapter we consider the classification into base forms, for its importance in any morphosyntactic processing, as well as the two transformations which produced the best alignment improvement on the full corpus:

base forms Also known as lemmas, base forms lack details on morphological derivation of the word (gender, number, tense, and so on) and only provide information on the

head of the word. Therefore, they represent a meaning-bearing reduced version of each word, especially in the case of high morphological derivation, such as verbs, nouns or adjectives in Spanish. In English, verbs and nouns are also reduced by taking the base form, even though in lesser degree.

stems Same as lemmatisation, stemming is another method of word transformation which truncates inflected word forms into a single stem without morphological suffixes or derivations. However, a stemmer may not necessarily produce any meaning-bearing word form, whereas a lemmatiser returns the base form, usually associated with a dictionary citation of the given word form. A remarkable difference between stems and base forms occurs for irregular forms, like auxiliary verbs in English. "am", "are" and "is" remain unchanged with stemming, but have the same base form: "be".

Table 5.2 gives examples of stemming and lemmatisation results illustrating the differences between the two processes.

full verb forms Undoubtedly, given a verb meaning, tense and person, each language *implements* each verbal form independently from the other language. For example, whereas the personal pronoun is compulsory in English unless the subject is present, this does not occur in Spanish, where the morphology of the verb expresses the same aspect.

Therefore, aiming at simplifying the work for the word alignment, another word classification strategy can be devised to address the rich variety of verbal forms. For this, we group all words that build up a whole verbal form (including pronouns, auxiliary verbs and head verb) into the lemma of the head verb. This is a knowledge-based detection taken using deterministic automata implementing a few simple rules. These rules require information on word forms, POS-tags and lemmas in order to map the resulting expression to the lemma of the head verb, as done in [Gis05]. Examples of such mappings can be found in Table 5.1.

English		Spanish	
full form → lemma		full form → lemma	
has been found	find	introdujeran	introducir
we will find	find	han cometido	cometer
do you think	think	dijo	decir
offered	offer	está haciendo	hacer
I am doing	do	haremos	hacer

Table 5.1: Full verb forms are mapped to the lemma of the head.

English	Asian countries have followed our example too .
base forms	Asian country have follow our example too .
stems	asian countri have follow our exampl too .
full verbs	Asian countries V[follow] our example too .
Spanish	Los países asiáticos han seguido también nuestro ejemplo .
base forms	El país asiático haber seguir también nuestro ejemplo .
stems	los país asiátic han segu también nuestr ejempl .
full verbs	Los países asiáticos V[seguir] también nuestro ejemplo .

Table 5.2: Some English and Spanish corpus transformations as described in corresponding sections.

5.1.2 Experimental work

5.1.2.1 Experiment setup

Experiments were carried out using an update of the Spanish–English European Parliament parallel corpus described in §4.3.1, containing debates proceedings from 1996 to May 2005. Table 5.3 shows the main statistics of this parallel corpus, which will be referred to as EPPS-05. The lower part of the table shows the statistics for the 1% division used in the small data track.

	sent	words	vocab.	avg len
English	1.28 M	34.9 M	106 k	27.2
Spanish		36.6 M	153 k	28.5
English 1%	13.4 k	366 k	16.3 k	27.4
Spanish 1%		385 k	22.4 k	28.8

Table 5.3: EPPS-05 parallel corpus statistics for large and small data tracks.

In order to extract the linguistic information needed to perform the presented corpus modifications, data was preprocessed as follows:

- English POS-tagging using freely-available *TnT* tagger [Bra00].
- English lemmatisation using *wnmorph*, included in the WordNet package [Mil91].
- Spanish POS-tagging and lemmatisation using *FreeLing* analysis tool [Car04].
- English and Spanish stemming using the Snowball stemmer[1], which is based on Porter's algorithm.

[1] http://www.snowball.tartarus.org/

For evaluation, the manual reference described in §3.1.5 was used. This alignment test set is a subset of the training data, both in the large and the small data tracks.

5.1.2.2 Alignment results

As word alignment core algorithm (baseline), GIZA++ was used. Two baseline configurations are compared.

On the one hand, model iterations were set to $1^5H^5 4^3 4^3$ (meaning 5 iterations of IBM model 1, 5 iterations of HMM model and 3 iterations of IBM models 3 and 4) without using word classes and respecting original case. On the other hand, we used the $1^4H^5 4^4$ configuration (meaning 4 iterations of IBM model 1, 5 iterations of HMM model and 4 iterations of IBM model 4), included 50 word classes per language as estimated by 'mkcls', a freely-available tool along with GIZA++[2], and worked with lowercase text before aligning.

As it will be seen in alignment results, the latter strategy (denoted simply as 'baseline') always produced significantly lower AER results that its true-case no-class counterpart (denoted as 'baseline*'), which is shown as a means of comparison. For this reason, this better configuration applies for all experiments that have been done, except the one noted as baseline*.

	Eng→Spa			Spa→Eng			Union		
	R_S	P_P	AER	R_S	P_P	AER	R_S	P_P	AER
baseline*	59.97	75.05	33.09	59.11	78.16	32.31	69.33	67.65	31.56
baseline	63.10	77.11	30.34	64.12	80.21	28.38	73.37	69.43	28.77
base forms	66.37	83.50	25.75	68.06	83.72	24.69	73.93	75.01	25.51
stems	67.02	84.30	25.01	68.61	83.80	24.32	74.66	75.65	24.82
full verbs	66.50	79.72	27.13	65.44	81.30	27.10	73.96	71.36	27.45
verbs + stems	69.58	83.17	23.89	67.33	83.96	24.85	75.47	75.17	24.69

Table 5.4: Word Alignment results for small-data task.

Results with the 1% data set are shown in Table 5.4, where both directions and the symmetrisation through union are evaluated. Each row refers to each of the corpus transformations presented.

As it can be seen, both **base forms** and **stems** produce a very significant quality improvement, especially reflected in a more than 5 point absolute precision improvement in union alignment, whereas recall is also very high in these two cases for all alignment

[2]See http://www.fjoch.com for details on both tools.

directions. It looks like their classifications reduce sparseness and help the word alignment algorithm perform better. This improvement is best in the case of stems.

'**Full verb**' classification achieves a 1.5 AER reduction, basically thanks to an important recall increase in all alignment directions, due to the grouping effect of this classification, so that all words belonging to a verb form become linked to the same tokens.

Combining full verb classification and stemming (of the words outside verb forms) we obtain the best AER results.

	Eng→Spa			Spa→Eng			Union		
	R_S	P_P	AER	R_S	P_P	AER	R_S	P_P	AER
baseline*	69.13	88.81	21.94	67.25	90.04	22.60	73.98	84.41	20.92
baseline	73.20	90.78	18.65	72.18	92.17	18.64	78.42	86.43	17.56
base forms	72.80	91.70	18.54	71.84	93.17	18.50	76.73	87.90	17.82
stems	73.56	92.40	17.79	72.72	93.78	17.68	77.81	88.94	16.74
full verbs	74.27	90.77	17.85	73.03	93.31	17.56	78.60	87.37	16.97
verbs + stems	74.74	91.83	17.14	73.23	93.84	17.23	78.36	88.82	16.42

Table 5.5: Word Alignment results for large-data task.

Results with the full parallel corpus are shown in Table 5.5. Interestingly, conclusions regarding base forms and stems do not hold in this case. Whereas **base forms** are not useful anymore and even degrade alignment quality, **stems** still provide significant improvement in AER. This is expressed in a 2.5 point absolute precision increase at a cost of 0.6 recall decrease of union alignment. One possible reason for this is the harder classification of stems, especially for English, where initial vocabulary of 95K words is reduced to 81K with base forms and only 69K for stems (in Spanish, from baseline 138K vocabulary we end up with 78K base forms and 79K stems). Apparently, this involves a sparseness reduction, which makes word alignment more robust to non-literal translations. On the other hand, frequent words such as auxiliary verbs are not mapped to the same stem, thus possibly helping the aligner to discriminate compared to the case with base forms.

'**Full verb**' classification is still producing significant improvements, again reflected in the best recall figures for all alignment directions. This recall can countermeasure the recall loss when stemming and achieves the best AER (16.42) when combining these two approaches.

5.1.2.3 Discussion

Remarkably, and even though quality improvements due to morphological information are bigger in case of data scarceness, alignment error rate can be reduced by using this information even in case large amounts of data are available. Specifically, stemming and verb forms classification achieve significantly better recall and precision figures in all situations.

These experiments provide different alignment sets which can contain complementary information, so alignment quality can be further improved if they are combined. For the large data task, the best 3, 4 and 5 best union sets were combined with a consensus criterion. For each link present in at least one of the sets, if this link is present in a majority of sets, then it is selected for the combined set. Otherwise it is absent from the combined set. For the combination of an even number of sets, the criterion can be strict (more than half of the sets must agree) or weak (a half is enough).

	R_S	P_P	AER
3 best	78.50	90.04	15.79
4 best (weak)	80.29	87.35	16.10
4 best (strict)	76.51	92.59	15.87
5 best	78.37	89.70	16.07

Table 5.6: Combination, with a consensus criterion, of the best union alignment sets obtained in the large data task (in order: the verbs+stems, stems, full verbs, spa adj base and baseline sets).

Results are shown in Table 5.6. While all combinations improve the best AER presented in Table 5.5 (that of the verbs+stems experiment), the combination of best 3 sets is particularly interesting since both recall and precision are also improved. In the 4 sets combinations, the weak criterion gives a high recall and lower precision combination, whereas the strict criterion gives a high precision but lower recall combination.

5.1.3 Correlation with SMT quality

In order to see the impact of the alignment improvement on translation quality, translation experiments were carried out by de Gispert [Gis06] by comparing 5 selected alignment configurations, ranging from worst to best AER for both large and small data tracks. Since this is very relevant to the present thesis, we report in this section his main results and conclusions.

Small data track

Results for both translation directions in the small data track are shown in Table 5.7, where the result when translating only with the Bilingual Model (onlyBM) and the full log-linear combination (full) are shown.

	AER	Eng→Spa				Spa→Eng			
		onlyBM		full		onlyBM		full	
		BLEU	NIST	BLEU	NIST	BLEU	NIST	BLEU	NIST
baseline*	31.56	0.2815	7.450	0.3198	7.922	0.3685	8.670	0.4159	9.163
baseline	28.77	0.2824	7.453	0.3215	7.943	0.3733	8.734	0.4209	9.204
full verbs	27.45	0.2884	7.507	0.3251	8.015	0.3651	8.603	0.4218	9.249
stems	24.82	0.2859	7.483	0.3254	8.031	0.3719	8.717	0.4283	9.319
verbs + stems	24.69	0.2897	7.491	0.3290	8.048	0.3597	8.567	0.4190	9.229

Table 5.7: Translation scores for small-data task.

At first glance, we can already conclude that strong variations in AER do not end up producing a strong variation in translation quality. While AER shows a nearly 22% relative decrease from worst to best alignment (about 6.87 points absolute), BLEU experiences an increase of at most 3% relative (Eng→Spa) and 4% relative (Spa→Eng). In Eng→Spa, biggest BLEU difference is 0.009 absolute (full system) whereas in Spa→Eng, it is about 0.014 absolute.

According to a 95% confidence level for this task, BLEU measures may have a variation of around ±0.015 in Eng→Spa and around ±0.012 in Spa→Eng. Therefore, the impact of improved AER figures in the Eng→Spa translation direction is below the 95% BLEU confidence level, even though there seems to be a tendency to positively correlate with AER in this task.

Opposite to that, in the Spa→Eng direction, BLEU variations do achieve the confidence level, though minimally. Unfortunately, in this case correlation with AER is unclear. Particularly, stems and baseline achieve pretty similar results in the onlyBM configuration even though they present a nearly 4-point absolute AER difference. In addition to that, the use of full verb forms for alignment, which always generates a lower-AER alignment solution, does not help and even harm translation performance when comparing 'stems' versus 'verbs+stems'.

Large data track

Results for the large data track are shown in Table 5.8. Again, the relatively big AER difference between both baselines (from 20.92 to 17.56, a 16% relative decrease) does not

yield any significant change in translation performance in any translation direction, as shown in the first two rows.

	AER	Eng→Spa				Spa→Eng			
		onlyBM		full		onlyBM		full	
		BLEU	NIST	BLEU	NIST	BLEU	NIST	BLEU	NIST
baseline*	20.92	0.4368	9.345	0.4794	9.887	0.4782	9.963	0.5519	10.797
baseline	17.56	0.4366	9.331	0.4802	9.909	0.4769	9.919	0.5526	10.763
full verbs	16.97	0.4293	9.219	0.4790	9.922	0.4728	9.871	0.5514	10.779
stems	16.74	0.4284	9.223	0.4787	9.883	0.4760	9.902	0.5553	10.788
verbs + stems	16.42	0.4264	9.187	0.4785	9.889	0.4748	9.882	0.5525	10.765

Table 5.8: Translation scores for large-data task.

Surprisingly, further improvements in AER achieved by 'full verbs', 'stems' and 'verbs+stems' configurations do not improve performance, but even produce worse translation results, as can be especially observed when only the bilingual translation model is used. In the case of full log-linear combination, BLEU differences are reduced to 0.4% relative (Eng→Spa) and 0.7% relative (Spa→Eng), or 0.0017 and 0.0039 absolute, far away from significance thresholds.

A careful study of this unexpected behaviour of the bilingual model has been carried out by de Gispert [Gis06].

5.2 Grouping Multi-word Expressions in Alignment

Despite the change from a word-based to a phrase-based translation approach, word to word approaches for inferring alignment models from bilingual data [Vog96; Och03c] continue to be widely used (see also §2.4.3).

On the other hand, from observing bilingual data sets, it becomes evident that in some cases it is just impossible to perform a word to word alignment between two phrases that are translations of each other. For example, certain combination of words might convey a meaning which is somehow independent from the words it contains. This is the case of bilingual pairs such as "fire engine" and "camión de bomberos".

Notice that a word-to-word alignment strategy would most probably[3] provide the following Viterbi alignments for words contained in the previous example: "camión:truck", "bomberos:firefighters", "fuego:fire", and "máquina:engine".

Of course, it cannot be concluded from these examples that a SMT system which uses a word to word alignment strategy will not be able to handle properly the kind of word expression described above. This is because there are other models and feature functions involved which can actually *help* the SMT system to get the right translation.

However these ideas motivate for exploring alternatives for using multi-word expression information in order to improve alignment quality and consequently translation accuracy. In this sense, our idea of a multi-word expression (hereafter MWE) refers in principle to word sequences which cannot be translated literally word-to-word. However, the automatic technique studied in this work for extracting and identifying MWEs does not necessarily follow this definition rigorously.

Experiments were performed with two type of data: a small corpus composed of a selection of spontaneous speech databases available from the Verbmobil project[4] and a large corpus consisting of parliamentary session transcriptions of the European Parliament (EPPS-04).

In §5.2.1, the method to extract and identify MWEs is presented. For the EPPS task, a few very basic rules based on part-of-speech have also been added to filter out noisy entries in the dictionary, and bilingual MWEs have been classified into three categories (nouns, verbs and others). In §5.2.2 and §5.2.3, results for the Verbmobil and the EPPS tasks are reported.

[3] Of course, alignment results strongly depends on corpus statistics.
[4] http://verbmobil.dfki.de/verbmobil

5.2.1 Experimental Procedure

In this section we describe the technique used to see the effect on the translation model of grouping multi-word expressions in alignment.

First, bilingual MWEs (BMWE) were automatically extracted from the parallel training corpus and the most relevant ones were stored in a dictionary. More details on this stage of the process are given in §5.2.1.1 and §5.2.1.2. In a second stage, BMWE present in the dictionary were detected in the training corpus in order to modify the word alignment (see §5.2.1.3 and 5.2.1.4 for more details). Every word of the source side of the BMWE was linked to every word of the target side.

Then the source words and target words of each detected BMWE were grouped in a unique "super-token" and this modified training corpus was aligned again with GIZA++. By grouping multi-words, we increased the size of the vocabulary and thus the sparseness of data. However, we expect that if the meaning of the multi-words expressions we grouped is effectively different from the meaning of the words it contains, the individual word probabilities should be improved. After re-aligning, we unjoined the super-tokens that had been grouped in the previous stage, correcting the alignment set accordingly. More precisely, if two super-tokens A and B were linked together, after ungrouping them into various tokens, every word of A was linked to every word of B. Note that even after re-aligning, the vocabulary and the sequence of words to train the translation model were the same as in the baseline model, since we unjoined the super-tokens. The difference comes from the alignment, and thus from the translation units and from their corresponding n-grams.

5.2.1.1 Bilingual Multi-words Extraction

Various methods to extract BMWE were experimented.

Asymmetry Based Extraction

Multi-word expressions were extracted with the method proposed by Lambert and Castell [Lam04]. This method is based on word-to-word alignments which are different in the source-target and target-source directions. Such alignments can be produced with the IBM Translation Models (see §2.1.1). We used GIZA++, which implements these models, to perform word-to-word alignments in both directions, source-target and target-source. Multi-words like idiomatic expressions or collocations can typically not be aligned word-to-word, and cause an asymmetry in the (source-target and target-source) alignment sets.

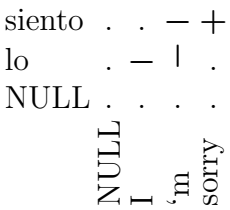


Figure 5.2: Asymmetry in the word-to-word alignments of an idiomatic expression. Source-target and target-source links are represented respectively by horizontal and vertical dashes.

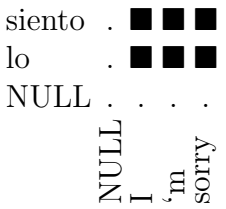


Figure 5.3: A multi-word expression has been detected in the asymmetry depicted in figure 5.2 and aligned as a group. Each word of the source side is linked to each word of the target side.

An asymmetry in the alignment sets is a subset where source-target and target-source links are different. An example is depicted in figure 5.2. A word does not belong to an asymmetry if it is linked to exactly one word, which is linked to the same word and is not linked to any other word. In the method proposed by Lambert and Castell, asymmetries in the training corpus are detected and stored as bilingual multi-words, along with their number of occurrences.

These asymmetries can be originated by idiomatic expressions, but also by translation errors or omissions. The method relies on the idea that if the asymmetry is caused by a language feature, it will be repeated various times in the corpus, otherwise it will occur only once. Thus only those bilingual multi-words which appeared at least twice are selected. Still, some bilingual multi-words, whose source side is not the translation of the target side, can appear various times. An example is "de que - you". To minimise this type of errors, we wanted to be able to select the N best asymmetry based BMWE, and ranked them according to their number of occurences.

Bilingual Phrase Extraction

Here we refer to *Bilingual Phrase* (BP) as the bilingual phrases used by Och and Ney [Och04b], described in §2.1.2.1. We extracted all BP of length up to three words. Again, we

established a ranking between them. In that purpose, we estimated the phrase translation probability distribution by relative frequency:

$$p(\tilde{t}|\tilde{s}) = \frac{N(\tilde{s},\tilde{t})}{N(\tilde{s})} \tag{5.1}$$

In equation 5.1, $\tilde{s}$ and $\tilde{t}$ stand for the source and target side of the BP, respectively. $N(\tilde{s},\tilde{t})$ is the number of times the phrase $\tilde{s}$ is translated by $\tilde{t}$, and $N(\tilde{s})$ is the number of times $\tilde{s}$ occurs in the corpus. Data sparseness can cause probabilities estimated in this way to be overestimated, and the inverse probability ($p(\tilde{s}|\tilde{t})$) has proved to contribute to a better estimation [Rui05]. To increase reliability, we took the minimum of both relative frequencies as probability of a BP, as shown in equation 5.2:

$$p(\tilde{s},\tilde{t}) = \min(p(\tilde{t}|\tilde{s}), p(\tilde{s}|\tilde{t})) \tag{5.2}$$

Many phrases occur very few times but always appear as the translation of the same phrase in the other language, so that their mutual probability as given by equation 5.2 is 1. However, this does not necessarily imply that they are a good translation of each other. To avoid to give a high score to these entries, we took as final score the minimum of the relative frequencies multiplied by the number of occurrences of this phrase pair in the whole corpus.

Intersection

Taking the intersection between the asymmetry based multi-word expressions and the BP presents the following advantages:

- BP imply a stronger constraint on the alignment between source and target side than asymmetries. In particular, entries which appear various times and whose source and target sides are not aligned together can't be selected as bilingual phrases and disappear from the intersection.
- Statistics of the BP set, which come from counting occurrences in the whole corpus, are more reliable than the statistics which come from counting occurrences in alignment asymmetries only. Thus, scoring asymmetry based BMWE with the BP statistics should be more reliable than with the number of occurrences in alignment asymmetries.
- Finally, since BP are extracted from all parts of the alignment (and not in asymmetries only), most BP are not BMWE but word sequences that can be decomposed

word to word. For example, in the 10 best BP, we find "una reunión - a meeting", which is naturally aligned word to word. So if we want to use a BMWE dictionary of N entries (i.e. the N best scored), in the case of BP this dictionary would contain, let's say, only a 60% of actual BMWE. In the case of the asymmetry based multi-words, it would contain a much higher percentage of actual BMWE, which are the only "usefull" entries for our purpose.

So we performed the intersection between the entire BP set and the entire asymmetry based multi-words set, keeping BP scores.

Extraction Method Evaluation

To compare these three methods, we evaluated for the Verbmobil task (see §5.2.2.1) the links corresponding to the BMWE grouped in the detection process, with the manual alignment reference. Table 5.9 shows the precision and recall (as defined in §2.4.1) for the multi-words detected in the corpus when the three different dictionaries where used. Here the computed links are only those of the BMWE detected. However the reference links are not restricted to multi-words. So the recall gives the proportion of detected multi-words links in the total set of links. In all three cases, only the best 650 entries of the dictionary were used.

We see from table 5.9 that taking the intersection with the BP set allows a nearly 6% improvement in precision with respect to the asymmetry based BMWE. The best precision is reached with the BP dictionary, which suggests than a better precision could be obtained for the intersection, for instance establishing a threshold pruning condition. Note that using a (manually built) dictionary of idiomatic expressions and with verb phrases detected with (manually specified) rules, de Gispert *et al.* [Gis04b] achieved a much higher precision.

Recall scores reflect in a way the number of actual BMWE present in the 650 entries of the dictionary, and how frequent they are in the alignment asymmetries, which are where the BMWE are searched (see §5.2.1.3). Logically, the asymmetry based dictionary, ranked according to the occurrence number, has got the higher recall. As explained in §5.2.1.1, many high score BP are not multi-words expressions. So in the particular 650 entries we selected, there are less BMWE than in the intersection and the asymmetry based selections, and the recall is much lower. Thus the impact of multi-words information is expected to be lower.

