

Rich Linguistic Knowledge for Empirical Machine Translation

Jesús Ángel Giménez Linares
jgimenez@lsi.upc.edu

directors

Lluís Màrquez Villodre i Núria Castell Ariño

Memòria del DEA i Projecte de Tesi
Programa de Doctorat en Intel·ligència Artificial
Departament de Llenguatges i Sistemes Informàtics
Universitat Politècnica de Catalunya

Juny de 2005

Contents

1	Introduction	3
1.1	History of MT	3
1.2	Classification of MT systems	3
1.3	Why is MT so difficult?	4
1.4	Our Proposal	5
2	Statistical Machine Translation	9
2.1	Fundamentals	9
2.2	Phrase-based SMT	10
2.3	Use of Linguistic Knowledge in SMT	11
2.4	MT Evaluation Metrics	13
3	Our Approach	15
3.1	System Description	16
3.2	Combining Linguistic Data Views	17
3.2.1	Building Linguistic Data Views	17
3.2.2	Building Combined Translation Models	18
3.2.3	Using the MCR	18
3.2.4	Experimenting with Linguistic Data Views	21
3.2.5	Conclusions	22
3.3	Automatic Translation of WordNet Glosses	23
3.3.1	Experimental Setting	23
3.3.2	Results for the Baseline System	24
3.3.3	Improved Language Modeling	24
3.3.4	Using the MCR	27
3.3.5	Tuning the System	29
3.3.6	Conclusions	30
4	Work Plan	33
4.1	Development of Resources	33
4.1.1	Corpora Collection	33
4.1.2	Development of NLP Tools	34
4.2	Further Work	35

4.2.1	Minor Improvements	36
4.2.2	Research Prospectives	37
4.2.3	Project Scheduling	39
5	Publications	41
5.1	Our approach to MT	41
5.2	Developing NLP Tools	41
5.3	Generation of Resources	42
	References	45

Abstract

In this work, a new architecture for Empirical Machine Translation is suggested. We build a state-of-the art Phrase-based Statistical Machine Translation system and study different alternatives so as to provide it with rich linguistic knowledge. Natural Language Processing technology and external knowledge sources are suavely integrated into the system. We utilize existing resources and develop new ones.

We have tested our approach in two tasks: the Spanish-to-English translation of parliament proceedings and the English-to-Spanish translation of dictionary definitions. Initial results, based on the utilization of shallow syntactic analysis techniques and lexical databases, suggest that significant improvements may be attained by working with rich linguistic knowledge.

Further steps are devised in a detailed work plan. We intend to move from syntax onto semantics by incorporating information about word senses and semantic roles. Additionally, we consider also the possibility of working on a hybrid system combining the best properties of empirical-based and rule-based approaches. We also plan to work on new language pairs, particulary Catalan-English and Catalan-Spanish.

Chapter 1

Introduction

Machine Translation (MT) is the use of a computer to translate a message from one natural language to another. Many different approaches to MT, either from a linguistic or an empirical point of view, have been tried in the past. In this work, we intend to exploit the best of these two irreconcilable approaches. We suggest new methods to incorporate current Natural Language Processing (NLP) technology into a state-of-the-art empirical MT system.

1.1 History of MT

MT history¹ is as relatively long as history of Artificial Intelligence itself (Nirenburg et al., 2003). Interest in MT began in the early 1950's right after World War II. A first decade of euphoria was followed by nearly two decades of oblivion after the ALPAC report (ALPAC, 1966). This report recommended stopping investment in MT and investing in NLP technology instead.

In the last decade, progress achieved in Artificial Intelligence in general, and particularly in NLP since the days of the ALPAC report together with the applicability of empirical methods on large amounts of data, have motivated again an enormous interest in MT. Great investments are being made from governments worldwide, as it is the case, for very different reasons, of the European Union, the United States of America and China. In the current context of globalization, MT is considered a very important technology because it is intended to help humanity to cross language barriers between cultures.

1.2 Classification of MT systems

Approaches to MT may be classified with respect to several criteria. Regarding the degree of human interaction, MT systems may be classified in Machine-aided Human Translation (MAHT), Human-aided Machine Translation (HAMT) and Fully Automatic Machine Translation (FAMT).

According to the level of linguistic analysis that is performed MT systems may be classified in three big groups: direct, transfer, and interlingua. See Figure 1.1. In the *direct* approach a

¹Consult John Hutchins' website at <http://ourworld.compuserve.com/homepages/WJHutchins/> for an excellent review of MT history.

word-by-word or phrase-by-phrase replacement is performed. In the *transfer* approach the input is syntactically and/or semantically analyzed to produce an abstract representation from which the output is generated. The *interlingua* approach is similar to the latter but with the difference that the abstract representation is language independent and deeply detailed.

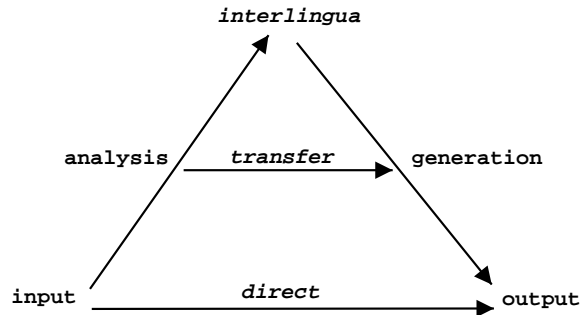


Figure 1.1: Classification of MT systems according to the level of linguistic analysis.