Finally, the intersection dictionary allows to detect BMWE with a high precision and a high recall (compared to the two other methods), so it is the dictionary we used.

	Precision	Recall
Asymmetry based	85.39	20.21
Bilingual phrases	92.98	13.41
Intersection	91.26	18.82

Table 5.9: Multi-word expressions quality.

5.2.1.2 Lexical and Morpho-syntactic Filters

In the EPPS task, multi-word expressions quality is much lower than in the Verbmobil one, so we had to implement filters to keep only the best entries.

In English and Spanish, a list of stop words[5] (respectively 19 and 26) was established. The BMWE dictionary was also processed by a Part-Of-Speech (POS) tagger and eight rules were written to filter out noisy entries. These rules depend on the tag set used. Examples of criteria to reject a BMWE include:

- Its source or target side only contains stop words
- Its source or target side ends with a coordination conjunction
- Its source or target side begins with a coordination conjunction (except "nor", in English)
- Its source or target side ends with an indefinite determiner

English data have been POS-tagged using the *TnT* tagger [Bra00], after the lemmas have been extracted with *wnmorph*, included in the Wordnet package [Mil91]. POS-tagging for Spanish has been performed using the *FreeLing* analysis tool [Car04].

Finally, the BMWE set has been divided in three subsets, according to the following criteria, applied in this order:

- If source AND target sides of a BMWE contain at least a verb, it is assigned to the "verb" class.
- If source AND target sides of a BMWE contain at least a noun, it is assigned to the "noun" class.
- Otherwise, it is assigned to the "misc" class (miscellaneous). Note that this class is mainly composed of adverbial phrases.

[5] frequently occurring, semantically insignificant words like "in", "of", "on".

5.2.1.3 Multi-Words Identification

Identification consists, first, of the detection of all possible BMWE(s) in the corpus, and second, of the selection of the relevant candidates.

The detection part simply means matching the entries of the dictionaries described in the previous subsections. In the example of figure 5.2, the following BMWEs would have been detected (the number on the right is the score):

```
i am sorry ||| lo siento ||| 1566
am sorry ||| siento ||| 890
it is ||| es ||| 1004407
it is ||| esto es ||| 269
true ||| es verdad ||| 63
```

Then, selection in a sentence pair runs as follows. First, the BMWE with highest score among the possible candidates is considered and its corresponding positions are set as covered. If this BMWE satisfies the selection criterion, the corresponding words in the source and target sentences are grouped as a unique token. This process is repeated until all word positions are covered in the sentence pair, or until no BMWE matches the positions remaining to cover.

The selection criterion rejects candidates whose words are linked to exactly one word. Thus in the example, "esto – this is" would not be selected. This is correct, because the subject "esto" (this) of the verb "es" (is) in Spanish is not omitted, so that "this is – es" does not act as BMWE ("esto" should be translated to "this" and "is" to "es").

At the end of the identification process the sentence pair of figure 5.2 would be the following: "lo_siento ; esto es verdad – I_'m_sorry , this is true".

In order to increase the recall, BMWE detection was insensitive to the case of the first letter of each multi-word. The detection engine also allows a search based on lemmas. Two strategies are possible. In the first one, search is first carried out with full forms, so that lemmas are resorted to only if no match is found with full forms. In the second strategy, only lemmas are considered.

5.2.1.4 Re-alignment

The modified training corpus, with identified BMWEs grouped in a unique "super-token" was aligned again in the same way as previously. By grouping multi-words, we increased the size of the vocabulary and thus the sparseness of data. However, we expect that if

the meaning of the multi-words expressions we grouped is effectively different from the meaning of the words they contain, the individual word probabilities should be improved.

After re-aligning, we unjoined the super-tokens that had been grouped in the previous stage, correcting the alignment set accordingly. More precisely, if two super-tokens A and B were linked together, after ungrouping them into various tokens, every word of A was linked to every word of B. Translation units were extracted from this corrected alignment, with the unjoined sentence pairs (*i.e.*the same as in the baseline). So the only difference with respect to the baseline lied in the alignment, and thus in the distribution of translation units and in lexical model probabilities.

5.2.2 Experimental Results for the Verbmobil Task

5.2.2.1 Training and Test Data

Training and test data come from a selection of spontaneous speech databases available from the Verbmobil project[6]. The databases have been selected to contain only recordings in US-English and to focus on the appointment scheduling domain. Then their counterparts in Catalan and Spanish have been generated by means of human translation [Arr03]. Dates and times were categorised automatically (and revised manually). A test corpus of 2059 sentences has been separated for the training corpus.

The alignment reference corpus consists of 400 sentence pairs manually aligned by a single annotator, with no distinction between ambiguous or unambiguous links, i.e. with only one type of links.

Some statistics of the data are displayed in table 5.10.

		Spanish	English
Training	Sentences	28000	
	Words	201893	209653
	Vocabulary	4894	3167
Test	Sentences	2059	
	Words	19696	20585
Align. Ref.	Sentences	400	
	Words	3124	3188

Table 5.10: Characteristics of Verbmobil corpus: training and translation test as well as alignment reference.

[6]http://verbmobil.dfki.de/verbmobil

5.2.2.2 Alignment and Translation Results

The effect of grouping multi-words before aligning the corpus is shown in figures 5.4 and 5.5, and in tables 5.11 and 5.12.

Figure 5.4 represents the Alignment Error Rate (AER, Equation 2.18) versus the number of times BMWE are grouped during our process. Nevertheless, here the whole set of links is evaluated, not only the links restricted to grouped BMWE.

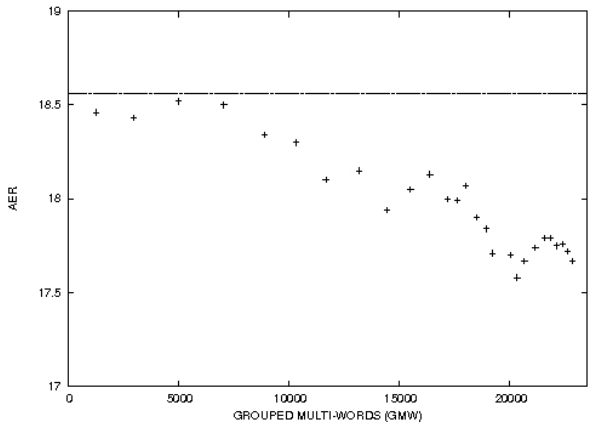


Figure 5.4: Alignment Error Rate versus the number of multi-words grouped (GMW).

In figure 5.4, the horizontal line represents the AER of the baseline system. We see a clear tendency in lowering the AER while more multi-words are grouped, although the total improvement is slightly less than one percent AER. An analysis of the precision and recall curves (not shown here) reveals that this AER improvement is due to a constant recall increase without precision loss.

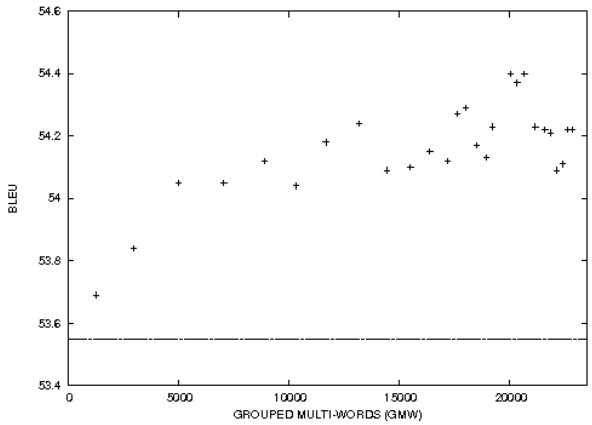


Figure 5.5: BLEU score versus the number of multi-words grouped (GMW), in the translation from Spanish to English.

In figure 5.5, BLEU score for the Spanish to English translation is plotted against

the number of multi-words grouped. The horizontal line is the baseline value. Again, it is clear that while more multi-words are grouped, the translation quality is improved, the overall effect being of 0.85% in absolute BLEU score. However, there is an inflexion point (occuring around 22000 GMW, which corresponds to a dictionary of 1000 BMWE entries), after which there is a saturation and even a decrease of BLEU score. This inflexion could be caused by the lower quality of the worst ranked BMWE entries in the dictionary.

Tables 5.11 and 5.12 show experimental results obtained for a size of the BMWE dictionary of 650 entries. In this case, 20671 multi-words were grouped before re-aligning.

As can be seen in table 5.11, the size of the Spanish and English vocabularies were increased respectively a 5% and 13%, while the number of running words was decreased respectiveley a 3.7% and 10.5%.

	Voc. Size		Running Words	
	Spa	Eng	Spa	Eng
Baseline	4851	3139	198245	207743
650 MW	5089	3551	190730	186066

Table 5.11: Effect on the vocabulary size and the number of running words with a dictionary of 650 bilingual multi-words.

		S→E		E→S	
	AER	WER	BLEU	WER	BLEU
Bas.	18.6	30.0	53.5	35.2	48.2
Sym.	18.0	30.0	53.7	35.1	48.3
650	17.7	29.6	54.4	34.8	48.5

Table 5.12: Effect on Alignment and Translation with a dictionary of 650 bilingual multi-words.

Table 5.12 shows the alignment and translation results. "Bas." stands for the baseline system, and "Sym." is the system trained with the alignment set calculated after symmetrising (with a dictionary of 650 BMWE), but before grouping and re-aligning. Because the other systems are trained on the union of alignment sets in both directions, for this particular result, when no multi-word matched uncovered positions, the union was taken instead of the intersection (see section 5.2.1.3). "650" stands for the system obtained with the dictionary of 650 BMWE entries. Results are shown for both translation directions, Spanish to English (S→E) and English to Spanish (E→S). First, it can be observed that the symmetrising process doesn't permit to improve significantly translation results. So the effect is due to the grouping of multi-word expressions and the improvement of

individual word alignment probabilities it implies. Secondly, the effect is smaller when translating from English to Spanish than in the other direction.

Linear Regressions and Significance Analysis

In order to study in more detail the incidence of the proposed multi-word extraction technique on both alignment quality and translation accuracy, linear regressions were computed among some variables of interest. This analysis allows to determine if the variations observed in AER, WER and BLEU are actually due to variations in the number of BMWE used during the alignment procedure; or, on the other hand, if such variations are just random noise.

We were actually interested in checking for two effects:

- the incidence of the total number of bilingual multi-words grouped in the training corpus (GMW) on the resulting quality measurement variations (AER, WER and BLEU), and
- the incidence of alignment quality variations (AER) on translation accuracy variations (WER and BLEU).

A total of nine regression analysis, which are defined in Table 5.13, were required to evaluate the mentioned effects. More specifically, Table 5.13 presents the translation direction, a reference number, and the independent and dependent variables considered for each of the nine regressions. For all regression analysis, only variable values corresponding to a maximum of 900 BMWE entries were considered. As seen from figure 5.5 the behaviour of variables changes drastically when more that 1000 BMWE entries (around 22000 GMW) in the dictionary are considered.

Table 5.14 presents the regression coefficients obtained, as well as the linear correlation coefficients and the significance test results, for each of the considered regressions.

From the significance analysis results presented in Table 5.14, it is observed that all regressions performed can be considered statistically significant; i.e. the probabilities for such value distributions occurring by pure chance are extremely low.

These results allow us to conclude that the proposed technique for extracting and using multi-word expressions has a positive incidence on both alignment quality and translation accuracy. However, as can be verified from slope values (β_1) presented in Table 5.14, this incidence is actually small. Although increasing the number of multi-words reduces AER and WER, and increases the BLEU, the absolute gains are lower than we expected.

Dir.	Ref.	Dependent variable	Independent variable
–	reg1	AER	GMW
S → E	reg2	BLEU	GMW
	reg3	WER	GMW
	reg4	BLEU	AER
	reg5	WER	AER
E → S	reg6	BLEU	GMW
	reg7	WER	GMW
	reg8	BLEU	AER
	reg9	WER	AER

Table 5.13: Linear regressions performed.

	β_1	β_0	ρ	F	p-value
reg1	-0.04	18.7	-0.93	159.9	$0.00\ 10^{-5}$
reg2	0.02	53.8	0.85	58.69	$0.01\ 10^{-5}$
reg3	-0.01	30.0	-0.84	53.31	$0.02\ 10^{-5}$
reg4	-0.45	62.3	-0.79	37.95	$0.28\ 10^{-5}$
reg5	0.31	24.2	0.84	54.18	$0.02\ 10^{-5}$
reg6	0.02	48.0	0.88	79.80	$0.00\ 10^{-5}$
reg7	-0.03	35.5	-0.89	91.64	$0.00\ 10^{-5}$
reg8	-0.45	57.3	-0.81	45.04	$0.08\ 10^{-5}$
reg9	0.62	23.7	0.82	49.43	$0.04\ 10^{-5}$

Table 5.14: Regression coefficients (β_1: slope, and β_0: intercept), linear correlation coefficients (ρ) and significance analysis results for the regression coefficients (F-test). In this table GMW unit was 1000 GMW.

5.2.3 Experimental Results for the EPPS Task

The translation and alignment data are those indicated in §4.3.1 (EPPS-04).

First, we describe the results of the BMWE extraction and detection techniques applied to the training data. Then we present results in alignment and translation.

5.2.3.1 Description of the BMWE dictionaries

Parameters of the extraction process have been optimised with the alignment development corpus available with the alignment test corpus. With these parameters, a dictionary of 60k entries was extracted. After applying the lexical and morpho-syntactic filters, 45k entries were left. The best 30k entries (hereinafter referred to as *all*) have been selected

for the experiments and divided in the three groups mentioned in section 5.2.1.2. *verb*, *noun* and *misc* (miscellaneous) dictionaries contained respectively 11797, 9709 and 8494 entries.

Table 5.15 shows recall (R_S) and precision (P_P) for the BMWEs identified with each dictionary. The first line is the evaluation of the MWEs obtained with the best 30k entries of the dictionary before filtering. Alignments evaluated in table 5.15 contained only links corresponding to the identified BMWEs. For an identified BMWE, a link was introduced between each word of the source side and each word of the target side. Nevertheless, the test data contained the whole set of links.

	Recall	Precision
Best 30k (no filters)	13.6	53.6
Best 30k (filters)	11.4	79.3
VERB (filters)	3.7	81.8
NOUN (filters)	4.0	72.8
MISC (filters)	4.1	80.8

Table 5.15: Quality of the BMWEs identified from the various dictionaries.

From table 5.15 we see the dramatic effect of the filters. The precision for nouns is lower than for the other categories because many word groups which were identified, like "European Parliament - Parlamento europeo", are not aligned as a group in the alignment reference. Notice also that the data in table 5.15 reflects the precision of bilingual MWE, which is a lower bound of the precision of "super-tokens" formed in each sentence, the quantity that matters in our experiment.

Identification of BMWE based on lemmas has also been experimented. However, with lemmas, the selection phase is more delicate. With our basic selection criterion (see section 5.2.1.3), the quality of MWEs identified was worse so we based identification on full forms.

Figure 5.6 shows the first 10 entries in the *misc* dictionary, along with their renormalised score. Notice that "the EU - la UE", "young people - jóvenes" and "the WTO - la OMC" have been incorrectly classified due to POS-tagging errors.

5.2.3.2 BMWE Identification Statistics

Table 5.16 shows, for each language, the MWE vocabulary size after the identification process, and how many times a MWE has been grouped as a unique token (instances). The different number of instances between Spanish and English correspond to one-to-many BMWEs. In general more MWEs are grouped in the Spanish side, because English

```
the EU ||| la UE ||| 770731
secondly ||| en segundo lugar ||| 610599
however ||| sin embargo ||| 443042
finally ||| por último ||| 421879
firstly ||| en primer lugar ||| 324396
thirdly ||| en tercer lugar ||| 286924
young people ||| jóvenes ||| 178571
the WTO ||| la OMC ||| 174496
once again ||| una vez más ||| 169317
once ||| una vez ||| 150139
```

Figure 5.6: Examples of BMWEs of the *misc* category.

is a denser language. However, the omission of the subject in Spanish causes the inverse situation for verbs.

	Vocabulary		Instances	
	ENG	SPA	ENG	SPA
ALL	12.2k	12.6k	1.28M	1.56M
VERB	6.0k	3.3k	738k	237k
NOUN	3.9k	5.9k	288k	827k
MISC	3.1k	4.3k	336k	557k

Table 5.16: Statistics for the BMWEs identified from the various dictionaries. ALL refers to the 30k best entries with filters.

5.2.3.3 Alignment and Translation Results

Tables 5.17 and 5.18 show the effect of aligning the corpus when the various categories of multi-words have been previously grouped.

IBM1 lexical probabilities	baseline	All
p(in_other_words|es_decir)	-	0.94
p(words|decir)	0.23	0.0013
p(other|decir)	0.026	$6\ 10^{-5}$
p(say|decir)	0.45	0.49

Table 5.17: Single word lexical probabilities of the alignment model in the baseline and after grouping MWE with all dictionary entries. The multi-word tokens "in_other_words" and "es_decir" do not exist in the baseline.

In table 5.17 we see how word-to-word lexical probabilities of the alignment model can be favourably modified. In the baseline, due to presence of the fixed expression "in

other words - es decir", the probability of "words" given "decir" ("say" in English) is high. With this expression grouped, probabilities p(words|decir) and p(other|decir) vanish, while p(say|decir) is reinforced. These observations allowed to expect that with many individual probabilities improved, a global improvement of the alignment would occur.

However, table 5.18 shows that alignment is not better when trained with BMWEs grouped as a unique token.

	Recall	Precision	AER
Baseline	76.3	85.0	19.4
All	78.0	82.0	19.9
Verb	77.0	84.5	19.3
Noun	76.8	83.0	20.0
Misc	77.0	84.1	19.4

Table 5.18: Alignment results

A closer insight into alignments confirms that they have not been improved globally. Changes with respect to the baseline are very localised and correspond directly to the grouping of the BMWEs present in each sentence pair.

Table 5.19 presents the automatic translation evaluation results. In the Spanish to English direction, BMWEs seem to have a negative influence. In the English to Spanish direction, no significant improvement or worsening is observed.

	S→E		E→S	
	mWER	BLEU	mWER	BLEU
Baseline	**34.4**	**0.547**	**40.2**	**0.472**
All	36.4	0.517	40.7	0.470
Verb	35.1	0.537	**40.2**	**0.472**
Noun	35.1	0.537	40.7	0.469
Misc	35.8	0.527	41.1	0.466

Table 5.19: Translation results in Spanish-to-English (S→E) and English-to-Spanish (E→S) directions.

In order to understand these results better, we performed a manual error analysis for the first 50 sentences of the test corpus. We analysed, for the experiment with all dictionary entries ("All" line of table 5.19), the changes in translation with respect to the baseline. We counted how many changes had a neutral, positive or negative effect on

translation quality. Results are shown in table 5.20. Notice that approximately half of these changes were directly related to the presence some BMWE.

This study permitted to see interesting qualitative features. First, BMWEs have a clear influence on translation, sometimes positive and sometimes negative, with a balance which appears to be null in this experiment. In many examples BMWEs allowed a group translation instead of an incorrect word to word literal translation. For instance, "Red Crescent" was translated by "Media Luna Roja" instead of "Cruz Luna" (cross moon).

Two main types of error were observed. The first ones are related to the quality of BMWEs. Determiners, or particles like "of", which are present in BMWEs are mistakenly inserted in the translations. Some errors are caused by inadequate BMWEs. For example "looking at – si analizamos" ("if we analyse") cannot be used in the sense of looking with the eyes. The second type of error is related to the rigidity and data sparseness introduced in the bilingual n-gram model. For example, when inflected forms are encapsulated in a BMWE, the model looses flexibility to translate the correct inflection. Another typical error is caused by the use of back-off (n-1)-grams in the bilingual language model, when the n-gram is not any more available because of increased data sparseness.

The error analysis did not give explanation for why the effect of BMWEs is so different for different translation directions. A possible hypothesis would be that BMWEs help in translating from a denser language. However, in this case, verbs would be expected to help relatively more in the Spanish to English direction, since there are more verb group instances in the English side.

	Neutral	Positive	Negative
S→E	43	20	22
E→S	49	19	17

Table 5.20: Effect on quality of differences in the translations between the baseline and the BMWE experiment with "ALL" dictionary. S and E stand for Spanish and English, respectively.

5.3 Conclusions and Further Work

In this chapter we discussed ways of improving word alignment by two types of corpus processing: linguistic classification and, reversely, grouping of multi-word expressions. The impact of this modified alignment on translation quality has also been evaluated.

Even with a large corpus, stemming and verb classification did clearly improve the

word alignment. However this improvement had no positive impact on translation quality.

We applied a technique for extracting and using BMWEs in Statistical Machine Translation. This technique is based on grouping BMWEs before performing statistical alignment. Although this technique yielded alignment and translation quality improvements on the Verbmobil corpus, it failed to clearly improve alignment quality or translation accuracy on a large corpus with real-life data.

After performing a detailed error analysis, we believe that when the considered MWEs are fixed expressions, grouping them before training helps for their correct translation in test. However, grouping MWEs which could in fact be translated word to word, doesn't help and introduces unnecessary rigidity and data sparseness in the models. Some errors were also caused by noise in the automatic generation of BMWEs. Thus filtering techniques should be improved, and different methods for extracting and identifying MWEs must be developed and evaluated. Resources build manually, like Wordnet multi-word expressions, should also be considered.

The proposed method considers the bilingual multi-words as units ; the use of each side of the BMWEs as independent monolingual multi-words must be considered and evaluated.

Some research work reported in this chapter was published in the following contributions:

- [Gis] A. de Gispert, D. Gupta, M. Popovic, P. Lambert, J. B. Mariño, M. Federico, H. Ney and R. Banchs, **Improving Statistical Word Alignments with Morpho-syntactic Transformations**, in *Proceedings of 5th International Conference on Natural Language Processing, FinTAL'06*, Springer Verlag, LNCS, pps. 368–379, August 2006.
- [Lam05a] Patrik Lambert and Rafael Banchs. 2005. **Data Inferred Multi-word Expressions for Statistical Machine Translation**. Proc. of Machine Translation Summit X, pp. 396-403. Phuket, Thailand.
- [Lam06a] Patrik Lambert and Rafael Banchs. 2006. **Grouping Multi-word Expressions According to Part-Of-Speech in Statistical Machine Translation**. Proc. of the EACL Workshop on Multi-Word-Expressions in a Multilingual Context, pp. 9-16. Trento, Italy, April 3rd.

Chapter 6

Coefficient Optimisation

This chapter relates work done in order to improve the optimisation of the SMT models scaling factors in the log-linear combination. This work was motivated by an experimental issue reported in §2.2.4: optimisations with different initial parameters can converge to fairly distinct local optima.

The choice of the automatic metric used to evaluate translation error is another important optimisation problem.

Thus two research lines were considered: the improvement of the metric used as objective function and the improvement of the optimisation algorithm itself. This chapter is structured as follows:

- The first research line was carried out in cooperation with UPC's software department (and UNED University in Madrid), where the IQmt translation evaluation framework has been developed [Ami06]. IQmt combines metrics with a stable and robust criterion. In §6.1, we consider the IQmt framework as metric to tune the SMT coefficients. By means of a manual evaluation, we compare the translation quality of the resulting configuration to that of a system tuned with the BLEU metric alone.

- In §6.2, the use of the Simultaneous Perturbation Stochastic Approximation (SPSA) [Spa92] algorithm is studied. Coefficient optimisation with the SPSA algorithm is compared to optimisation with the downhill simplex method in terms of performance and variability.

6.1 Machine Translation System Development Based on Human Likeness

6.1.1 Introduction

In this section, a novel approach for SMT parameter adjustment is presented. Instead of relying on a single evaluation metric, or in an ad-hoc linear combination of metrics, our method works over metric combinations with maximum descriptive power, aiming to maximise the Human Likeness of the automatic translations.

By far, the most widely used metric in the recent literature is BLEU, which computes lexical matching accumulated precision for n-grams up to length four [Pap01]. However, it presents several deficiencies which cast serious doubts on its usefulness, both for sentence-level error analysis [Tur03] and for system-level comparison [CB06]. Moreover, optimising over an error measure based on a single metric presents a major drawback. The system may end strongly biased towards configurations which maximise this metric score but may not necessarily maximise the scores conferred by other metrics [Cre05a]. We refer to this problem as system *over-tuning*. Some authors have tried to overcome its negative effects by defining error measures over linear combinations of metrics [Hew05; Che05]. However, in these cases metric combinations are selected arbitrarily, or based on uncertain or ad-hoc criterion.

In our work, we suggest a tuning procedure based on a robust and stable criterion. We aim to maximise the *'Human Likeness'* of automatic translations [Ami06]. Translations are evaluated in terms of the probability that they could have been generated by a human translator, instead of the probability that they could look acceptable to human judges (*'Human Acceptability'*). We approach this target with the QARLA Framework [Ami05]. We apply our methodology to optimise a state-of-the-art SMT system [Cre06b]. We show, through a rigorous manual evaluation process, how tuning based on Human Likeness provides more reliable parameter configurations.

The rest of the paper is organised as follows. Section 6.1.2 describes the fundamentals of QARLA and its application to MT. In Section 6.1.3 the SMT system used is described. Experimental work is deployed in Section 6.1.4.

6.1.2 QARLA for Machine Translation

Inside the QARLA Framework, metrics are ranked according to their descriptive power, i.e. their capability to discern between human and automatic translations [Ami05]. Given

a set of test cases A, a set of similarity metrics X, and sets of human references R, QARLA provides three measures:

- **QUEEN**$_{X,R}(A)$**,** a measure to evaluate the quality of a translation using a set of similarity metrics. QUEEN operates under the assumption that a good translation must be similar to all human references according to all metrics. QUEEN is defined as the probability, over $R \times R \times R$, that for every metric in X the automatic translation $a \in A$ is closer to a reference than two other references to each other.
- **KING**$_{A,R}(X)$, a measure to evaluate the quality of a set of similarity metrics. KING represents the probability that, for a given set of human references R, and a set of metrics X, the QUEEN quality of a human reference is greater than the QUEEN quality of *any* automatic translation in A. Thus, KING accounts for the proportion of cases in which a set of metrics has been able to fully distinguish between automatic and manual translations.
- **JACK**(A, R, X)**,** a measure to evaluate the reliability of a test set, defined as the probability over all human references $r \in R$ of finding a couple of automatic translations a, a' which are (i) close to all human references (QUEEN > 0) and (ii) closer to r than to each other, according to all metrics. In other words, JACK measures the heterogeneity of system outputs with respect to human references. A high JACK value means that most references are closely and heterogeneously surrounded by automatic translations. Thus, it ensures that R and A are not biased.

The QARLA Framework for MT Evaluation is publicly and freely available under the name of IQ$_{\text{MT}}$[1] (Inside Qarla Machine Translation Evaluation Framework) [Gim06]. IQ$_{\text{MT}}$ provides a useful mechanism for MT evaluation based on 'Human Likeness' [Ami06]. We use QARLA in two complementary manners. First, we determine the set of metrics with highest descriptive power by maximising the KING measure. Second, we use QUEEN to measure MT quality according to the optimal metric set. Furthermore, for completeness, we estimate test set reliability by means of the JACK measure.

6.1.3 Translation System

In addition to the translation model, the translation system implements a log-linear combination of six additional feature functions: a 5-gram language model of the target language (denoted TM); a 5-gram language model of target POS-tags (TTM), a 5-gram language model of reordered source POS-tags (TSM), used to support a pattern-based reordering

[1] `http://www.lsi.upc.edu/~nlp/IQMT`.

strategy [Cre06b]; a word bonus feature (WB); and finally, two lexicon models (L1 and L2).

6.1.4 Experimental Work

6.1.4.1 Settings

Experiments were performed in both English-to-Spanish and Spanish-to-English translation directions. 33 variants from 7 families of metrics (BLEU, NIST, WER and PER, GTM, ROUGE, and METEOR)[2] were considered.

We have used the EPPS-05 parallel corpus described in §5.3. Table 6.1 shows statistics of the development and test data used.

		sent	words	vocab.	avg len
Dev. (3 refs)	English	1008	26070	3173	25.9
	Spanish		25778	3895	25.6
Test (2 refs)	English	1094	26917	3958	24.6
	Spanish	840	22774	4081	27.1

Table 6.1: Development and test sets statistics.

6.1.4.2 Procedure

We optimised the contribution of each feature function in the SMT system using our tool based on the Downhill Simplex method (see §4.2.5). Note that in this problem, only a local optimum is usually found. Tuning was performed according to two different MT quality measures, evaluated over development data: (i) BLEU and (ii) QUEEN.

To reduce the possibility of having an initial set of weights which would happen to be particularly bad for one of the two objective functions (leading to a particularly poor local optimum), optimisations were started from three initial parameter sets: 1) all free parameters are set to 1; 2) they are all set to 0; and 3) they are alternatively set to 1 and 0. Thus, for the objective function corresponding to each metric, we got three sets of final parameters. Between these three, we chose the final set which corresponded to the best local optimum in the development set.

In both cases (BLEU and QUEEN), optimal parameters were used to translate the test data, and a manual comparison of the resulting two sets of translations was performed by 4 different human evaluators. Each evaluator compared 150 randomly extracted translation pairs, and assessed in each case whether one system produced a better translation,

[2] A detailed list of the variants incorporated may be found in [Gim05]

		TM	TTM	TSM	WB	L1	L2
E→S	B	0.49	0.24	0.96	1.12	0.58	0.41
	Q	0.65	0.23	1.6	1.58	0.97	0.88
S→E	B	0.38	0.22	1.0	0.9	0.76	0.4
	Q	0.31	0.25	0.72	1.9	0.25	0.76

Table 6.2: Final parameters obtained in Spanish-to-English (S→E) and English-to-Spanish (E→S) directions. The translation model weight is set to 1 and kept fixed. B and Q stand for system optimised respectively with BLEU and QUEEN.

or whether both were of equivalent quality. Strictly equal outputs were removed before choosing the 150 pairs. Each judge evaluated a different set of (possibly overlapping) sentences. In order to avoid any bias in the evaluation, the respective position in the display of the sentences corresponding to each system was also random.

6.1.4.3 Results

As described in Section 6.1.2, the first step deals with finding the optimal metric set, based on the KING measure optimisation. In the case of Spanish-to-English the optimal metric set is: {MTR_{wnsyn}, MTR_{stem} and $\text{RG}_{W\_1.2}$} (KING = 0.1472), where MTR refers to METEOR and RG to ROUGE. Whereas for the English-to-Spanish the optimal metric set is: {MTR_{exact}, MTR_{stem} and $\text{RG}_{W\_1.2}$} (KING = 0.2593). These metric sets are used to compute the QUEEN measure.

The systems optimised with BLEU and QUEEN are then compared at various levels. The final model weights obtained from tuning are indicated in Table 6.2. According to this table, the main characteristic of QUEEN optimisation is its tendency to favour the word bonus model with respect to the translation model and the word and POS tags target language models. Thus, the QUEEN measure rated long sentences more favourably than BLEU.

Automatic results are presented in Table 6.3. According to all metrics, both English-to-Spanish systems are equivalent, whereas the Spanish-to-English system optimised with BLEU achieves better translations (1.7 absolute BLEU [3] and nearly 2 absolute WER above the other system). This was expected. After all, conventional metrics have been developed on the basis of Human Acceptability.

In order to clarify this scenario a manual evaluation has been conducted as described in Subsection 6.1.4.2. Table 6.4 shows, for each evaluator, the results of its manual comparison, along with the results of the comparison of the same sentences with respect to WER

[3] In a scale from 1 to 100

		BLEU	WER	PER	MTR	RG
E→S	B	0.486	40.3	31.4	0.7004	0.3974
	Q	0.480	40.2	31.2	0.7000	0.3972
S→E	B	0.562	33.3	25.3	0.7084	0.4310
	Q	0.545	35.4	26.6	0.7154	0.4330

Table 6.3: Automatic translation evaluation results. MTR and RG stand respectively for METEOR and ROUGE. We had only 2 references so QUEEN was not measured (see below).

		EVAL 1		EVAL 2		EVAL 3		EVAL 4	
		H	W	H	W	H	W	H	W
E→S	B>Q	33	55	37	72	56	52	32	54
	Q>B	41	57	57	51	78	65	60	57
	B=Q	76	38	56	29	16	33	52	39
S→E	B>Q	35	79	31	83	46	91	37	85
	Q>B	41	36	52	37	36	33	46	33
	B=Q	74	35	67	30	68	26	67	32

Table 6.4: Number of sentences that the system optimised with BLEU has translated better (B>Q), worse (Q>B) or with equivalent quality (B=Q) as that optimised with QUEEN, according to Human Experts (H) and WER scores (W). Evaluators of translation into Spanish were different from those of translation into English.

scores. Manual comparisons are in strong disagreement with conventional automatic evaluation metrics. For example, in negative sentences the negation was sometimes omitted by the system tuned with BLEU but not by that tuned with QUEEN. The omission of a word present in *all* references implies indeed a stronger penalty in QUEEN. Table 6.6 shows automatic scores for the translations in Table 6.5. Although the translation of the system tuned with BLEU is more fluent, it has the opposite meaning as the source sentence. When the correct meaning is restored ('B corrected'), BLEU slightly worsen, and QUEEN improves.

Evaluators have clearly considered that the English-to-Spanish system optimised by QUEEN performed better. For translation into English, human judges have globally preferred the system optimised on QUEEN, but with less contrast. The fact that QUEEN favoured longer sentences may provide an explanation, since English is denser than Spanish. As a second possible explanation, translation into Spanish is far more difficult than into English. This difficulty would benefit QUEEN. First, because in that case metrics would become more expressive, i.e. there would be more features to capture in order to distinguish automatic from human translations. Second, because the English-to-Spanish test set would exhibit a higher degree of heterogeneity. Fortunately, we may test this

Source	Creo que **no** se puede pensar que lo que gana una institución lo pierde la otra .
Translation Q	I believe that it is **not** possible to think that what wins a institution **what** loses the other .
Translation B	I believe that it is possible to think that what wins a institution loses the other .
B corrected	I believe that it is **not** possible to think that what wins a institution loses the other .
Ref. 1	I do **not** believe there is any mileage in imagining that what the one institution gains , the other loses .
Ref. 2	I believe that we should **not** think that what one institution wins , the other loses .
Ref. 3	I think that we can **not** think that what an institution wins is lost by the other .

Table 6.5: Example from development data: source, translation of systems tuned with QUEEN (Q), BLEU (B), and references.

Translation	BLEU	WER	QUEEN
Q	0.203	48.2	0.222
B	0.213	48.2	0.056
B corrected	0.204	48.2	0.222

Table 6.6: Evaluation of the translations of Table 6.5.

hypothesis by inspecting the reliability of the test sets according to the JACK measure. The JACK measure for the Spanish-to-English test set is 0.2189, whereas for English-to-Spanish the JACK value is significantly higher, 0.3122. This confirms our intuitions.

Nevertheless, notice that in both cases the level of reliability is low. This was expected. All systems are indeed different parameterisations of the same original system.

Problems caused by the minor reliability of the Spanish-to-English test set could be alleviated by enriching it with outputs by different MT systems implementing other approaches (e.g. rule-based, or word-based SMT), and by working on more sophisticated metrics which discriminate to a greater extent between human and automatic translations.

Finally, we must note some limitations of IQ_{MT}: (i) at least three human references per sentence must be available for the purpose of QUEEN computation, (ii) QUEEN computations depend cubically on the size of reference sets, and linearly on the size of test and metric sets, thus in our current experimental setup there is a severe associated time overhead.

6.2 Tuning Machine Translation Parameters with SPSA

This section is structured as follows. First the essential features of the SPSA method are presented. Then in section 6.2.2, objectives and details of the experimental work are given. In section 6.2.3, results are shown and discussed.

6.2.1 Presentation of SPSA algorithm

The SPSA method has been successfully applied in areas including statistical parameter estimation, simulation-based optimisation, signal and image processing [Spa98b]. It is based on a gradient approximation which requires only two evaluations of the objective function, regardless of the dimension of the optimisation problem. This feature makes it especially powerful when the number of dimensions is increased.

The SPSA procedure is in the general recursive stochastic approximation form:

$$\hat{\lambda}_{k+1} = \hat{\lambda}_k - \mathbf{a}_k \hat{\mathbf{g}}_k(\hat{\lambda}_k) \tag{6.1}$$

where $\hat{\mathbf{g}}_k(\hat{\lambda}_k)$ is the estimate of the gradient $\mathbf{g}(\lambda) \equiv \partial E/\partial\lambda$ at the iterate $\hat{\lambda}_k$ based on the previous mentioned evaluations of the objective function. a_k denotes a positive number that usually gets smaller as k gets larger. Two-sided gradient approximations involve evaluations of $E(\hat{\lambda}_k + \text{perturbation})$ and $E(\hat{\lambda}_k - \text{perturbation})$. In the simultaneous perturbation approximation, all elements of $\hat{\lambda}_k$ are randomly perturbed together and the approximated gradient vector is:

$$\frac{E(\hat{\lambda}_k + c_k\mathbf{\Delta_k}) - E(\hat{\lambda}_k - c_k\mathbf{\Delta_k})}{2c_k} \begin{bmatrix} 1/\Delta_{k1} \\ 1/\Delta_{k2} \\ \vdots \\ 1/\Delta_{kN} \end{bmatrix} \tag{6.2}$$

In equation 6.2, $\mathbf{\Delta}_k$ is a perturbation vector of same dimension N as λ, whose values Δ_i are computed randomly. c_k denotes a small positive number that usually gets smaller as k gets larger. Compared to a finite-difference gradient approximation, involving N times more function evaluations, the simultaneous approximation causes deviations of the search path. These deviations are averaged out in reaching a solution and according to [Spa98b], under reasonably general conditions, both gradient approximations achieve the same level of statistical accuracy for a given number of iterations. Notice that in general, SPSA converges to a local minimum.

The general form of the algorithm consists of the following steps (see section 6.2.2.4 for further implementation details):

Step 1 Calculate gain sequences a_k and c_k.

Step 2 Generate the simultaneous perturbation vector $\mathbf{\Delta}_k$.

Step 3 Evaluate $E(\hat{\lambda}_k + c_k\mathbf{\Delta_k})$ and $E(\hat{\lambda}_k - c_k\mathbf{\Delta_k})$.

Step 4 Approximate the gradient as in equation 6.2

Step 5 Update λ estimate as in equation 6.1

Step 6 Iteration or termination. Return to Step 2 with $k + 1$ replacing k. Terminate if the maximum number of iterations have been reached or if there is little change in several successive iterates.

6.2.2 Experimental Settings

6.2.2.1 Translation system used

In these experiments, in addition to a bilingual 4-gram translation model, the translation system implemented a log linear combination of five feature functions: a 4-gram language model of the target language (denoted TM); a 4-gram language model of target POS-tags (TTM); a word bonus feature (WB); and finally, two lexicon models (L1 and L2).

6.2.2.2 Objectives

Thus we have a translation system whose outcome depends on a set of parameters λ (in this experiment, parameters were restricted to the scaling factors of the various models). We want to minimise a function $E(\lambda)$, which measures the translation errors over a given development set, made by the system with the parameter vector λ. In this experiment, each evaluation of $E(\lambda)$ implies computing full translation of the development corpus (see §4.2.5), which is computationally intensive (the number of evaluations of $E(\lambda)$ to achieve convergence is in the order of 100).

The objective of the experiment was to perform the optimisation of $E(\lambda)$ with the downhill simplex method and the SPSA method, and to compare the consistency of the results over changes in initial settings. For this, we ran the algorithms from 7 different initial points and for each point, for 10 slightly different realisations. For both algorithms, an evaluation of $E(\lambda)$ implied a translation of the development corpus by exactly the same system (except the model weights). Thus, an objective function evaluation had the same computational cost for both algorithms.

We aimed at choosing initial points well distributed in parameter space, but nevertheless realistic. Notice that in the log-linear combination, weights can be rescaled to set one of the parameters to some value, so the translation model was set to 1 and kept fixed during optimisation. In the first initial point, all parameters are also 1, so that all models start with equal weights. In the second initial point, all parameters are equal to 0.5. The other points were chosen in the following way. We collected sets of optimal parameters obtained previously on another development corpus, and noted down in which range the scaling factor of each model behaved. We selected the initial value of each parameter randomly within its corresponding range. Table 6.7 displays the initial points used in the experiments.

ID	TM	TTM	WB	L1	L2
1	0.29	0.52	0.32	1.7	0.84
2	0.5	0.5	0.5	0.5	0.5
3	0.58	0.42	1.4	0.2	0.075
4	1	1	1	1	1
5	1.1	0.22	1.5	1.6	0.29
6	1.2	0.53	1.5	1.3	0.89
7	1.3	0.34	1.2	0.85	0.44

Table 6.7: Sets of initial parameters used in the experiments. In table 6.9 points are referred to by their ID number.

The error function we choose is the BLEU score [Pap01]. Actually it does not measure an error but a translation accuracy, so its opposite is to be minimised.

6.2.2.3 Downhill simplex implementation details

We implemented the simplex method as explained in §4.2.5.

Given a starting point $\mathbf{P}_0$ (see section 6.2.2.2), the other N points of the initial simplex were taken to be $\mathbf{P}_i = \mathbf{P}_0 + \alpha_i \mathbf{e}_i$, where the $\mathbf{e}_i$ are unit vectors. The N constants α_i were chosen randomly such that the perturbed parameter $P_{0i}+\alpha_i$ be in the range corresponding to this scaling factor (as defined in section 6.2.2.2). For each of the seven starting points $\mathbf{P}_0$, we ran the algorithm from 10 different initial simplexes. Different initial simplexes were obtained by varying the seed of the random generator used to compute the α_i constants.

6.2.2.4 SPSA implementation details

After some experiments, we adopted slight changes to the form of the algorithm presented in section 6.2.1. The algorithm presented in section 6.2.1 does not restrict the updated set of parameters λ at a new iteration. This means that if we are unlucky with the $\mathbf{\Delta_k}$ vector, we can go back from an good λ vector to a bad λ vector. This process will eventually converge, but it can take many iterations. Thus we introduced a function evaluation after step 5, to be able to determine whether the new parameters led to an improvement or not. A new set of parameters λ_{k+1} which was worse than the current one (λ_k) was accepted according to the following probability distribution:

$$\exp\left[-\frac{|E(\hat{\lambda}_{k+1}) - E(\hat{\lambda}_k)|}{T\ (k+1)}\right], \tag{6.3}$$

were T was empirically set to 0.005. According to this probability distribution, the worse the new set of lambdas, the less probability to accept it.

Since we introduced an evaluation of $E(\hat{\lambda}_k)$, we replaced the two-sided gradient approximation by a one-sided one, which involves evaluations of $E(\hat{\lambda}_k)$ and $E(\hat{\lambda}_k + \text{perturbation})$. If the noise caused by the one-sided approximation (as opposed to the two-sided approximation) is small compared to the noise arising from the simultaneous approximation, one-sided approximation can be more efficient since it saves one function evaluation at each iteration.

Finally, after a certain number of iterations without improving the optimum, we also changed the distribution in Step 2 to a $0, \pm 1$ distribution with probability 1/3 for each outcome. Although this distribution seems to slow down the algorithm, it allows for a finer approximation of the gradient in a subpart of the parameter space.

At the end the SPSA algorithm was implemented as follows:

Step 1 Calculate gain sequences a_k and c_k. Notice that the choice of the gain sequence a_k and c_k are critical to the performance of SPSA. We choose the basic parameters according to [Spa98a], and tuning the algorithm over various development sets (distinct from the one used in this experiment) from different machine translation tasks. Thus these parameters are expected to be valid for experiments with the same objective function on other language pairs and corpora. We used $a_k = 8/(2+k+1)^{0.602}$ and $c_k = 0.25/(k+1)^{0.101}$.

Step 2 Generate the simultaneous perturbation vector $\mathbf{\Delta}_k$. Each component of $\mathbf{\Delta}_k$ was a Bernoulli ± 1 distribution with probability of 1/2 for each ± 1 outcome (or a $0, \pm 1$ distribution with probability 1/3 for each outcome, as mentioned above).

Step 3 Evaluate $E(\hat{\lambda}_k + c_k\mathbf{\Delta_k})$

Step 4 Approximate the gradient as in equation 6.2, but replacing $E(\hat{\lambda}_k - c_k\mathbf{\Delta_k})$ by $E(\hat{\lambda}_k)$ and dividing by c_k instead of $2c_k$.

Step 5 Update λ estimate as in equation 6.1, and evaluate the objective function with this new set of parameters. Accept the new parameters according to the probability distribution in equation 6.3.

Step 6 Iteration or termination (as in section 6.2.1).

In our comparative experiment, for each starting point $\mathbf{P}_0$, we ran the algorithm with 10 different seeds for the random generator which computes the simultaneous perturbation vector in step 2. These seeds were the same as those used to generate the 10 initial simplexes (see subsection 6.2.2.3).

6.2.2.5 Data set

The translation system was trained with the Chinese-English data provided for IWSLT'06 evaluation campaign, and the parameters were tuned over the development set provided for the same evaluation (dev4). These parameters were then used to translate the test set, which was a selection of 500 sentences among the development sets of previous evaluations (dev1, dev2 and dev3). Table 6.8 shows the main statistics of the training, development and test data used, including number of references, number of sentences, number of words, vocabulary and average sentence length for each language.

	sent	words	vocab.	avg len
Training set				
Chinese	46k M	314k	9725	6.7
English		326k	9643	7.0
Development set (7 references)				
Chinese	489	5478	1096	11.2
Test set (16 references)				
Chinese	500	3005	909	6.0

Table 6.8: *Training, development and test sets statistics.*

6.2.3 Results

We report in table 6.9 the average BLEU score and standard deviation obtained after running 20, 40, 60 and 90 function evaluations of the simplex algorithms and SPSA algorithm. When an optimisation converged before the given number of function evaluations, the optimum value was taken. In each cell of table 6.9, the upper number refers to the simplex, and the lower refers to SPSA. For each initial set of parameters, average and standard deviation are calculated over the 10 slightly different realisations controlled with the random seeds.

	Function Evaluations			
Pt	20	40	60	90
1	17.9±0.45 17.6±0.74	18.4±0.55 18.9±0.53	18.7±0.47 19.4±0.22	18.9±0.47 19.7±0.15
2	18.8±0.40 18.9±0.48	19.1±0.41 19.3±0.15	19.2±0.41 19.5±0.11	19.3±0.42 19.6±0.16
3	19.1±0.22 19.1±0.44	19.3±0.26 19.4±0.21	19.5±0.29 19.5±0.19	19.5±0.30 19.6±0.10
4	18.9±0.45 18.7±0.63	19.4±0.39 19.5±0.28	19.6±0.34 19.6±0.18	19.7±0.34 19.7±0.11
5	19.4±0.23 19.4±0.19	19.6±0.23 19.5±0.17	19.6±0.24 19.5±0.17	19.6±0.25 19.6±0.14
6	19.5±0.19 19.3±0.20	19.7±0.20 19.5±0.20	19.7±0.20 19.6±0.16	19.8±0.20 19.7±0.10
7	19.6±0.13 19.4±0.19	19.7±0.15 19.6±0.12	19.8±0.14 19.7±0.11	19.8±0.15 19.7±0.11
	19.0±0.58 18.9±0.64	19.3±0.45 19.4±0.22	19.4±0.37 19.5±0.09	19.5±0.33 19.7±0.08

Table 6.9: Average BLEU score and standard deviation obtained with the simplex method (above) and SPSA method (below) in the development set, after 20, 40, 60, and 90 function evaluations, for each initial point in parameter space (referred to as pt). In the last row, separated from the rest of the table, the average and standard deviation of the averages are displayed.

First, from table 6.9 it seems that both algorithms have similar performance in terms of the optimum value achieved after a given number of function evaluations. Nevertheless, it is remarkable that from 60 function evaluations on, the standard deviation is always smaller for the SPSA algorithm, which suggests that this is a more stable method. Since the implementation of the different realisations cannot be the same for both algorithms, we cannot be totally certain that it was fair. A change of seed to generate the simultaneous

perturbation for the SPSA may be less significant than a change of initial simplex. To verify this, we need to fix the seed and see how the algorithm behaves across several initial points. A first indication is given by the last row of table 6.9. In this row, the average and standard deviation of the averages $< BLEU >_{Ni}$ taken after N function evaluations, for each point i, are calculated. The standard deviation of the averages, after 60 function evaluations, is much lower with SPSA method, which confirms that on this data set, the simplex optimal value is more dependent on the starting parameters than SPSA.

Table 6.10 explores this point in greater detail. For a given seed used, determining a given realisation, the only varying factor is the starting point. For each seed, the average BLEU score and standard deviation over all 7 starting points, after 20, 40, 60, and 90 function evaluations, are shown. For all seeds, after at least 60 function evaluations, the optimum value obtained with SPSA is less sensitive to the choice of initial parameters, which should lead to more consistent results. For SPSA, the highest standard deviation after 90 function evaluations is 0.18. For the simplex, it reaches 0.67. Thus doing two successive optimisations, one can expect in average up to 0.4 percent BLEU difference with SPSA and up to more than 1.3 percent BLEU different with the simplex. The conjugated effect of two SPSA properties not shared by the simplex method may contribute to explain this difference in stability. Firstly, SPSA search path follows in average the direction of the gradient, whereas the simplex orientation is blind. Secondly, SPSA has always a probability to go away from a zone close to a minimum, which allows it to find a lower minimum elsewhere in the search space. On the contrary, when the simplex shrinks in a zone close to a minimum, it is stuck in that zone.

While optimal objective function values lie in a pretty close range, as seen in Table 6.9, Figure 6.1 show that final values in parameter space are very dispersed. Thus different parameter sets lead to similar scores. Surprisingly, the value of the lexical (L1) model weight does not seem to be determinant, although this model has got a big impact in translation quality [Mar06b]. This is an indication of the interdependence of the various models. Figure 6.1 also suggests that there would be no point in averaging parameter values in order to gain generalisation power.

As ultimate goal, we need to see if the stability of SPSA optimisations is conserved when translating new text (*i.e.* the test corpus). For each initial point and seed, and after a given number of function evaluations, we collected the optimum parameter set over the development corpus, and translated the test corpus with these parameters. Results are brought together in Table 6.11. Table 6.11 instructs, as expected, that dispersion of scores is higher in test than in development. It also reveals that the standard deviation in test is similar for both algorithms. Thus the stability gain observed in the development corpus for SPSA was not conserved with new data. Finally, the last row of Table 6.11 indicates

ID	Function Evaluations			
	20	40	60	90
1	19.0±0.76	19.5±0.22	19.6±0.20	19.7±0.22
	18.9±1.09	19.5±0.31	19.7±0.14	19.7±0.06
2	18.8±0.65	19.1±0.69	19.1±0.68	19.1±0.67
	19.0±0.68	19.5±0.26	19.6±0.16	19.6±0.14
3	19.1±0.44	19.2±0.36	19.4±0.30	19.5±0.28
	19.1±0.40	19.4±0.23	19.6±0.25	19.6±0.18
4	18.9±0.88	19.3±0.73	19.6±0.43	19.7±0.26
	18.8±0.60	19.3±0.29	19.5±0.16	19.6±0.14
5	19.1±0.61	19.3±0.57	19.5±0.41	19.6±0.38
	18.5±1.11	19.2±0.74	19.5±0.23	19.7±0.11
6	19.1±0.38	19.4±0.40	19.5±0.40	19.6±0.38
	18.9±0.98	19.4±0.18	19.6±0.19	19.7±0.18
7	18.8±0.73	19.1±0.66	19.2±0.64	19.3±0.60
	19.1±0.42	19.3±0.33	19.5±0.18	19.6±0.14
8	19.1±0.47	19.4±0.36	19.5±0.40	19.6±0.36
	18.9±0.89	19.4±0.26	19.5±0.16	19.7±0.14
9	19.0±0.72	19.3±0.68	19.5±0.58	19.5±0.57
	18.9±0.54	19.5±0.19	19.6±0.17	19.7±0.11
10	19.1±0.71	19.4±0.59	19.5±0.36	19.6±0.32
	18.9±0.65	19.4±0.28	19.5±0.16	19.6±0.16

Table 6.10: Average BLEU score and standard deviation obtained with the simplex method (above) and SPSA method (below) in the development set, after 20, 40, 60, and 90 function evaluations, for each seed used to generate different algorithm conditions. Only seed ID numbers are displayed.

that the average BLEU score in test is similar for both algorithms.

Since we have got 10 realisations of the algorithms for each initial set of parameters, it is interesting to select the parameters corresponding to the best score out of the 10 realisations, and translate the test corpus with these parameters. The results are plotted in Figure 6.2. Firstly, we can notice that the selected parameters do not lead to substantially better results than the average of all results (reported in the last row of Table 6.11). The largest gain is 0.2 point BLEU for SPSA after 60 function evaluations. Secondly, while no over-fitting seems to occur, continuing the optimisation over development data after 30 function evaluations does not lead to significant translation quality improvement in test. Even 10 function evaluations were sufficient to obtain nearly optimal parameters with both algorithms.

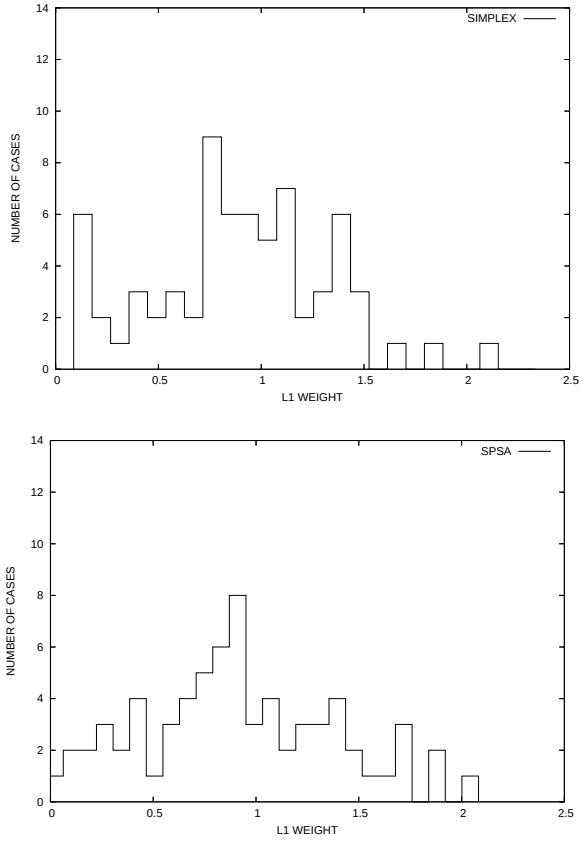


Figure 6.1: Histogram of L1 model weights for simplex (above) and SPSA (below)

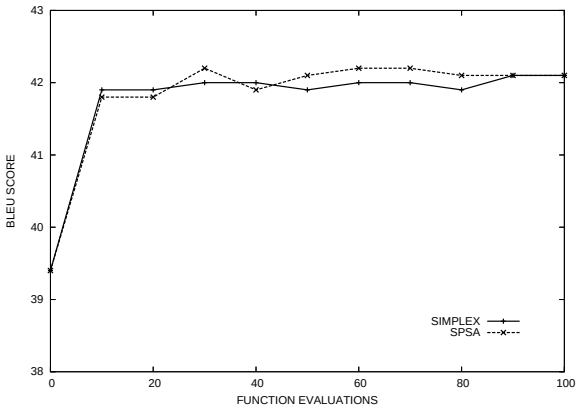


Figure 6.2: Average, over the initial points, of the results in test for the best parameters obtained in development among the 10 different realisations.

Pt	Function Evaluations			
	20	40	60	90
1	41.1±1.08	41.4±1.10	41.4±1.03	41.4±1.09
	40.3±1.12	41.2±0.61	41.6±0.37	41.7±0.52
2	42.3±0.56	42.4±0.65	42.5±0.68	42.5±0.70
	42.1±0.84	42.3±0.78	42.3±0.91	42.3±0.87
3	41.8±0.55	42.0±0.60	42.0±0.60	42.0±0.65
	41.9±0.77	41.9±0.66	42.0±0.69	41.9±0.70
4	41.5±0.55	41.8±0.27	41.8±0.35	41.8±0.34
	41.9±0.65	42.2±0.34	42.1±0.37	42.0±0.38
5	42.5±0.37	42.3±0.46	42.3±0.57	42.3±0.51
	42.2±0.47	42.1±0.60	41.8±0.50	42.0±0.44
6	41.8±0.55	41.9±0.33	42.1±0.29	42.1±0.35
	42.0±0.56	42.1±0.63	42.0±0.56	42.0±0.47
7	42.4±0.32	42.5±0.34	42.5±0.33	42.5±0.31
	42.2±0.50	42.2±0.35	42.1±0.30	42.3±0.33
	41.9±0.52	42.0±0.40	42.1±0.39	42.1±0.38
	41.8±0.69	42.0±0.38	42.0±0.25	42.0±0.22

Table 6.11: Average BLEU score and standard deviation obtained with the simplex method (above) and SPSA method (below) in the TEST set, after 20, 40, 60, and 90 function evaluations, for the parameters obtained in development for each initial point (referred to as pt). In the last row, separated from the rest of the table, the average and standard deviation of the averages are displayed.

6.3 Conclusions and Further Work

Two aspects of the system coefficients optimisation process were studied in this chapter: the metric used as objective function and the optimisation algorithm itself.

We have provided an effective methodology for MT system development based on 'Human Likeness'. We have shown that this approach provides more reliable and robust system configurations than a tuning strategy based on BLEU alone. The disagreement between conventional metric scores and manual evaluation has shown one more evidence of the need for this type of methodology.

Further work include performing similar experiments on other data sets and language pairs, such as Chinese-English and Arabic-English, for example in the framework of NIST evaluations. Another direction, currently explored by Giménez, consists in considering a wider set of partial metrics working at different linguistic levels further than lexical, i.e. at the syntactic and shallow semantic levels.

We have also presented experiments in which the SPSA algorithm has been used to tune SMT parameters. These experiments have been repeated with the downhill simplex method for comparison. According to the results obtained in the IWSLT task, both methods seem to have similar performance. However, SPSA was more robust than the simplex with respect to the choice of initial parameters and with respect to slightly different realisations of the algorithm. This conclusion is not restricted to a particular tuning setup. However, this SPSA advantage was not conserved when using the optimal parameters to translate new data. We also observed a high dispersion in parameter space, showing that various sets of parameters led to similar scores.

While no over-fitting was noticed, nearly optimum results in test could be obtained after only 10 function evaluations over the development corpus. The dispersion of results in test may be overvalued because of the task, which allows particularly poor generalisation since training, development and test corpora are small. Thus, it would be interesting to repeat these experiment with more data, such as those of European Parliament corpus. Furthermore, SPSA being expected to perform better for a problem of higher dimensionality, we should carry out experiments with a system including more feature functions.

The research work reported in this chapter was published in the following contributions:

- [Lam06b] Patrik Lambert and Rafael E. Banchs. 2006. **Tuning Machine Translation Parameters with SPSA**. Proc. of the International Workshop on Spoken Language Translation, pp. 190-196. Kyoto, Japan, November 27-28.

- [Lam06c] Patrik Lambert, Jesús Giménez, Marta R. Costa-jussà, Enrique Amigó, Rafael E. Banchs, Lluís Màrquez and J.A. R. Fonollosa. 2006. **Machine Translation System Development Based on Human Likeness**. Proc. of the IEEE/ACL Workshop on Spoken Language Technology, pp. 246-249. Palm Beach, Aruba, December 10-13.

Chapter 7

Minimum-Translation-Error Discriminative Alignment Training

In this chapter, we propose a novel framework for discriminative training of alignment models with automated translation metrics as maximization criterion. In this approach, alignments are optimized for the translation task. In addition, no link labels at the word level are needed. This framework is evaluated in terms of automatic translation evaluation metrics on a small corpus: the IWSLT06 task (see Table 6.8). Then, the approach is extended to large corpora and evaluated on the English-to-Spanish EPPS-05 task (described in Table 5.3) and on the Arabic-to-English United-Nations corpus task (Table 7.8).

7.1 Related Work

As detailed in §2.1, in the first SMT systems [Bro93], word alignment was introduced as a hidden variable of the translation model. When word-based translation models have been replaced by phrase-based models, alignment and translation model training have become two separated tasks. The alignment task is performed first. In a second stage, these alignments are used to extract translation units and their corresponding probabilities.

Within the maximum entropy approach described in §2.2.3, translation quality can be tuned by adjusting the weight of each feature function in the log-linear combination. In order to improve translation quality, this tuning can be effectively performed by minimizing translation error over a development corpus for which manually translated references are available [Och03a]. As a separate first stage of the process, alignment is not in practice directly tuned in function of the machine translation task. As a consequence, the alignment used is not the one which most benefits to the translation system.

Tuning alignment for an MT system is subject to two main difficulties. The first one is the difficulty of optimizing alignment parameters themselves, which depends on the approach used. Most of the systems cited in recent literature follow either an unsupervised perspective [Och00a; Lia06] or a supervised one [Liu05; Moo05b; Tas05; Itt05; Blu06]. Fraser and Marcu [Fra06] proposed a semi-supervised strategy, interleaving discriminative training with EM training. The unsupervised approach is based on generative models which are trained with the EM algorithm. It has two drawbacks with respect to adapting alignment to the translation task: it is difficult to incorporate new features to the system, and it requires large computational resources. Thus, in practice, external model parameters, such as smoothing parameters for the distortion probabilities, or fertility probabilities, are rarely tuned. In contrast, adding new features to some supervised systems [Liu05; Moo05b; Itt05] is easy, but the need of annotated data is a problem. Although most systems can achieve good performance with only a small amount of manually aligned data, even this small amount is currently difficult to get compared to the availability of bilingual corpora. Furthermore, some systems [Liu05; Tas05] still need to incorporate generative models as features in order to produce quality alignments.

The second difficulty and main obstacle, however, to achieve alignments that would be optimal for machine translation, is common to any approach. It is that of finding an alignment evaluation metric favoring alignments which benefit Machine Translation. The fact that the required alignment characteristics depend on each particular system makes it even more difficult. It seems that high precision alignments are better for phrase-based SMT [Che06a; Aya06], whereas high recall alignments are more suited to n-gram SMT (as suggested by Table 4.3). In this context, alignment quality improvements do not necessarily imply translation quality improvements. This is in agreement with the observation of a poor correlation between word alignment error rate (AER [Och00b]) and automatic translation evaluation metrics [Itt05; Vil06] (see also §5.1.3). In some cases, a clearly worse alignment can even help to achieve better translations [Vil06].

Recently some alignment evaluation metrics have been proposed which are more informative when the alignments are used to extract translation units [Fra06; Aya06]. Fraser and Marcu [Fra06] reported that tuning the trade-off between precision and recall in the F-Measure formula leads to better BLEU scores. This is an interesting result, but it should be noticed that the trade-off obtained depends on the alignment reference, since the values of precision and recall depend on the reference, as was shown in Section 3.1.3. Ayan and Dorr [Aya06] designed the consistent phrase error rate (CPER) for phrase-based SMT, an indicator on how a set of phrases differ from one alignment to the next. Nevertheless, these metrics assess translation quality very indirectly.

We propose a novel framework for discriminative training of alignment models with

automated translation metrics as maximization criterion. First we use a basic beam-search alignment decoder similar to that of Moore [Moo05b], but with features intended to benefit to our translation system. We train the alignment parameters directly on the entire (small) training corpus. In a second stage, we extend the minimum-translation-error alignment training framework to make it usable with large corpora. The alignment system itself is improved to handle the ambiguity introduced by large sentences or free translations, such as in the EPPS or United Nations corpora. Furthermore, the alignment parameters are trained on a small subset of the training corpus and used to align the whole corpus.

As noted by Blunsom and Cohn [Blu06], a weakness of this procedure is that it does not provide an optimal solution to the problem. However, its big advantage is that no observation sequence for training is required. It only requires a way of scoring each alignment a posteriori (in our case, with some translation evaluation metric). Thus we just need a reference aligned at the sentence level instead of link labels at the word level. Furthermore, finding a globally optimal solution of the alignment problem is perhaps the most important thing only when alignment is the end-product of the system. Otherwise, it may be more important to be able to tune alignment in function of the end-product criteria. As mentioned before, in the case of machine translation, finding an optimal solution of the alignment problem would not necessarily imply an improvement of the translation problem.

This chapter is structured as follows. §7.2 explains the models used in our basic word aligner, focusing on the features designed to account for the specificities of the SMT system. In §7.3, our minimum error training procedure for small corpora is described and experimental results are shown. §7.4 describes our improved aligner, focusing on the new findings with respect to similar systems reported in the literature. In §7.5, our extended minimum error training procedure is described and experimental results are shown. Finally, some concluding remarks and lines of further research are given.

7.2 Basic Bilingual Word Aligner

For versatility and efficiency requirements, we implemented BIA, a BIlingual word Aligner similar to that of Moore [Moo05b]. BIA consists in a beam-search decoder searching, for each sentence pair, the alignment which minimizes the cost of a linear combination of various models. The algorithm will be described later (§7.4.3).

The differences with the system of Moore lie in the features, which we specially designed to suit our translation model (see §4.1). Two issues regarding this translation model can be dealt with at the alignment stage.

Firstly, in order to estimate the bilingual n-gram model, only one monotonic segmentation of each sentence pair is performed. Thus long reorderings cause long and sparse tuples to be extracted. For example, if the first source word is linked to the last target word, only one tuple can be extracted, which contains the whole sentence pair. This kind of tuple is not reusable, and the data between its two extreme words are lost.

Secondly, as explained in §4.1, alignments are preprocessed before tuple extraction such that any unlinked target word is attached to either its precedent or its following word.

Taking theses issues into account, we implemented the following features:

- distinct source and target unlinked word penalties: since unlinked words have a different impact whether they appear in the source or target language, we introduced an unlinked word feature for each side of the sentence pair.
- link bonus: in order to accommodate the n-gram model preference for higher recall alignment, we introduced a feature which adds a bonus for each link in the alignment.
- embedded word position penalty: this feature penalizes situations like the one depicted in figure 7.1. In this example, the bilingual units s2-t2 and s3-t3 cannot be extracted because word positions t2 and t3 are embedded between links s1-t1 and s1-t4. Thus the link s1-t4 may introduce data sparseness in the translation model, although it may be a correct link. So we want to have a feature which counts the number of embedded word positions in an alignment[1].

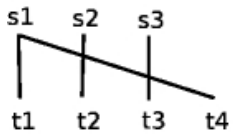


Figure 7.1: Word positions embedded in a tuple.

In addition to the embedded word position feature, we used the same two distortion features as Moore to penalize reorderings in the alignment (one sums the number of crossing links, and the other one sums the amplitude of crossing links). We also used the χ^2 score (see §2.4.2) as a word association model, and as a POS-tags association model. A maximum of one cooccurrence was considered per sentence pair, as in Equation 2.20.

[1]Notice that as part of our translation system reordering strategy (§4.4), the source words are reordered before translation units extraction so as to make alignment more monotonic. Thus embedded word positions can be avoided in the source side of the corpus. We still need a target embedded word position penalty.

7.3 Experimental Work on Small Corpora

For these experiments we used the Chinese-English data provided for IWSLT'06 evaluation campaign [Pau06] (see Table 6.8).

7.3.1 Optimisation Procedure

Once the alignment models were computed, a set of optimal log-linear coefficients was estimated via the optimisation procedure depicted in Figure 7.2.

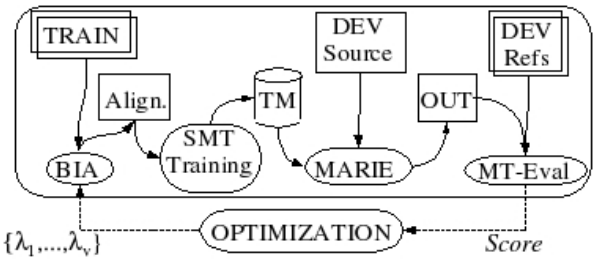


Figure 7.2: Optimisation loop.

The training corpus was aligned with a set of initial parameters $\lambda_1, \ldots, \lambda_7$. This alignment was used to extract tuples and build a bilingual n-gram translation model (TM). A baseline SMT system, consisting of MARIE decoder and this translation model as unique feature[2], was used to produce a translation (OUT) of the development source set. Then, translation quality over the development set is maximized by iteratively varying the set of coefficients.

The optimisation procedure was performed by using the SPSA algorithm (see §6.2.1).

Each function evaluation required to align the training corpus and build a new translation model. The algorithm converged after about 80 evaluations, lasting each 17 minutes with a 3 GHz processor. Alignment decoding was performed with a beam of 10 (it took 50 seconds and required 8 MB memory).

Finally, the corpus was aligned with the optimum set of coefficients, and a full SMT system was build, with a target language model (trained on the provided training data), a word bonus model and two lexical models. SMT models weights were optimized with a standard Minimum Error Training (MET) strategy[3] and the test corpus was translated

[2] As mentioned in §4.2, an n-gram SMT system can produce good translations without additional target language model.

[3] SMT parameters are not optimized together with alignment parameters for two main reasons. Firstly, translation is more sensitive to variations of SMT parameters. Secondly, alignment is optimized over the full training set, whereas SMT is tuned over the development set.

with the full system. To contrast the results, full translation systems were also build extracting tuples from various combinations of GIZA++ alignments (trained with 50 classes and respectively 4,5 and 4 iterations of models 1,HMM and 4). In order to limit the error introduced by MET, we translated the test corpus with three sets of SMT model weights, and took the average and standard deviation.

7.3.2 Results

Table 7.1 shows results obtained with the full SMT system on the test corpus, with GIZA++ alignments, and BIA alignments optimized in function of three metrics: BLEU, NIST, and BLEU+4*NIST. The standard deviation is indicated in parentheses. Although results for systems trained with different BIA alignments present more variability than systems trained with GIZA++ alignments, they achieve better average scores, and one of them obtains much higher scores. Unexpectedly, BIA alignments tuned with NIST yield the system with worse NIST score.

System	BLEU	NIST	PER	WER
GIZA++ union	42.7 (1.1)	8.82 (0.07)	34.7 (0.2)	43.7 (0.4)
GIZA++ intersection	42.4 (0.9)	8.53 (0.07)	37.0 (0.9)	45.0 (1.3)
GIZA++ Zh→En	43.7 (0.9)	8.90 (0.2)	37.2 (1.4)	45.5 (2.0)
BIA (BLEU)	44.8 (0.4)	9.00 (0.04)	35.7 (0.07)	43.8 (0.09)
BIA (BLEU+4*NIST)	47.0 (1.5)	8.83 (0.4)	32.9 (0.8)	40.9 (0.5)
BIA (NIST)	44.8 (0.1)	8.55 (0.14)	33.0 (0.2)	41.4 (0.5)

Table 7.1: *Automatic translation evaluation results.*

7.4 Improved Bilingual Word Aligner

The basic alignment system is good enough to align parallel corpora with short sentences or with translations which closely mirror the original. However, it can not handle the ambiguity introduced by large sentences or free translations, such as in the EPPS or United Nations corpora. Thus in a first stage, we tried to correct the main weaknesses of this baseline alignment system, looking at the recall, precision and AER over a freely available alignment reference. As discussed in Subsections 7.4.2, 7.4.3 and 7.4.4, we studied various types of association scores, improved the search algorithm, and introduced other models, including fertility models. Notice that since our aim was not to obtain the best possible AER, we have not optimised the complete system parameters according to AER, and we have limited ourselves to mentioning the most striking issues.

7.4.1 Alignment Data

We trained the system with the EPPS-05 corpus (§5.3). The test alignment data described in Chapter 3 was divided in two sets, as presented in Table 7.2.

(a) Alignment development data set

Lang.	Sentences	Words	Vocab.	Aver.
Eng	246	7.0 k	1.9 k	28.6
Spa	246	7.4 k	2.2 k	29.9

(b) Alignment test data set

Lang.	Sentences	Words	Vocab.	Aver.
Eng	245	7.3 k	2.0 k	29.6
Spa	245	7.7 k	2.3 k	31.5

Table 7.2: Basic statistics for the considered alignment development (a) and test (b) data sets (M and k stands for millions and thousands, respectively).

Interestingly, experiments of §5.1 were performed on these data and no improvement in translation was observed when reducing the AER from 21.9 to 17.1. We trained with Giza++ tool the best straightforward system presented in §5.1: aligning stems with 50 classes, and respectively 4,5 and 4 iterations of models IBM 1, HMM and IBM 4. The corresponding results (referred to as Giza++ union and intersection) are shown in table 7.4.

7.4.2 Association Score

In order to compare in terms of AER various ways of computing the association scores and various types of association measures, for each variant the system's weights were optimised as a function of development's corpus AER, and were then used to align the test data. To highlight the effect of the association model, we added only the link bonus, the unlinked word penalty and the two distortion features. The values presented are the average (and standard error between parentheses) of values from three optimisations of the system's weights. Only one possible association was considered for each word.

How to Calculate the Association Score

As noted in §2.4.2, the χ^2 score considers both positive and negative associations. Since we are only interested in positive associations, we only considered token pairs for which

$p(s,t) > p(s)p(t)$. We also implemented the two ways mentioned in §2.4.2 to count cooccurrences for a pair token (s,t). In the first one, multiple cooccurrences in the same sentence pair are taken into account (Equation 2.19). In the second one, the maximum cooccurrence count in the same sentence pair is 1, whatever the number of occurrences of the considered tokens (Equation 2.20). In Table 7.3, the AER obtained in each case is presented.

AER results when taking into account positive associations only or positive and negative are compared in the first two rows of Table 7.3. The considered associations, positive ("+") and negative ("-"), are indicated in the "asso" field. The results show that the fact of restricting the association table to only positive association has very little impact in terms of AER. It can be surprising that when negative associations, and thus the corresponding links, are pruned out from the table, recall increases. This can nevertheless happen if bad links which prevented good links to be computed (for example because of distortion costs), were pruned out.

tokens	asso	cooc	Rs	Pp	AER
words	+	single	59.9 (0.3)	85.0 (0.4)	29.4 (0.1)
words	+ -	single	59.7 (0.5)	85.0 (0.8)	29.5 (0.1)
words	+	multiple	52.7 (0.3)	79.2 (0.2)	36.4 (0.2)
stems	+	single	62.4 (0.8)	86.7 (1.5)	27.1 (0.1)

Table 7.3: Recall (Rs), Precision (Pp) and AER in EPPS test corpus for various implementations of the χ^2 score.

Looking at the first and third rows of Table 7.3, it can be concluded that it is better to count only a single cooccurrence for a pair type and a given sentence pair (referred to as "single" in the "cooc" field) than counting multiple cooccurrences. This is because counting multiple cooccurrences gives relatively more weight to associations between frequent words, which can generate noisy links because they have a high association probability with many different words, with respect to associations between content words, which in any case seldom appear more than once in the same sentence pair.

Finally, a basic idea to limit the ambiguity of the alignment task is to restrict the vocabulary by aligning stems instead of full forms[4]. The corresponding AER results are presented in the last row of Table 7.3. As expected, aligning stems is an easier task than aligning words, and all metrics are improved by 1.5 to 2.5 points.

Taking into account the conclusions of this section, we will calculate association scores between stems, considering only positive associations and counting a maximum of one

[4]In this purpose stems are more useful than base form because frequent words such as auxiliary verbs remain unchanged [Gis].

cooccurrence of a pair type in the same sentence pair.

Comparison of Different Association Measures

Another issue to consider is the choice of word association model, since it is used both to determine the order in which links are introduced in the alignment, and as a central model in the log-linear combination. In the basic aligner, χ^2 score was used. However, as seen in §2.4.2, it overestimates the significance of rare events and the log-likelihood ratio (LLR) is a better statistics to account for rare events occurring in large samples. For example, the probability of cooccurrence of two singletons in the same sentence pair is 1 according to χ^2, and less than 0.0001 according to the LLR statistics[5].

We also experimented using IBM1 probabilities [Bro93] (trained with Giza++ tool), since they benefit from various iterations of the EM algorithm with a reasonable training cost increase (106 min, compared to 50 min for χ^2 or LLR training).

	Rs	Pp	AER
χ^2	62.4 (0.8)	86.7 (1.5)	27.1 (0.1)
LLR	59.4 (0.1)	75.7 (0.5)	33.2 (0.3)
IBM1	65.9 (0.7)	90.3 (1.4)	23.5 (0.3)
Giza++ U	77.9	89.2	16.6
Giza++ GDF	76.4	92.8	15.9
Giza++ I	68.0	98.5	19.2

Table 7.4: Recall (Rs), Precision (Pp) and AER in EPPS test corpus for various types of association scores (for stems). Scores for union (U), Grow-forward-diag refined symmetrisation (GDF) and intersection (I) of source-target and target-source Giza++ alignments are also indicated.

As can be observed from Table 7.4, the substitution of χ^2 score by the more accurate LLR yielded surprising results, notably a 11 points precision drop. In fact, this result makes sense. To come back to our example, it would certainly be an error to give some significance to two singletons cooccurring in the same sentence of a monolingual corpus. Nevertheless, in a large and reasonably parallel corpus, wouldn't you think that a singleton in the source language appearing in the same sentence pair as another singleton in the target language, have a high probability to be the translation of each other ? This is true in a vast majority of cases. The IBM model 1 probability in this case is actually equal to 1. Despite of this, the LLR scores gives a very low probability to these links. On the contrary, it gives a (comparatively to χ^2) higher importance to links involving frequent

[5]We normalised the LLR scores so that the best one has a score equal to 1.

words, which are often stop words. Thus LLR score misses some links between content words, and can facilitate links between content words and stop words, which produces noisier alignments. As to IBM1 probability, they are better than association scores and yield a 3.5 points improvement over χ^2 stem association scores.

7.4.3 Search

The search strategy described by Moore [Moo05b] and that we used in the basic aligner (§7.2), is depicted in Figure 7.4. We will refer to this strategy as *baseline search*. Figure 7.4 corresponds to the example of Figure 7.3.

Sentence pair:

0	1	2	3	4			
europ	is	at	a	crossroad			
europ	se	encuentr		en	un	moment	crucial
0	1	2		3	4	5	6

Possible links (association score cost order):

Link	Cost	Corresponding words
0-0	0.0984768	europ-europ
3-4	0.389904	a-un
2-5	0.909069	at-moment
1-2	1.29639	is-encuentr
1-1	1.31789	is-se
2-3	1.79945	at-en
4-2	2.76726	crossroad-encuentr
4-6	2.76726	crossroad-crucial

Figure 7.3: Possible links list example. Here we considered only the links corresponding to the best target word for each source word in the sentence pair or the best source word for each target word ($N_{sp} = 1$). Corresponding words are actually stemmed forms.

Figure 7.3 shows the list of links considered in search (referred to as the list of *possible links*[6]). This list is obtained as follows. First, the association model table is built by pruning IBM 1 probability tables. This pruning is performed by using histogram counts with two thresholds $N_\cap$ and $N_\cup$. Links belonging to the $N_\cup$-most probable ones for a given source word *or* the $N_\cup$-most probable ones for a given target word are selected, as well as links belonging to the $N_\cap$-most probable ones for a given source word *and* the $N_\cap$-most probable ones for a given target word. $N_\cup$ is on the order of 10 and $N_\cap$ on the order of 100.

[6]The notion of possible link here has nothing to do with the notion of possible link in an alignment reference introduced in Chapter 3.

Among the links present in the association model table, a second histogram pruning is performed withing each sentence pair, with a third threshold N_{sp}. In this second pruning, only links belonging to the N_{sp}-most probable ones for each source word *or* the N_{sp}-most probable for each target word are considered in the list of possible links.

Possible links are arranged in a certain number of stacks of links to be expanded during search. Search starts with an initial alignment, for instance the empty alignment. Then all the hypotheses of the current alignment stack are expanded with each link of the current link stack. Histogram pruning is applied to the alignment hypothesis stack to reduce complexity. Threshold pruning can also be applied, but it did not help in preliminary experiments and we did not use it.

As depicted in Figure 7.4, in the baseline search strategy the possible links, sorted in function of their cost, are arranged one link per stack, together with the "empty" link $\emptyset$. Baseline search always begins with the empty alignment (alignment stack 0). This hypothesis is expanded with each link of link stack 1, forming two new hypotheses (the empty alignment and the alignment containing the link 0-0) which are copied into alignment stack 1. Then each hypothesis of alignment stack 1 is expanded with each link of link stack 2. The action of the empty link ensures that all the hypotheses present in stack i (after histogram pruning) remain in stack $i + 1$. The dotted line in alignment stack 2 illustrates the histogram pruning threshold for a beam size of 3. In this case only the best three hypothesis (hypotheses stacks are sorted) would be expanded with the links of link stack 3.

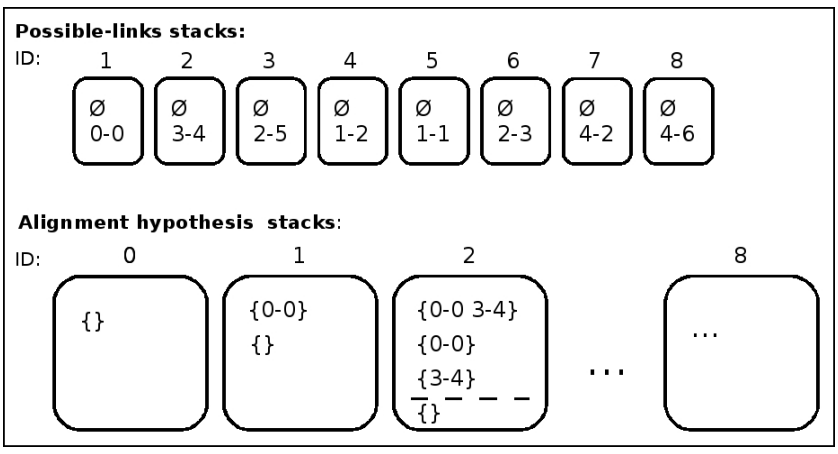


Figure 7.4: Baseline search: link-by-link search following association score order.

In our view, the main drawback of the baseline search strategy is that the final alignment depends on the order in which links are introduced. A very probable but incorrect link to some word can indeed be introduced in alignment hypotheses before the correct link to that word. When expanding these hypotheses with the correct link, the presence

of the incorrect link may produce a distortion cost which causes the new hypotheses containing the correct link to be pruned out. In our example, suppose that link 2-5 (which is actually wrong) has a lower cost than link 3-4 and thus would be introduced first in the alignment hypotheses. If after histogram pruning all the hypotheses in the stack contain link 2-5[7], the correct link 3-4 might not be introduced in any hypothesis since it will generate a distortion cost because of the crossing with link 2-5.

The adequate solution to this problem would be to estimate the remaining cost of each hypothesis, but this would probably be either too time consuming or too inaccurate. To help overcoming this problem, we perform two successive iterations of the alignment algorithm. In the second one, we start from the final alignment of the first iteration instead of the empty alignment. In the context given by the first iteration's final links, incorrect links are more easily penalised with distortion cost. With an initial alignment which is not empty, expanding an hypothesis with some link still means introducing this link in the alignment hypothesis if the link is not present yet, but also means removing it if it is already present.

A related issue in the baseline search is that links to the same word cannot compete on a fair basis since they are introduced successively in the alignment hypotheses. To solve this problem, we implemented a search strategy following the source (or target) sentence words. In this scheme, possible links are organised in one stack for each source (or target) word, as in Figure 7.5. The links of each stack are used to expand the same hypotheses. In the example of Figure 7.5, there is a stack for each source word, sorted according to the cost of the best link in the stack. We will refer to this strategy as "source-word-score" search. Another possibility is to follow strictly the source sentence, without reordering the stacks. This other possibility will be referred to as "source-word-position" search.

The total number of alignment hypotheses created during search is the same with baseline and "source-word" strategies, since the number of "possible links" is equal. However, the search following the source sentence words, in the way depicted in Figure 7.5, only allows many-to-one links since each hypothesis can only be expanded with one of the various possible links to the same source word. To allow many-to-many links, the stacks of possible links associated to a given word must also contain combinations of these links. Each combination represents an additional alignment hypothesis to create compared to the baseline search. To limit complexity, we allow only combinations with consecutive target positions, except one admitted empty position in between.

The three search strategies described in this section are compared in Table 7.5. In

[7]Of course, the solution to this seems to be an adequate choice of the pruning threshold. However, this threshold cannot be infinite and after a few iterations, it is frequent that some links are present in all hypotheses.

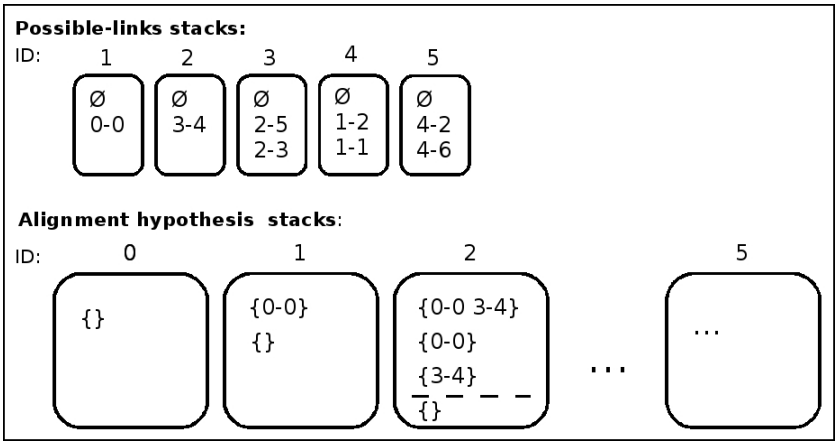


Figure 7.5: "source-word-score search": word-by-word search following source words reordered according to each word's most probable link.

the case of the "source-word-position" strategy, since alignment hypotheses are expanded at consecutive words, it makes sense to recombine the alignment hypotheses with equal recent history. Although hypothesis recombination helps, this strategy gives the worst results because the first links introduced are not the best ones. The best strategy is the "source-word-score" one, in which links to the same words are compared fairly, but keeping the idea of introducing the best links first.

	Rs	Pp	AER	Hyp
Baseline	65.9 (0.7)	90.3 (1.4)	23.5 (0.3)	279k
source-word-position	64.3 (0.4)	88.5 (1.2)	25.2 (0.2)	265k
source-word-score	66.8 (0.3)	90.2 (0.3)	22.9 (0.3)	349k

Table 7.5: Recall (Rs), Precision (Pp), AER and number of hypotheses created in the test data for various search strategies. The baseline model here ("Base") is the best system of table 7.4.

The number of hypotheses created to align the development set is indicated in the last column. As mentioned above, the complexity is higher for "source-word-score" search than for baseline search. "source-word-position" search involves less hypotheses because some of them were recombined. If "source-word-score" search is slightly more expensive than baseline search, it offers also more flexibility to control complexity. The link stacks (possibly containing link combinations) can be sorted and pruned by histogram and/or threshold. We can also set the maximum number of words in the combinations. "source-word" searches also make it possible to expand in the first place words with less links than the average (which are often content words). This gives a context of links which helps aligning the other words (which are often stop words).

7.4.4 Final Alignment Strategy

The IBM1 model also considers for each word the probability to remain unlinked (or linked to NULL). For the system with "source-word-score" search presented in Table 7.5, we substituted the uniform unlinked word penalty by a penalty proportional to the IBM1 NULL link probability. As can be seen from table 7.6, this yielded a more than 3 points precision gain without losing recall. Increasing the number of possible associations improves the recall but also worsens the precision.

	Rs	Pp	AER
up ($N_{sp} = 1$)	66.8 (0.3)	90.2 (0.3)	22.9 (0.3)
um ($N_{sp} = 1$)	67.1 (0.2)	93.5 (0.5)	21.6 (0.0)
um ($N_{sp} = 3$)	69.4 (0.6)	91.5 (0.8)	20.8 (0.3)

Table 7.6: Effect of the unlinked model based on IBM1 NULL probabilities. N_{sp} is the maximum number of possible associations considered for each word. "up" is the best model of Table 7.5, and "source-word-score" search is used in all three cases. "um" refers to the system with unlinked model instead of uniform unlinked penalty.

We now describe the alignment strategy adopted in the experiments reported in §7.5. We considered until 3 possible associations per word in the list of possible links. To help selecting the good links, we added source and target ranking features. A value r of the target ranking feature means that a target word is involved in the r most probable association given a source word. We also added, in addition to the models described in §7.4.2, an embedded word penalty model.

With these models we ran an iteration of source-word-score search, and an iteration of target-word-score search (*i.e.* along the target sentence). The output of each iteration is the best alignment found during search. We took the intersection of these two alignments. This intersection was the initial alignment for another iteration of source-word-score and target-word-score searches. At the end we took an "intelligent" union of the last two searches. This constituted the first pass of our alignment strategy. Our union method is "intelligent" in the sense that when two links to the same word leave empty positions on the other side of the sentence, only one link is kept, depending on the cost of the corresponding alignment hypothesis.

In order to improve alignment quality, we performed a second alignment pass in which the association score model with IBM1 probabilities and the unlinked model were substituted by two improved models benefiting from the first pass links: an association score model with relative link probabilities as suggested by Melamed [Mel00], and source and target fertility models (giving the probability for a given word to have one, two, three or

four links). In the second pass we also used the following features: link bonus, distortion features, and source and target embedded word position penalties. With these models we performed again source and target-word-score searches to obtain the intersection, using it as initial alignment for another iteration along each side of the corpus to obtain the "intelligent" union.

The "possible" links combinations in the stack associated to each word were restricted to basic links, with at most 12 per stack. Links with a cost difference of more than 3 with respect to the best link were also discarded.

Notice that we didn't optimize the full system in function of AER score. However, when optimized in function of BLEU (see §7.5), it ranged between 17 and 20.

In some experiments, we used a chunk based filtering feature, proposed by Crego[8]. Mainly, we use the idea that words in a source chunk are typically aligned to words in a single target chunk to discard alignments which link words from distant chunks. Considering too strict permitting only one-to-one chunk alignments, we extend the number of allowed alignments by permitting words in a chunk be aligned to words in a target range of words, which is computed as projection of the considered source chunk. The resulting refined set contains all the Intersection alignments and some of the Union.

7.5 Minimum-Translation-Error Alignment Training

7.5.1 Data set

We realised experiments on two data sets and language pairs: the EPPS corpus described in §5.3, translating from English (Eng) to Spanish, and a subset of the United Nations (UN) corpus distributed by the Linguistic Data Consortium (reference LDC2004E13), from Arabic (Ar) to English. Statistics for the EPPS-05 training corpus are shown in §5.3. Statistics for the translation development and test data are collected in table 7.7. Statistics for the UN corpus are collected in table 7.8. UN-b, UN-c and UN-d come respectively from test data of NIST evaluation[9] in 2002, 2004 and 2005. Although the UN corpus is in a slightly different domain as NIST test data (which are mainly in the news domain), we decided to use a subset of it as training corpus for the following reasons:

- we did not want to mix domains in the training corpus to avoid side-effect related to this mixture

[8] unpublished to this date
[9] http://www.nist.gov/speech/tests/mt/

- we wanted a corpus of more than 30 millions words, and the corpus in the news domain that was available to us had less than 15 millions words.

As a consequence of the training and test sets being in slightly different domains, scores are much lower than for the same system trained with news data.

EPPS-a) Development data set used for alignment parameter optimisation (maximising BLEU score)

Lang.	Sentences	Words	Vocab.	Aver.
Eng	1008	26.1 k	3.2 k	25.9

EPPS-b) Development data set used for SMT parameter optimisation (maximising BLEU score)

Lang.	Sentences	Words	Vocab.	Aver.
Eng	735	18.7 k	3.2 k	25.5

EPPS-c) Test data set

Lang.	Sentences	Words	Vocab.	Aver.
Eng	1094	26.9 k	4.0 k	24.6

Table 7.7: Basic statistics for EPPS development (a) and (b) and test (c) data sets (M and k stands for millions and thousands, respectively).

7.5.2 Optimisation Procedure

Once the alignment models were computed, a set of optimal log-linear coefficients was estimated via the optimisation procedure depicted in Figure 7.2. First, we selected a small subset of the parallel corpus. The selection criterion was based on two keys: the number of occurrences in the training corpus of the least frequent word in the sentence pair, and to decide between draws, the same criterion only among words present in the development data set.

The training corpus subset was aligned according to the strategy detailed in §7.4.4, with a set of initial parameters $\lambda_1, \ldots, \lambda_\nu$. The parameters of the first and second alignment pass were optimised together. This alignment was used to extract translation units and build a bilingual n-gram translation model (TM). A baseline SMT system, consisting of MARIE decoder and this translation model as unique feature, was used to produce a translation (OUT) of the development source set. In a variant of this procedure, the full n-gram SMT system (with a target language model, a word bonus model and two lexical

UN-a) Training data set

Lang.	Sentences	Words	Vocab.	Aver.
Ar	1.43 M	45.2 M	191 k	31.7
Eng	1.43 M	43.8 M	134 k	30.7

UN-b) Development data set used for alignment parameter optimisation (maximising BLEU score)

Lang.	Sentences	Words	Vocab.	Aver.
Eng	1043	29.1 k	5.9 k	27.9

UN-c) Development data set used for SMT parameter optimisation (maximising BLEU score)

Lang.	Sentences	Words	Vocab.	Aver.
Eng	950	29.8 k	6.8 k	31.3

UN-d) Test data set

Lang.	Sentences	Words	Vocab.	Aver.
Eng	1056	32.1 k	6.3 k	30.4

Table 7.8: Basic statistics for UN corpus training (a), development (b) and (c) and test (c) data sets (M and k stands for millions and thousands, respectively).

models along with the translation model) was used to translate the development source set. In this variant, an additional internal optimisation of SMT system weights was realised with another development set. Then, for both variants, translation quality over the development set (EPPS-a or UN-b in Tables 7.7 and 7.8) was maximised by iteratively varying the set of alignment coefficients.

The optimisation procedure was performed by using the SPSA algorithm (see Chapter 6).

Each function evaluation required to align the training corpus subset and build a new translation model. We performed about 80 evaluations of the objective function Training times are indicated in the results table.

Finally, the whole training corpus was aligned with the optimum set of coefficients, and a full SMT system was build, with a target language model, a word bonus model and two lexical models. SMT models weights were optimised with a the rescoring loop scaling factor optimisation strategy described in §4.2.5, and the test corpus was translated with the full system. To contrast the results, a full translation system was also build extracting

translation units from the union of source-target and target-source GIZA++ alignments (trained with 50 classes and respectively 4,5 and 4 iterations of models 1,HMM and 4). In previous experiments, better results were achieved with the union than with a refined symmetrisation method, at least with the translation model as unique feature (see Table 4.3 in §4.3.2). Despite of this, we also contrasted here the results with the "refined" symmetrisation method of Och and Ney [Och03c] and the "grow-diag-final" method of Koehn *et al.* [Koe].

7.5.3 Results

The ideal testing scheme for our alignment system would have been to evaluate in terms of automated translation metrics the effect of each feature and each parameter. Since this was too costly, what we did is to evaluate the impact on AER, with the EPPS alignment development and test data, of most parameters and features. The most remarkable results of these experiments were shown in §7.4. These AER based experiments helped to define our final alignment system (§7.4.4). Nevertheless we also measured the impact on translation metrics of some parameters, as detailed in §7.5.3.1. In §7.5.3.2, the impact of some characteristics of the minimum error training procedure is discussed. Recall that optimal alignment parameters were used to align the *whole* training corpus, and all the results were obtained with translation systems built on the *whole* training corpus (except when otherwise explicitly indicated). In this section, BLEU score was the metric used for the alignment model weights tuning. In order to limit the error introduced by SMT system Minimum Error Training, we translated the test corpus with three sets of SMT model weights, and the results shown are the average and standard deviation (in parenthesis).

7.5.3.1 Effect of some Characteristics of the Alignment System

As mentioned in §7.4.4, in the second alignment pass the links obtained in the first pass are used to estimate fertility models as well as an improved association model based on relative link probabilities. However, it doubles the alignment training time. The number of feature function weights to be tuned is also doubled, which increases the optimisation problem complexity. Therefore, it is crucial to check that the second pass is useful for the resulting SMT system.

Another key parameter of the alignment system is the threshold N_{sp} for the possible links pruning at sentence pair level. According to preliminary AER results, this parameter was set to 3. However, we noted that some frequent words are involved in many links, which slows down the algorithm and can produce noise. Even if only the 3 best links for

this word are considered, if this word is indeed included in the 3 best of many other words, the total number of links involving it can be much larger than 3. To alleviate the problem without limiting the number of links involving less frequent words, we restrict the number of associations for words involved in more links than the average for this sentence pair. In the example of Figure 7.3, the average number of links per source word is 1.6 and the words at position 0 and 3 are involved in less links than the average (l link), whereas the words at positions 1, 2 and 4 are involved in more links than the average (2 links). In this case, though, the maximum number of associations N_{sp} is 1 and thus cannot be limited further for words at positions 1, 2 and 4. Notice also that this is a bad example since a rare word with semantic content ("crossroad") is involved in more links than the average, whereas a very frequent word without semantic content ("a") is involved in less links than the average. In real life examples, however, words like "a" are usually involved in many links, and rare words like "crossroad" are usually involved in few links.

Table 7.9 shows results obtained with 1 and 2 passes, and with the maximum number of associations limited to 1, 2 and not limited (that is, equal to N_{sp}, which here is 3) for words involved in more links than the average.

Passes	>avge assoc.	BLEU	NIST	PER	WER	METEOR	Hours
1	1	25.0 (0.03)	7.88 (0.02)	43.5 (0.4)	66.3 (0.3)	53.4 (0.1)	15
	2	24.5 (0.3)	7.87 (0.06)	43.5 (0.4)	66.1 (0.8)	53.1 (0.4)	22
	3	24.2 (0.3)	7.82 (0.1)	44.0 (1.3)	66.4 (0.9)	52.9 (0.3)	25
2	1	25.2 (0.1)	8.02 (0.01)	42.5 (0.1)	64.9 (0.4)	53.5 (0.1)	32
	2	26.7 (0.2)	8.19 (0.1)	42.0 (0.6)	64.3 (1.3)	54.3 (0.6)	47
	3	25.4 (0.2)	8.10 (0.15)	42.0 (1.0)	64.3 (1.8)	53.4 (0.7)	58

Table 7.9: Effect of performing only one alignment pass and of limiting the number of maximum associations for words involved in more links than the sentence pair average. The number of maximum associations for these words is referred to as ">avge assoc.". These results are for the full AR→EN UN-a corpus (see Table 7.8), with a full SMT system. However, alignment system weight optimisation was performed on a 25000 sentences subset of the training corpus, using only the translation model as SMT system. BLEU and METEOR scores were multiplied by 100 to make reading easier.

From Table 7.9 it is clear that the second alignment pass allows to build better translation systems, since every cell of the table concerning systems with two passes is better than the corresponding cell for systems with only one pass.

Regarding the maximum number of associations for words involved in more links than the average, results are surprisingly different for systems with 1 and 2 passes. Whereas for one-pass systems the smaller the maximum number of associations, the better the score

are, for two-pass systems, a value of 2 of this number gives clearly better scores (especially BLEU score) and a value of 3 seems to be better than a value of 1. This difference between one and two-pass systems might be due to the fertility models, which indicate which words should benefit from the larger number of links available and which not. Finally, the computational complexity reduction when the number of possible associations is restricted is indicated in terms of time in the last column.

7.5.3.2 Impact of the Alignment Minimum Error Training Procedure

In this section we study the influence of the size of the sub-corpus chosen to optimise the alignment system parameters. We also compare the results obtained with two different SMT systems used during alignment parameter optimisation: the full SMT system (which requires an additional internal tuning at each iteration, that of the SMT features parameters) and the one consisting of the translation model alone (referred to as "BM").

Results for the Arabic to English task are presented in table 7.10. In this task we optimised the alignment parameters on both a 25000 and a 50000 sentences subset of the training corpus (referred to respectively as "25" and "50"). Contrarily to what we thought since the experiments reported in Table 4.3, the best Giza++ system in this case is not given by the union but by the "grow-diag-final" symmetrisation method.

Concerning BIA (our aligner) systems, as expected the size of the corpus seems to be important, since both the "BM" and "full" variants get better average scores when tuned on a 50000 sentences corpus than on a 25000 sentences corpus. Nevertheless, all the differences lie within the error of the SMT weights optimisation process, except for BLEU score. On the contrary, there is no clear pattern regarding BM and full SMT systems. For the 25000 sentences corpus, the full system gets slightly better score (except METEOR) than the BM one, but the situation is reversed for the 50000 sentences corpus. These variations might be due to alignment weights optimisation process error. Anyway in this case all differences between two system's scores lie within the standard deviation of the SMT MET process. Notice also that in the experiments presented in Table 7.10, the chunk based filtering feature was used. This filtering is the only difference between the "25 BM" system and the system with 2 passes and with a maximum number of 2 associations for words with more links than the average in Table 7.9. The average BLEU score for this latter system is 0.7 points better than for the "25 BM" system. This could be due to the fact that chunk filtering slightly harms translation quality. It could also be a typical variation of the alignment weights optimisation procedure.

Alignment parameters training time, referred to as "Hours", is also indicated. The value before the plus sign corresponds to 80 objective function evaluations. The value

after the plus sign corresponds to the time needed to align the entire training corpus. In this task training is really faster with our method. Two further advantages of this kind of system, which aligns the corpus linearly sentence by sentence, are that memory usage is considerably low (except at the very beginning, to train the association score table), and aligning new data for a corpus for which alignment parameters are known is straightforward.

	BLEU	NIST	PER	WER	METEOR	Hours
Giza++ U	26.0 (0.2)	8.10 (0.2)	42.8 (2.0)	64.8 (2.4)	53.5 (0.7)	339
Giza++ GDF	26.9 (0.3)	8.09 (0.1)	42.5 (1.0)	64.9 (1.0)	54.4 (0.1)	339
Giza++ R	25.0 (0.2)	7.91 (0.09)	43.2 (0.7)	66.4 (0.6)	53.5 (0.2)	339
25 BM	26.0 (0.2)	8.14 (0.1)	41.7 (0.7)	64.3 (1.2)	54.0 (0.5)	88+45
25 full	26.3 (0.1)	8.18 (0.15)	41.8 (0.9)	63.6 (1.4)	53.6 (0.4)	141+45
50 BM	27.0 (0.2)	8.15 (0.07)	42.6 (0.6)	64.8 (0.8)	54.4 (0.2)	157+45
50 full	26.7 (0.3)	8.19 (0.06)	42.0 (0.6)	63.8 (0.7)	53.9 (0.4)	203+45

Table 7.10: AR→EN United Nations Results. U, GDF and R refer respectively to the following symmetrization methods: union, grow-diag-final and Och and Ney's refined.

	BLEU	NIST	PER	WER	METEOR
Giza++ U	21.4 (0.4)	7.15 (0.13)	48.8 (2.0)	71.8 (2.2)	51.4 (0.4)
Giza++ GDF	22.5 (0.7)	7.41 (0.2)	46.8 (2.4)	70.7 (2.4)	52.9 (0.4)
25 BM	23.2 (0.1)	7.76 (0.03)	43.0 (0.3)	66.6 (0.4)	52.8 (0.1)
25 full	23.2 (0.5)	7.59 (0.2)	44.9 (2.4)	68.5 (2.5)	52.8 (0.5)
50 BM	23.8 (0.1)	7.77 (0.1)	43.4 (1.1)	67.1 (1.3)	53.2 (0.2)
50 full	23.5 (0.2)	7.69 (0.09)	43.8 (0.9)	67.6 (1.1)	52.9 (0.2)

Table 7.11: AR→EN United Nations Results on the 50000 sentences sub-corpus.

In order to see how the aligner weights tuned on a small part of the training corpus scaled to the entire corpus, we compare the results of Table 7.10, for which SMT systems were trained on the entire corpus, with results for SMT systems trained on the small part. For each system[10] in Table 7.10, the sentences (and alignment) corresponding to the 50000 sentences sub-corpus were picked, and translation units were extracted from them. Results are presented in Table 7.11. Alignment weights of "50 BM" and "50 full" systems were thus tuned on the entire 50000 sentences sub-corpus. Since the 25000 sentences sub-corpus is included into the 50000 sentences one, we can say than weights of "25 BM" and "25 full" systems were tuned on half of the total sub-corpus considered for SMT training. Looking at Giza++ results, we can evaluate to what extent the small corpus is more

[10]Except the refined symmetrisation, which was worse than the other Giza++ combinations.

sensitive to alignment quality than the large corpus. The 1.1 BLEU, 0.26 NIST and 1.5 METEOR average score differences between union and GDF combinations observed on the 50000 sentences part are respectively reduced to 0.9, 0.01 and 0.9 on the entire corpus. Comparing BIA results corresponding to the 25k and 50k sentences systems, it can be noticed that tuning alignment weights on only half of the corpus produces a small loss of the average translation metric scores (except for WER and PER, but these two metrics are so noisy in this experiment that we don't know if some conclusions should be drawn from their scores). Now comparing systems coming from BIA and Giza++ alignments, an important result is that the "50 BM" system is better than Giza++ GDF system. This is mainly the case for BLEU and NIST scores, with a 1.3 average BLEU improvement (and still 0.5 points considering the worst case given by the standard deviation), and a 0.36 average NIST improvement. This result is another evidence that when the alignment system weights are optimised on the entire corpus, BIA can achieve better alignments, in terms of the resulting SMT system, than with Giza++ combinations, at least for small corpora. When weights are optimised on only a small part of the training corpus, as it is the case in Table 7.10, the difference between the same two systems is cancelled. Although alignment system weights optimised on the corpus subset did not perfectly scale to the whole corpus, we think that they nicely scaled to the whole corpus since we still obtained alignments able to generate an SMT system as good as for the best Giza++ combination. This is an important result because our method is much more flexible, and requires less computational resources than Giza++ training.

Automated translation metric results for the EPPS task are shown in Table 7.12. Although the system "25 full" has got a better score than "25 BM" for each translation metric, the differences are not statistically significant. The same can be concluded from the comparison with Giza++ systems. The differences lie within the error given by standard deviation, and all systems can actually be considered as of equivalent quality.

	AER	BLEU	NIST	PER	WER	Hours
Giza++ U	15.2	49.1 (0.1)	9.83 (0.02)	31.4 (0.1)	40.9 (0.2)	114
Giza++ GDF	14.4	49.5 (0.3)	9.84 (0.04)	31.5 (0.2)	40.6 (0.3)	114
Giza++ R	14.5	49.2 (0.3)	9.80 (0.05)	31.4 (0.3)	40.6 (0.3)	114
25 BM	20.9 / 18.9	48.8 (0.3)	9.81 (0.08)	31.4 (0.4)	40.6 (0.6)	73+36
25 full	23.0 / 19.1	49.3 (0.2)	9.88 (0.01)	31.1 (0.05)	40.3 (0.06)	133+36

Table 7.12: Automatic translation results on EPPS Corpus (English-to-Spanish). 25 refers to the size of the sub-corpus (in thousand sentence pairs) used to train the aligner weights.

In this particular task, total training time is of the same order as for Giza++. Re-

markably, Giza++ training lasted three times less than for the UN-a corpus to align only a 15% of sentences less. This must be due to the difference in vocabulary size. In contrast, training time with our method was in the same order as for the UN-a task. However, if the corpus was larger, or if the alignment parameters were already known, our training method would be faster.

The AER in the alignment development set of the alignments corresponding to each system is also indicated for completeness, although this metric was not used in this experiment (the maximisation criterion for alignment parameter optimisation was BLEU score). The two values for our systems stand for "first pass / second pass" alignments AER.

7.6 Conclusions and Further Work

We proposed a novel framework for discriminative training of alignment models with automated translation metrics as maximization criterion. On small corpora, according to this type of metrics, the translation systems trained from the optimized alignments clearly performed better than the ones trained from Giza++ alignment combinations.

We successfully applied this framework to large corpora. We trained the alignment parameters on a small subset of the training corpus and used these parameters to align the whole corpus. The obtained set of parameters scaled nicely to the entire corpus since the translation systems built from the corresponding alignments were as good as the system built with alignments produced by Giza++ training, for two different tasks.

This alignment technique is very flexible, requires no labelled data at the word level and requires very low average memory usage. It is also faster than Giza++ for large corpora or when the alignment system parameters are already known.

In these experiments, we optimised the BLEU score of translations produced by an n-gram SMT system, but any other SMT system could be used.

We also found out interesting results for discriminative alignment systems. We saw that the Likelihood Ratio statistics was not adequate for alignment because rare bilingual events are not noise in most cases, as opposed to events occurring in a monolingual corpus. In contrast, association score models derived from IBM1 probabilities give very good results, and the corresponding training time increase is reasonable. We also implemented an improved search strategy.

In the future we would like to analyse the differences between alignments produced after optimisation in function of AER or BLEU score. We also plan to study more sys-

tematically the effect of the size of the sub-corpus used to train alignment parameters, as well as the effect of the sentences selection method.

We think the alignment system might be improved in various ways. We could try to use IBM model 2, or an improved IBM model 1 [Moo04] as association model for the first pass. In the second pass, when relative link probabilities are calculated, we could consider link probabilities between groups of words instead of only words. We should see the effect of the main features and parameters according to translation metrics and not AER. We could also try to symmetrise BIA alignment following source and target sentences with the grow-diag-final method instead of our "intelligent" union.

This research work was also reported in the following publication:

- [Lam07] Patrik Lambert, Rafael Banchs and Josep M. Crego. 2007. **Discriminative Alignment Training without Annotated Data for Machine Translation**. In Proc. of the North American Chapter of the Association for Computational Linguistics, Human Language Technologies Conference (NAACL-HLT), pp 85–88. Rochester, NY, USA, April 23-25.

Chapter 8

Conclusions and Further Work

In this last chapter, we summarise the main results and contributions of this thesis' research, and we discuss directions for future work.

8.1 Conclusions

A first contribution of this work was the elaboration of resources for word alignment and machine translation. We carried out a study on word alignment evaluation. In this study, we observed the impact of the evaluation procedure on the metric scores. In particular, we stressed that the ratio of Sure and Possible links in the test data has got a great influence on the alignment error rate: low Sure/Possible link ratios favour high precision alignments, whereas high Sure/Possible link ratios favour high recall alignments. This statement relativises many recent word alignment results, which were obtained calculating the alignment error rate of the system against test data with a particularly low Sure/Possible link ratio. We wrote guidelines for alignment evaluation, including guidelines for the manual alignment task. We provided freely available word alignment test data and software to handle a corpus aligned at the word level.

The general goal of the thesis was to improve statistical machine translation techniques. In particular, the research was performed with the n-gram-based machine translation system developed in our group. As detailed in Chapter 4, some components of this system are a contribution of this thesis. In many international evaluations, this system has proved to be competitive with other statistical machine translation approaches.

We presented experiments in which an attempt was made to improve translation qual-

ity keeping the same translation system, but processing the corpus. The impact of these corpus transformations on alignment quality were also evaluated. We observed that stemming and verb classification have a large positive impact on alignment error rate in a Spanish-English task. However, we also reported that this better alignment does not yield better translations.

Another type of preprocessing performed was the data-inferred extraction, the detection and substitution of multi-word expressions before alignment. In the small, "clean" Verbmobil corpus, results were encouraging. However, with "real-life" data, the data-inferred extraction method used was probably too noisy to yield any improvement. Nevertheless, after performing a detailed error analysis, we believe that when the considered multi-word expressions were fixed expressions which could not be translated word to word, grouping them before alignment training helped for their correct translation in test.

We raised the issue of the experimental error caused by the minimum error training strategy. This problem makes it difficult to measure small improvements of the system. We proposed to use the Simultaneous Perturbation Stochastic Approximation (SPSA) as optimisation algorithm and presented experiments in which the variability of the results is reduced with respect to using the downhill simplex method.

Another direction of research to improve minimum error training is the choice of the metric used as objective function in the optimisation process. We presented an alternative solution to a strategy based on the BLEU score alone. Our approach is based on the QUEEN metric of the Qarla framework. Instead of relying on a single evaluation metric, or in an ad-hoc linear combination of metrics, this method works over metric combinations with a stable and robust criterion. By means of a manual evaluation, we showed that this strategy was better, in terms of translation quality, than using the BLEU metric alone. The disagreement between conventional metric scores and manual evaluation has shown one more evidence of the need for this type of methodology.

Finally, after having highlighted the ambiguity of the alignment evaluation task and after having observed a lack of correlation between the alignment error rate and translation quality, we proposed a framework to train word alignment directly in function of automated translation metrics. Following a discriminative approach, we implemented a log-linear combination of alignment models which were designed to suit the characteristics of the n-gram-based translation system. We also presented interesting results for discriminative alignment systems: improved search strategy and comparison of various probabilities for the association model. The model coefficients were trained in a minimum translation error strategy with the SPSA optimisation algorithm. In this approach, no

link labels at the word level are needed.

On a small corpus, we adjusted word alignment parameters over the entire corpus. This method clearly allowed to improve translation quality. We successfully applied this framework to large corpora. We trained the alignment parameters on a small subset of the training corpus and used these parameters to align the whole corpus. The obtained set of parameters scaled nicely to the entire corpus since the translation systems built from the corresponding alignments were as good as the system built with alignments produced by Giza++ training, for two different tasks.

This alignment technique is very flexible and requires very low average memory usage. It is also faster than Giza++ for large corpora or when the alignment system parameters are already known. It represents a step further for the introduction of statistical machine translation into systems used for real life applications.

8.2 Further Work

We explored the possibility of treating multi-word expressions as a unique "super-token" during alignment training. After performing an error analysis, we believe that this technique is useful when a clean list of multi-word expressions is extracted, in the sense of expressions which could not be translated word to word. Thus filtering techniques for our data-inferred extraction method should be improved, and different methods for extracting and identifying multi-word expressions must be developed and evaluated. Maybe methods that focus on the extraction of specific types of multi-word expressions would be more effective than a general method as the one presented here. Resources build manually, like Wordnet multi-word expressions, should also be considered. The method we proposed considers the bilingual multi-words as units ; the use of each side of the bilingual multi-word expressions as independent monolingual multi-words must also be considered and evaluated.

As for the optimisation process, more tuning experiments with the IQmt framework should be performed, on other data sets and language pairs, such as Chinese-English and Arabic-English. Systematic experiments should also be performed to confirm the results obtained with the SPSA algorithm. Furthermore, SPSA being expected to perform better for a problem of higher dimensionality, experiments with a system including more feature functions must be carried out.

The proposed framework for discriminative training of alignment models in function of automated translation metrics is an encouraging starting point for various improvements

and extensions:

- We plan to study more thoroughly the effect of the size of the sub-corpus used to train alignment parameters
- We plan to study the effect of the sentences selection method.
- The information provided by clean bilingual multi-word expression dictionaries should be included in the discriminative alignment system. Verb classification for highly inflected languages and stemming would also be useful for our new alignment system.
- This framework could be used to optimise alignment for a phrase-based SMT system

Appendix A

Association Measures Derivation Details

In this appendix we give derivation details of some formula introduced in Chapter 2.

A.1 Chi-squared test

Here we derive Formula A.6 from Formula 2.23:

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{A.1}$$

Using $E_{ij} = R_i C_j / N$, $R_i = O_{i1} + O_{i2}$, $C_j = O_{1j} + O_{2j}$, $N = \sum_{i,j} O_{ij}$, we can transform the difference present in the denominator of Equation A.1:

$$\begin{aligned} O_{11} - E_{11} &= O_{11} - \frac{R_1 C_1}{N} \\ &= \frac{O_{11}(O_{11} + O_{12} + O_{21} + O_{22}) - (O_{11} + O_{12})(O_{11} + O_{21})}{N} \\ &= \frac{O_{11}O_{22} - O_{12}O_{21}}{N} \end{aligned} \tag{A.2}$$

Similarly,

$$O_{12} - E_{12} = \frac{-O_{11}O_{22} + O_{12}O_{21}}{N}$$

Thus the following identity holds for any contingency table:

$$(O_{11} - E_{11})^2 = (O_{12} - E_{12})^2 = (O_{21} - E_{21})^2 = (O_{22} - E_{22})^2 \tag{A.3}$$

With this identity, we can factor the numerator in Equation A.1 out from the sum:

$$\begin{aligned}\chi^2 &= (O_{11} - E_{11})^2 \sum_{i,j} \frac{1}{E_{ij}} \\ &= N(O_{11} - E_{11})^2 \frac{R_2C_2 + R_2C_1 + R_1C_2 + R_1C_1}{R_1R_2C_1C_2}\end{aligned}$$

Since $R_1 + R_2 = C_1 + C_2 = N$, we obtain the following simpler form of Equation A.1:

$$\chi^2 = \frac{N(O_{11} - E_{11})^2}{E_{11}E_{12}} \tag{A.4}$$

Using Equation A.2, this result can also be rewritten as follows:

$$\chi^2 = \frac{N(O_{11}O_{22} - O_{12}O_{21})^2}{R_1R_2C_1C_2} \tag{A.5}$$

Substituting by the value of R_i and C_j, we obtain Equation A.6.

$$\chi^2 = \frac{N(O_{11}O_{22} - O_{12}O_{21})^2}{(O_{11} + O_{12})(O_{21} + O_{22})(O_{11} + O_{21})(O_{12} + O_{22})} \tag{A.6}$$

A.2 Likelihood Ratio Test

The likelihood ratio test is based on the ratio between the maximum likelihood of the observed data under the null hypothesis and its unconstrained maximum likelihood (without making any assumptions about the parameters). Thus the log-likelihood ratio (LLR) has the following expression:

$$LLR = -2 \log \frac{\max P_{H_0}(\vec{X} = \vec{O} \mid N)}{\max P(\vec{X} = \vec{O} \mid N)} \tag{A.7}$$

The probability of a contingency table with cell values $k_{11}, k_{12}, k_{21}, k_{22}$ is given by the multinomial distribution:

$$P(\vec{X} = \vec{k} \mid N) = N! \prod_{i,j} \frac{p_{ij}^{k_{ij}}}{k_{ij}!}, \tag{A.8}$$

where the p_{ij} are the probability parameters corresponding to each cell of the contingency table. In Equation A.7, the numerator's likelihood function is maximized with the probability parameters corresponding to the null hypothesis, *i.e.* $p_{ij} = E_{ij}/N$. Maximum-likelihood estimates for the denominator are $p_{ij} = O_{ij}/N$. Replacing by these values in the expression of Equation A.8, the likelihood ratio can be written as in Equation A.9.

$$LLR = 2 \sum_{i,j} O_{ij} \log \frac{O_{ij}}{E_{ij}} \tag{A.9}$$

Replacing E_{ij} by its value under the point null hypothesis, as in §A.1, we obtain:

$$LLR = 2 \sum_{i,j} O_{ij} \log \frac{NO_{ij}}{(O_{i1} + O_{i2})(O_{1j} + O_{2j})} \tag{A.10}$$

Appendix B

Detailed Spanish-English Manual Alignment Guidelines

B.1 Phrasal constructions

Idiomatic expressions should be linked as a whole, but also other types of fixed expressions, that are unique to one language or the other. Examples include "at least"–"por lo menos", "let's see"–"a ver", "the thing is"–"lo que pasa".

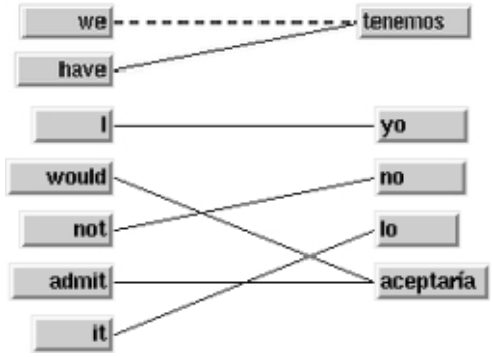
Sometimes a phrase can be considered as a fixed expression but a correspondence between elements of the phrase is also acceptable. In these cases the sub-elements are S-linked and the many-to-many alignment is completed with "Possible" links, as in the example below.

B.2 General verb constructions

B.2.1 Verbs and personal pronouns

If we divide verb expressions into two parts, namely personal pronoun (if any), and rest of words (including auxiliary verbs), each of these parts should be dealt as an indivisible group, and we should link personal pronouns in Spanish and English with a Sure link, and the group of remaining words in Spanish and English with Sure links too. If the pronoun is omitted in Spanish, Possible links should be introduced between the personal pronoun

in English and the remaining form in Spanish (and vice-versa). Below are two examples:



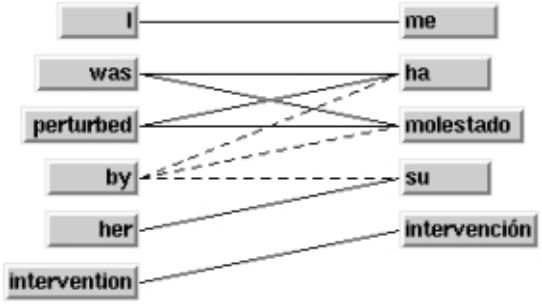
B.2.2 Verbs followed by a preposition

Phrasal verbs in English, or in general verbs where the preposition or adverb combined with the verb is part of the meaning of the verb, must be taken as a unit, i.e. the preposition or adverb cannot be separated from the verb. Ex: "to get on"–"subir", "to carry out"–"realizar".

The other verbs are considered to be distinct from the preposition and the alignment depend on the situation. For example, with "to wait-esperar", you could have a preposition in both languages ("wait for Maria"–"espera a Maria"), only in one ("wait for the bus"–"espera el autobÃºs") or in none of them ("wait a second"–"espera un segundo"). In the first case, the two prepositions should be linked together. When the preposition is omitted in the other language, it can be linked to the verb with a Possible link.

B.2.3 Passive voice

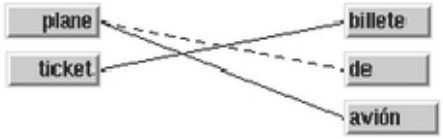
It is difficult to generalise the variety of passive voice examples, but in principle there is no problem in aligning words that represent the same action or object, but that correspond to a different function, tense or part of speech. An example is given in the following figure.



B.3 General noun constructions

B.3.1 Noun complement constructions

In constructions like "plane ticket"–"billete de avión" or "ceiling fan"–"ventilador en el techo", the omitted prepositions in the complement can be linked or not to the noun:

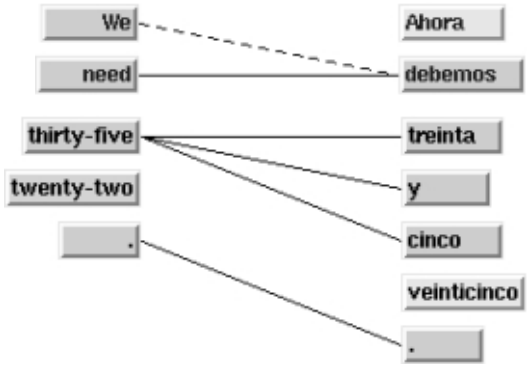


B.3.2 Enumerating

When enumerating in Spanish, a determiner is often added, like "el" in "it's K F seven five R six"–"es el K F siete cinco R seis". In these cases the determiner should remain without link.

B.4 Words not or incorrectly translated

They must remain unlinked (see "ahora" or "twenty-two"–"veinticinco" in the examples).



Some complementary words often occur (in particular in the Spanish translation, like "ya" in "it would be good"–"ya me irÃa bien" or "si que" in "yes we accept that"–"si que la aceptamos"). If these words have no correspondent and if removing them does not affect the meaning of the sentence, they can be left unlinked.

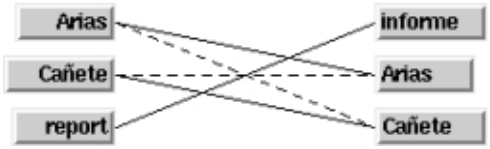
B.5 Repetitions

When a piece of text is repeated in a language but not in the other, Melamed [Mel98a] aligns all the repeated instances to the only instance of the other sentence. We prefer the alternative, which consists in aligning the first instance in each language and consider that the subsequent instances have no correspondence.

B.6 Special expressions

B.6.1 Proper nouns

The proper nouns of various words can be considered indivisible or linked word to word:

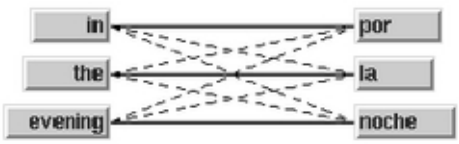


B.6.2 Numerals

A string representing a number (like "sixty-six" or "two hundred five") must be considered as a group. In numbers like "two hundred eighty eight"–"doscientos ochenta y ocho", we could establish the correspondences "two hundred"–"doscientos" and "eighty eight"–"ochenta y ocho". However, since it is all the same number, we link it as a group.

B.6.3 Date and time

Some constructions expressing date or time are rather fixed expressions (like "in the evening"–"por la noche", "nine o'clock"–"las nueve"). So we consider acceptable to link them as groups. Since often there is a clear correspondence between the words, we consider acceptable to link them only at the word level, so the final reference alignment should be:

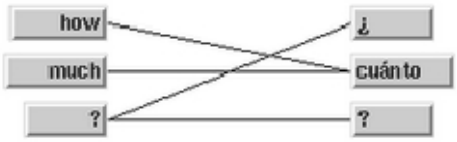


Typically time is expressed with a preposition (at, before, about, etc) followed by the time itself (twelve, one o'clock). In general these parts remain separated in the alignment:

B.7 Punctuation marks

Punctuation can give rise to unintuitive links, for instance when a dot is translated by a semi-column, or a coma is translated by a coordination conjunction like "and". If both parts of the link represent an accurate translation, they should be S-linked. If they represent an acceptable translation but not an usual way of translating them (for instance, "and" or "or" is usually not an accurate translation of a comma), you should link them with a P link.

In Spanish, exclamation and interrogation marks are double. You should align both parts to the corresponding English mark.



Although some expressions are always followed by a punctuation mark (like the question tags), the expression and the punctuation mark(s) form two distinct units.

References

[Ahr98] Lars Ahrenberg, Mikael Andersson, and Magnus Merkel, "A simple hybrid aligner for generating lexical correspondences in parallel texts", *Proc. of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL'98)*, pags. 29–35, Montreal, Canada, August 10–14 1998.

[Ahr00] Lars Ahrenberg, Magnus Merkel, Anna Sågvall Hein, and Jörg Tiedemann, *Proc. of the 2nd International Conference on Linguistic Resources and Evaluation (LREC)*, pags. Vol. III: 1255–1261, Athens, Greece, 31 May – 2 June 2000.

[Als96] H. Alshawi, "Head automata for speech translation", *Proc. of the 4th Int. Conf. on Spoken Language Processing, ICSLP'06*, pags. 2360–2364, October 1996.

[Ami05] E. Amigó, J. Gonzalo, A. Peñas, and F. Verdejo, "QARLA: A framework for the evaluation of text summarization systems", *Proc. of the 43rd Annual Meeting of the Association for Computational Linguistics*, pags. 280–289, Ann Arbor, Michigan, June 2005.

[Ami06] E. Amigó, J. Giménez, J. Gonzalo, and L. Màrquez, "MT evaluation: Human-like vs. human acceptable", *Proc. of COLING-ACL*, pags. 17–24, Sydney, Australia, July 2006.

[Arr03] V. Arranz, N. Castell, and J. Giménez, "Development of language resources for speech-to-speech translation", *Proc. of the International Conference on Recent Advances in Natural Language Processing (RANLP)*, Borovets, Bulgary, September, 10-12 2003.

[Arr04] Victoria Arranz, Núria Castell, Josep M. Crego, Jesús Giménez, Adrià de Gispert, and Patrik Lambert, "Bilingual Connections for Trilingual Corpora: An XML Approach", *Proc. of the 4th International Conference on Linguistic Resources and Evaluation (LREC)*, Lisbon, Portugal, May 26-28 2004.

[Aya06] Necip Fazil Ayan, and Bonnie J. Dorr, "Going Beyond AER: An Extensive Analysis of Word Alignments and Their Impact on MT", *Proc. of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pags. 9–16, Sydney, Australia, July 2006.

[Ban01] S. Bangalore, and G. Riccardi, "A finite state approach to machine translation", *Proc. IEEE ASRU*, Madonna di Campiglio, Italy, 2001.

[Ban05a] Rafael E. Banchs, Josep M. Crego, Adrià de Gispert, Patrik Lambert, and José B. Mariño, "Statistical machine translation of Euparl data by using bilin-

gual n-grams", *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pags. 133–136, Ann Arbor, Michigan, June 2005.

[Ban05b] S. Banerjee, and A. Lavie, "METEOR: An automatic metric for mt evaluation with improved correlation with human judgments", pags. 65–72, Ann Arbor, MI, June 2005.

[Ban06] Rafael E. Banchs, Josep M. Crego, Patrik Lambert, and José B. Mariño, "A Feasibility Study For Chinese-Spanish Statistical Machine Translation", *Proc. of the 5th Int. Symposium on Chinese Spoken Language Processing (ISCSLP)CONLL*, pags. 681–692, Kent Ridge, Singapore, December 13–16 2006.

[Bea95] S. Beale, S. Nirenburg, and K. Malesh, "Semantic analysis in the Mikrokosmos machine translation project", *Proc. of the 2nd Symposium on Natural Language Processing*, Bangkok, Thailand, August 1995.

[Ben85] W.S. Bennett, and J. Slocum, "The LRC machine translation system", *Computational Linguistics*, Vol. 11, pags. 111–121, 1985.

[Ber96] Adam L. Berger, Stephen A. Della Pietra, and Vincent J. Della Pietra, "A maximum entropy approach to natural language processing", *Computational Linguistics*, Vol. 22, nº 1, pags. 39–72, March 1996.

[Ber06] Nicola Bertoldi, "Minimum error training (updates)", Slides of the JHU Summer Workshop (http://www.statmt.org/jhuws), 2006.

[Blu06] Phil Blunsom, and Trevor Cohn, "Discriminative word alignment with conditional random fields", *Proc. of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pags. 65–72, Sydney, Australia, July 2006.

[Bra00] T. Brants, "Tnt — a statistical part-of-speech tagger", *Proc. of Applied Natural Language Processing (ANLP)*, Seattle, WA, 2000.

[Bro90] Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin, "A statistical approach to machine translation", *Computational Linguistics*, Vol. 16, nº 2, pags. 79–85, 1990.

[Bro93] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer, "The mathematics of statistical machine translation: Parameter estimation", *Computational Linguistics*, Vol. 19, nº 2, pags. 263–311, 1993.

[Car04] Xavier Carreras, I. Chao, L. Padró, and M. Padró, "Freeling: An open-source suite of language analyzers", *Proc. of the 4th International Conference on Linguistic Resources and Evaluation (LREC)*, Lisbon, Portugal, May 2004.

[Cas01] F. Casacuberta, "Finite-state transducers for speech input translation", *Proc. IEEE ASRU*, Madonna di Campiglio, Italy, 2001.

[Cas04] F. Casacuberta, and E. Vidal, "Machine translation with inferred stochastic finite-state transducers", *Computational Linguistics*, Vol. 30, nº 2, pags. 205–225, 2004.

[CB04] Ch. Callison-Burch, D. Talbot, and M. Osborne, "Statistical machine translation with word- and sentence-aligned parallel corpora", *Proc. of the 42th Annual Meeting of the Association for Computational Linguistics*, pags. 176–183, July 2004.

[CB06] C. Callison-Burch, M. Osborne, and P. Koehn, "Re-evaluating the role of bleu in machine translation research", *Proc. of the 11th Conference of the European Chapter of the ACL (EACL)*, pags. 249–256, Trento, Italy, 2006.

[CB07] C. Callison-Burch, C. Fordyce, P. Koehn, C. Monz, and J. Schroeder, "(meta-) evaluation of machine translation", *Proc. of the Second Workshop on Statistical Machine Translation*, pags. 136–158, Prague, 2007.

[Cet05] Mauro Cettolo, Marcello Federico, Nicola Bertoldi, Roldano Cattoni, and Boxing Chen, "A look inside the itc-irst smt system", *Proc. of Machine Translation Summit X*, pags. 451–457, Phuket, Thailand, September 2005.

[Che77] M. Chevalier, J. Dansereau, and G. Poulin, "TAUM-METEO: Description du système.", Tech. rep., Groupe de recherche en Traduction Automatique, Université de Montréal, 1977.

[Che03] Colin Cherry, and Dekang Lin, "A probability model to improve word alignment", *Proc. of the 41th Annual Meeting of the Association for Computational Linguistics*, 2003.

[Che05] B. Chen, R. Cattoni, N. Bertoldi, M. Cettolo, and M. Federico, "The ITC-irst SMT System for IWSLT-2005", *IWSLT*, pags. 98–104, 2005.

[Che06a] Boxing Chen, and Marcello Federico, "Improving phrase-based statistical translation through combination of word alignment", *Proc. of FinTAL - 5th International Conference on Natural Language Processing*, Turku, Finland, 2006.

[Che06b] C. Cherry, and D. Lin, "Soft syntactic constraints for word alignment through discriminative training", pags. 105–112, Association for Computational Linguistics, Sydney, Australia, July 2006.

[Cre05a] J. M. Crego, A. de Gispert, and J. Mariño, "The TALP ngram-based SMT system for IWSLT'05", *Proc. of the Int. Workshop on Spoken Language Translation, IWSLT'05*, pags. 191–198, Pittsburgh, USA, October 2005.

[Cre05b] J. M. Crego, J. Mariño, and A. de Gispert, "A ngram-based statistical machine

translation decoder", *Proc. of the 9th European Conf. on Speech Communication and Technology (Interspeech)*, pags. 3185–88, Lisbon, Portugal, 2005.

[Cre05c] J.M. Crego, J.B. Mariño, and A. de Gispert, "Reordered search and tuple unfolding for ngram-based smt", pags. 283–89, Phuket, Thailand, September 2005.

[Cre06a] J.M. Crego, and J.B. Mariño, "Reordering experiments for n-gram-based smt", Palm Beach, Aruba, December 2006.

[Cre06b] Josep M. Crego, Adrià de Gispert, Patrik Lambert, Marta R. Costa-jussà, Maxim Khalilov, Rafael Banchs, José B. Mariño, and José A. R. Fonollosa, "N-gram-based smt system enhanced with reordering patterns", *Proc. of the HLT-NAACL Workshop on Statistical Machine Translation*, pags. 162–165, Association for Computational Linguistics, New York City, June 2006.

[Cre06c] Josep M. Crego, Adrià de Gispert, Patrik Lambert, Maxim Khalilov, Marta R. Costa-jussà, José B. Mariño, Rafael Banchs, and José A.R. Fonollosa, "The TALP Ngram-based SMT System for IWSLT 2006", *Proc. of the International Workshop on Spoken Language Translation*, pags. 116–122, Kyoto, Japan, 2006.

[Dem77] A.P. Dempster, N.M.Laird, and D.B.Rubin, "Maximum likelihood from incomplete data via the EM algorithm", *Journal Royal Stat. Soc., Series B*, Vol. 39, nº 1, pags. 1–38, 1977.

[Dia02] Mona Diab, and Philip Resnik, "An unsupervised method for word sense tagging using parallel corpora", *Proc. of the Annual Meeting of the Association for Computational Linguistics*, pags. 255–262, Philadelphia, PA, 2002.

[Dod02] G. Doddington, "Automatic evaluation of machine translation quality using n-gram co-occurrence statistics", 2002.

[Dun93] Ted Dunning, "Accurate methods for the statistics of surprise and coincidence", *Computational Linguistics*, Vol. 19, nº 1, pags. 61–74, 1993.

[DY01] Grace Ngai David Yarowsky, and Richard Wicentowski, "Inducing multilingual text analysis tools via robust projection across aligned corpora", *Proc. of the 1st International Conference on Human Language Technology Research (HLT)*, pags. 161–168, 2001.

[Eck05] M. Eck, and Ch. Hori, "Overview of the IWSLT 2005 Evaluation Campaign", pags. 11–32, Pittsburgh, USA, October 2005.

[Eve05] Stefan Evert, *The Statistics of Word Cooccurrences: Word Pairs and Collocations*, PhD Thesis, Institut für maschinelle Sprachverarbeitung, University of Stuttgart, 2005.

[Far91] D. Farwell, and Y. Wilks, "ULTRA: A multi-lingual machine translator.", *Proc. of Machine Translation Summit III*, pags. 19–24, Washington DC, USA, 1991.

[For07] C. Fordyce, "Overview of the IWSLT 2007 Evaluation Campaign", *IWSLT07*, pags. 1–12, Trento, Italy, 2007.

[Fra05] A. Fraser, and D. Marcu, "Isi's participation in the romanian-english alignment task", *Proc. of the ACL Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond*, pags. 91–94, June 2005.

[Fra06] Alexander Fraser, and Daniel Marcu, "Semi-supervised training for statistical word alignment", *Proc. of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pags. 769–776, Sydney, Australia, July 2006.

[Fun94] Pascale Fung, and Kathleen McKeown, "Aligning noisy parallel corpora across language groups: Word pair feature matching by dynamic time warping", *Proc. of the Conference of the Association for Machine Translation in the Americas*, pags. 81–88, Columbia, Maryland, 1994.

[Gal91] W. Gale, and K. W. Church, "Identifying word correspondences in parallel texts", *DARPA Speech and Natural Language Workshop*, Asilomar, CA, 1991.

[Gal04] M. Galley, and Hopkins M., "What's in a translation rule?", *HLTNAACL04*, pags. 273–280, Boston, MA, May 2004.

[Ger01] U. Germann, M. Jahr, K. Knight, D. Marcu, and K. Yamada, "Fast decoding and optimal decoding for machine translation", pags. 228–235, July 2001.

[Gil03] D. Gildea, "Loosely tree-based alignment for machine translation", *ACL03*, pags. 80–87, Sapporo, Japan, July 2003.

[Gim05] J. Giménez, E. Amigó, and C. Hori, "Machine translation evaluation inside qarla", *Proceedings of the International Workshop on Spoken Language Technology (IWSLT'05)*, pags. 199–206, 2005.

[Gim06] J. Giménez, and E. Amigó, "Iqmt: A framework for automatic machine translation evaluation", *Proceedings of the 5th LREC*, pags. 685–690, 2006.

[Gis] Adrià de Gispert, Deepa Gupta, Maja Popovic, Patrik Lambert, José B. Mariño, Marcello Federico, Hermann Ney, and Rafael E. Banchs, "Improving statistical word alignments with morpho-syntactic transformations", .

[Gis02] 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.

[Gis04a] A. de Gispert, and J. Mariño, "Talp: Xgram-based spoken language translation system", *Proc. of the Int. Workshop on Spoken Language Translation,*

IWSLT'04, pags. 85–90, October 2004.

[Gis04b] 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*, pags. 107–114, October 2004.

[Gis05] A. de Gispert, "Phrase linguistic classification and generalization for improving statistical machine translation", pags. 67–72, Ann Arbor (Michigan), June 2005.

[Gis06] Adrià de Gispert, *Introducing Linguistic Knowledge into Statistical Machine Translation*, PhD Thesis, Signal Theory and Communications Department, Universitat Politècnica de Catalunya (UPC), 2006.

[Gra04] J. Graehl, and K. Knight, "Training tree transducers", *HLTNAACL04*, pags. 105–112, Association for Computational Linguistics, Boston, Massachusetts, USA, May 2 - May 7 2004.

[GV02] Ismael García Varea, Franz Josef Och, Hermann Ney, and Francisco Casacuberta, "Improving alignment quality in statistical machine translation using context-dependent maximum entropy models", *Proc. 19thInt. Conf. on Computational Linguistics*, pags. 1051–1054, Taipei,Taiwan, 2002.

[GV03] I. García Varea, *Traducción automática estadística: modelos de traducción basados en máxima entropía y algoritmos de búsqueda*, PhD Thesis in Informatics, Dep. de Sistemes Informàtics i Computació, Universitat Politècnica de València, 2003.

[Hew05] S. Hewavitharana, B. Zhao, A.S. Hildebrand, M. Eck, C. Hori, S. Vogel, and A. Waibel, "The CMU SMT System for IWSLT 2005", *IWSLT*, pags. 63–70, 2005.

[Hut92] W.J. Hutchins, and H.L. Somers, *An Introduction to Machine Translation*, Academic Press, 1992.

[Itt05] Abraham Ittycheriah, and Salim Roukos, "A maximum entropy word aligner for arabic-english machine translation", *Proc. of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pags. 89–96, Vancouver, British Columbia, Canada, October 2005.

[Jur00] D. Jurafsky, and J.H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics.*, Prentice-Hall, 2000.

[Kne95] Kneser, and Ney, "Improved backing-off for m-gram language modeling", *IEEE Inte. Conf. on Acoustics, Speech and Signal Processing*, pags. 49–52, Detroit, MI, May 1995.

[Kni99] K. Knight, "Decoding complexity in word replacement translation models", *Computational Linguistics, Squibs and Discussion*, Vol. 26, nº 2, pags. 607–615, 1999.

[Koe] Philipp Koehn, Amittai Axelrod, Alexandra Birch Mayne, Chris Callison-Burch, Miles Osborne, and David Talbot, .

[Koe03] P. Koehn, F.J. Och, and D. Marcu, "Statistical phrase-based translation", *Proc. of the 41th Annual Meeting of the Association for Computational Linguistics*, 2003.

[Koe04] Ph. Koehn, "Pharaoh: a beam search decoder for phrase-based statistical machine translation models", pags. 115–124, October 2004.

[Koe06] Ph. Koehn, and C. Monz, "Manual and automatic evaluation of machine translation between european languages", *Proceedings of the Workshop on Statistical Machine Translation*, pags. 102–21, Association for Computational Linguistics, New York City, June 2006.

[Koe07] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst, "Moses: Open source toolkit for statistical machine translation", *Proc. of the 45th Annual Meeting of the Association for Computational Linguistics (Demo and Poster Sessions)*, pags. 177–180, Association for Computational Linguistics, Prague, Czech Republic, June 2007.

[Kuh04] Jonas Kuhn, "Experiments in parallel-text based grammar induction", *Proc. of the 42th Annual Meeting of the Association for Computational Linguistics*, pags. 470–477, Barcelona, Spain, July 2004.

[Lam] Patrik Lambert, Marta R.Costa-jussà, Josep M. Crego, Maxim Khalilov, José B. Mariño, Rafael E. Banchs, José A. R.Fonollosa, and Holger Schwenk, "The talp ngram-based smt system for iwslt 2007", .

[Lam04] P. Lambert, and N. Castell, "Alignment of parallel corpora exploiting asymmetrically aligned phrases", *Proc. of the LREC 2004 Workshop on the Amazing Utility of Parallel and Comparable Corpora*, Lisbon, Portugal, May 25 2004.

[Lam05a] P. Lambert, and R. Banchs, "Data inferred multi-word expressions for statistical machine translation", *Proc. of Machine Translation Summit X*, pags. 396–403, Phuket, Tailandia, 2005.

[Lam05b] P. Lambert, A. de Gispert, R. Banchs, and J. Mariño, "Guidelines for word alignment evaluation and manual alignment", *Language Resources and Evaluation*, Vol. 39, nº 4, pags. 267–285, 2005.

[Lam06a] Patrik Lambert, and Rafael E. Banchs, "Grouping multi-word expressions according to part-of-speech in statistical machine translation.", *Proc. of the EACL Workshop on Multi-Word-Expressions in a Multilingual Context*, pags. 9–16, Trento, Italy, 2006.

[Lam06b] Patrik Lambert, and Rafael E. Banchs, "Tuning machine translation parameters with SPSA", *Proc. of the Int. Workshop on Spoken Language Translation, IWSLT'06*, pags. 190–196, Kyoto, Japan, 2006.

[Lam06c] Patrik Lambert, Jesús Giménez, Marta R. Costa-jussa, Enrique Amigó, Rafael E. Banchs, Lluís Màrquez, and J.A. R Fonollosa, "Machine translation system development based on human likeness", *Proc. of the IEEE/ACL Workshop on Spoken Language Technology*, pags. 246–249, Palm Beach, Aruba, 2006.

[Lam07] Patrik Lambert, Rafael E. Banchs, and Josep M. Crego, "Discriminative alignment training without annotated data for machine translation", *Proc. of the Human Language Technology Conference of the NAACL*, Rochester, NY, USA, 2007.

[Lia06] Percy Liang, Ben Taskar, and Dan Klein, "Alignment by agreement", *Proc. of the Human Language Technology Conference of the NAACL*, pags. 104–111, New York City, USA, June 2006.

[Liu05] Yang Liu, Qun Liu, and Shouxun Lin, "Log-linear models for word alignment", *Proc. of the 43rd Annual Meeting of the Association for Computational Linguistics*, pags. 459–466, Ann Arbor, Michigan, June 2005.

[Man99] C. Manning, and H. Schütze, *Foundations of Statistical Natural Language Processing*, MIT Press, Cambridge, MA, 1999.

[Mar02] Daniel Marcu, and William Wong, "A phrase-based, joint probability model for statistical machine translation", *Proc. of the Conference on Empirical Methods in Natural Language Processing*, Philadelphia, PA, July 6-7 2002.

[Mar05] Joel Martin, Rada Mihalcea, and Ted Pedersen, "Word alignment for languages with scarce resources", *Proc. of the ACL Workshop on Building and Using Parallel Texts*, pags. 65–74, Ann Arbor, Michigan, June 2005.

[Mar06a] D. Marcu, Wong. W, A. Echihabi, and K. Knight, "Spmt: Statistical machine translation with syntactified target language phrases", pags. 44–52, Sydney, Australia, July 2006.

[Mar06b] José B. Mariño, Rafael E. Banchs, Josep M. Crego, Adrià de Gispert, Patrik Lambert, José A.R. Fonollosa, and Marta R. Costa-jussà, "N-gram based machine translation", *Computational Linguistics*, Vol. 32, nº 4, pags. 527–549, 2006.

[Mel98a] I. Dan Melamed, "Annotation style guide for the blinker project", Tech. Rep. 98-06, IRCS, 1998.

[Mel98b] I. Dan Melamed, "Manual annotation of translational equivalence", Tech. Rep. 98-07, IRCS, 1998.

[Mel00] I. Dan Melamed, "Models of translational equivalence among words", *Computational Linguistics*, Vol. 26, nº 2, pags. 221–249, 2000.

[Mel04] D. Melamed, "Statistical machine translation by parsing", *Proc. of the 42th Annual Meeting of the Association for Computational Linguistics*, pags. 653–661, July 2004.

[Mih03] Rada Mihalcea, and Ted Pedersen, "An evaluation exercise for word alignment", Rada Mihalcea, Ted Pedersen (eds.), *HLT-NAACL 2003 Workshop: Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, pags. 1–10, Edmonton, Alberta, Canada, May 31 2003.

[Mil91] G.A. Miller, R. Beckwith, C. Fellbaum, D. Gross, K. Miller, and R. Tengi, "Five papers on wordnet", *Special Issue of International Journal of Lexicography*, Vol. 3, nº 4, pags. 235–312, 1991.

[Mn06] J.B. Mariño, R.E. Banchs, J.M. Crego, A. de Gispert, P. Lambert, J.A.R Fonollosa, M.R. Costa-jussà, and M. Khalilov, "Upc's bilingual n-gram translation system", *Proc. of the TC-STAR Workshop on Speech-to-Speech Translation*, pags. 43–48, Barcelona, Spain, June 2006.

[Moo04] Robert C. Moore, "Improving ibm word alignment model 1", *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, pags. 518–525, Barcelona, Spain, July 2004.

[Moo05a] Robert C. Moore, "Association-based bilingual word alignment", *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pags. 1–8, Association for Computational Linguistics, Ann Arbor, Michigan, June 2005.

[Moo05b] Robert C. Moore, "A discriminative framework for bilingual word alignment", *Proc. of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pags. 81–88, Vancouver, British Columbia, Canada, October 2005.

[Moo06] R.C. Moore, W. Yih, and A. Bode, "Improved discriminative bilingual word alignment", pags. 513–520, Association for Computational Linguistics, Sydney, Australia, July 2006.

[Mos] D. Mostefa, O. Hamon, N. Moreau, and K. Choukri, .

[Nel65] J.A. Nelder, and R. Mead, "A simplex method for function minimization", *The Computer Journal*, Vol. 7, pags. 308–313, 1965.

[Ney05] Hermann Ney, Volker Steinbiss, Richard Zens, Evgeny Matusov, Jorge González, Young-Suk Lee, Salim Roukos, Marcello Federico, Muntsin Kolss, and Rafael E. Banchs, "SLT progress report, tc-star deliverable d5", European Community project no. FP6-506738, 2005, available on line at: http://www.tc-star.org/pages/f_documents.htm.

[Ney06] H. Ney, "Overview of the SLT evaluation: Spoken language translation", *Proc. of the TC-STAR Workshop on Speech-to-Speech Translation*, Barcelona, Spain, June 2006.

[Och99a] Franz Josef Och, "An efficient method for determining bilingual word classes", *Proc. of the 9th Conference of the European Chapter of the ACL (EACL)*, pags. 71–76, Association for Computational Linguistics, 1999.

[Och99b] Franz Josef Och, Christoph Tillmann, and Hermann Ney, "Improved alignment models for statistical machine translation", *Proc. of the Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pags. 20–28, University of Maryland, College Park, MD, June 1999.

[Och00a] Franz Josef Och, "Giza++: Training of statistical translation models", http://www.isi.edu/~och/GIZA++.html, 2000.

[Och00b] Franz Josef Och, and Hermann Ney, "A comparison of alignment models for statistical machine translation", *Proc. of the 18th Int. Conf. on Computational Linguistics*, pags. 1086–1090, Saarbrucken,Germany, August 2000.

[Och00c] Franz Josef Och, and Hermann Ney, "Improved statistical alignment models", *Proc. of the 38th Annual Meeting of the Association for Computational Linguistics*, pags. 440–447, Hongkong, China, October 2000.

[Och01] F.J. Och, N. Ueffing, and H. Ney, "An efficient A* search algorithm for statistical machine translation", *Data-Driven Machine Translation Workshop, 39th Annual Meeeting of the Association for Computational Linguistics (ACL)*, pags. 55–62, July 2001.

[Och02] F.J. Och, and H. Ney, "Dicriminative training and maximum entropy models for statistical machine translation", *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics*, pags. 295–302, Philadelphia, PA, July 2002.

[Och03a] F.J. Och, "Minimum error rate training in statistical machine translation", *Proc. of the 41th Annual Meeting of the Association for Computational Linguistics*, pags. 160–167, 2003.

[Och03b] 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, "Syntax for

statistical machine translation", Tech. Rep. Summer Workshop Final Report, Johns Hopkins University, Baltimore, USA, 2003.

[Och03c] F.J. Och, and H. Ney, "A systematic comparison of various statistical alignment models", *Computational Linguistics*, Vol. 29, nº 1, pags. 19–51, March 2003.

[Och04a] 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*, pags. 161–168, May 2004.

[Och04b] F.J. Och, and H. Ney, "The alignment template approach to statistical machine translation", *Computational Linguistics*, Vol. 30, nº 4, pags. 417–449, December 2004.

[Ort05] D. Ortiz, I. García-Varea, and F. Casacuberta, "Thot: a toolkit to train phrase-based statistical translation models", pags. 141–148, Phuket, Thailand, September 2005.

[Pap98] 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*, pags. 189–192, Seattle, WA, May 1998.

[Pap01] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, "Bleu: a method for automatic evaluation of machine translation", IBM Research Report, RC22176, September 2001.

[Pau06] Michael Paul, "Overview of the IWSLT 2006 Evaluation Campaign", *Proc. of the Int. Workshop on Spoken Language Translation, IWSLT'06*, pags. 1–15, Kyoto, Japan, 2006.

[Ped03] Ted Pedersen, and Brian Rassier, "Aligner for parallel corpora", http://www.d.umn.edu/~tpederse/parallel.html, 2003.

[Pow64] M. J. D. Powell, "An efficient method for finding the minimum of a function of several variables without calculating derivatives", *The Computer Journal*, Vol. 7, pags. 155–162, 1964.

[Pre02] W.H. Press, S. Teukolsky, W. Vetterling, and B. Flannery, *Numerical Recipes in C++: the Art of Scientific Computing*, Cambridge University Press, 2002.

[Rib01] António Ribeiro, Gabriel Lopes, and João Mexia, "Extracting translation equivalents from portuguese-chinese parallel texts", *Journal of Studies in Lexicography*, Vol. 11, nº 1, pags. 118–194, 2001.

[Rui05] M. Ruiz, and J.A. Fonollosa, "Improving phrase-based statistical translation by modifying phrase extraction and including several features", *(to be published) ACL05 workshop on Building and Using Parallel Corpora: Data-driven*

Machine Translation and Beyond, 2005.

[She04] L. Shen, A. Sarkar, and F.J. Och, "Discriminative reranking for machine translation", pags. 177–184, Association for Computational Linguistics, Boston, Massachusetts, USA, May 2004.

[Sma96] Frank A. Smadja, Kathleen R. McKeown, and Vasileios Hatzivassiloglou, "Translating collocations for bilingual lexicons: A statistical approach", *Computational Linguistics*, Vol. 22, nº 1, pags. 1–38, 1996.

[Sno05] M. Snover, B. Dorr, R. Schwartz, J. Makhoul, L. Micciula, and R. Weischedel, "A study of translation error rate with targeted human annotation", Tech. Rep. LAMP-TR-126,CS-TR-4755,UMIACS-TR-2005-58, University of Maryland, College Park and BBN Technologies, July 2005.

[Spa92] James C. Spall, "Multivariate stochastic approximation using a simultaneous perturbation gradient approximation", *IEEE Trans. Automat. Control*, Vol. 37, pags. 332–341, 1992.

[Spa98a] James C. Spall, "Implementation of the simultaneous perturbation algorithm of stochastic optimization", *IEEE Trans. Aerospace and Electronic Systems*, Vol. 34, nº 3, pags. 817–823, 1998.

[Spa98b] James C. Spall, "An overview of the simultaneous perturbation method for efficient optimization", *Johns Hopkins APL Technical Digest*, Vol. 19, nº 4, pags. 482–492, 1998.

[Sto02] A. Stolcke, "SRILM: an extensible language modeling toolkit.", *Proc. of the Int. Conf. on Spoken Language Processing*, pags. 901–904, Denver, CO, 2002.

[Tas05] Ben Taskar, Lacoste-Julien Simon, and Klein Dan, "A discriminative matching approach to word alignment", *Proc. of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pags. 73–80, Vancouver, British Columbia, Canada, October 2005.

[Tie03] Jörg Tiedemann, "Combining clues for word alignment", *Proc. of the 10th Conference of the European Chapter of the ACL (EACL)*, Budapest, Hungary, April 12–17 2003.

[Til00] C. Tillmann, and H. Ney, "Word re-ordering and dp-based search in statistical machine translation", pags. 850–856, Saarbrucken, Germany, July 2000.

[Til03a] Tillmann, and Xia, "A phrase-based unigram model for statistical machine translation", *Proc. of the Human Language Technology Conference of the NAACL*, Edmonton, Canada, 2003.

[Til03b] C. Tillmann, "A projection extension algorithm for statistical machine translation", *Proc. of the Conf. on Empirical Methods in Natural Language Processing,*

EMNLP'03, pags. 1–8, July 2003.

[Tom76] P. Toma, "An operational machine translation system", R.W. Brislin (ed.), *Translation: Applications and Research*, pags. 247–259, Gardner Press, New York, 1976.

[Tou02] Kristina Toutanova, H. Tolga Ilhan, and Christopher D. Manning, "Extensions to hmm-based statistical word alignment models", *Proc. of the Conference on Empirical Methods in Natural Language Processing*, Philadelphia, PA, July 6-7 2002.

[Tur03] J.P. Turian, L. Shen, and I.D. Melamed, "Evaluation of machine translation and its evaluation", *Proteus technical report 03-005*, 2003.

[VÓ0] Jean Véronis, and Philippe Langlais, "Evaluation of parallel text alignment systems: the arcade project.", *Parallel text processing: Alignment and use of translation corpora*, pags. 369–388, Kluwer Academic Publishers, 2000.

[Ven03] A. Venugopal, S. Vogel, and A. Waibel, "Proc. of the 41th annual meeting of the association for computational linguistics", *Effective Phrase Translation Extraction from Alignment Models*, pags. 319–326, 2003.

[Vid97] E. Vidal, "Finite-state speech-to-speech translation", *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, pags. 111–114, April 1997.

[Vil06] David Vilar, Maja Popovic, and Hermann Ney, "AER: Do we need to "improve" our alignments?", *Proc. of the Int. Workshop on Spoken Language Translation, IWSLT'06*, pags. 205–212, Kyoto, Japan, 2006.

[Vog96] S. Vogel, H. Ney, and C. Tillmann, "HMM-based word alignment in statistical translation", *COLING'96: The 16thInt. Conf. on Computational Linguistics*, pags. 836–841, Copenhagen, Denmark, August 1996.

[VR79] C. J. Van Rijsbergen, *Information Retrieval, 2nd edition*, London, Butterworths, 1979.

[Wan98] Ye-Yi Wang, and Alex Waibel, "Modeling with structures in statistical machine translation", *Proc. of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, pags. 1357–1363, Montreal, Canada, 1998.

[Way05] A. Way, and N. Gough, "Comparing example-based and statistical machine translation", *Natural Language Engineering*, Vol. 11, nº 3, pags. 295–309, 2005.

[Wu97] D. Wu, "Stochastic inversion transduction grammars and bilingual parsing of parallel corpora", *Computational Linguistics*, Vol. 23, nº 3, pags. 377–403, September 1997.

[Yam01] Kenji Yamada, and Kevin Knight, "A syntax-based statistical translation

model", *Proc. of the Annual Meeting of the Association for Computational Linguistics*, Toulouse, France, 2001.

[Zen02] R. Zens, F.J. Och, and H. Ney, "Phrase-based statistical machine translation", Springer Verlag (ed.), *Proc. German Conference on Artificial Intelligence (KI)*, september 2002.