With respect to the core technology MT systems may be classified in two types: rule-based and empirical. In *rule-based systems*, a set of rules describing the translation process are specified by human experts. In contrast, *empirical systems* extract the knowledge of the translation process automatically from a collection of translation examples.

Figure 1.2 shows the architecture of an empirical MT system. A parallel corpus produced by a set of human translators is used to train knowledge models. At translation time, these models are used to suggest a set of possible translations for a given (pre-processed) input. MT is seen as the problem of deciding the best target text matching a given source text. At the end, the output may be post-processed.

1.3 Why is MT so difficult?

MT is considered an *NLP-complete/AI-complete* problem. Clearly, MT is difficult because of Natural Language complexity. Words in a sentence may have different meanings, and even when the meaning of all words is known, the meaning of the sentence may not be compositional. Still sentences may have different readings. Further these readings may have different interpretations in the context of the real world.

Natural Language presents different kinds of ambiguity:

Categorial ambiguity Words having more than one possible part-of-speech.

Word sense ambiguity Words having more than one possible meaning or sense.

Syntactic ambiguity Sentences having more than one possible syntactic parsing, leading to multiple alternative semantic interpretations.

Semantic ambiguity Sentences syntactically unambiguous having still different possible semantic interpretations.

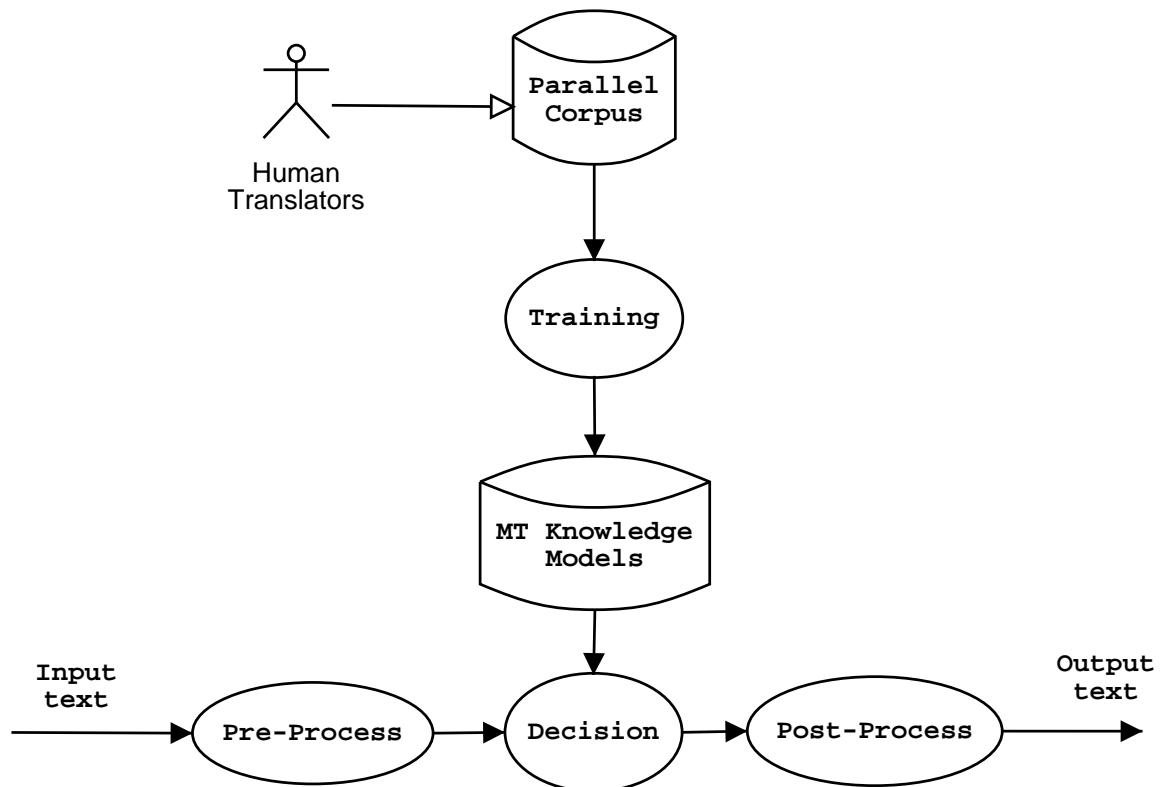


Figure 1.2: Architecture of an empirical MT system.

Referential ambiguity Anaphoric noun phrases having more than one possible referent.

Ellipsis Incomplete Sentences in which the missing constituent is not clear.

Pragmatic ambiguity When the meaning depends on the context of the current situation.

It is impossible to produce high quality translations without understanding the meaning of the message. Dorr (1994) presented an excellent report on MT divergences. Thus, High-quality Fully Automatic Machine Translation is still today an unrealistic scenario except for very restricted domains, as it is the case of the METEO system (Chandioux & Grimalia, 1996), which translates weather forecasts from English into French.

However, there are a number of good applications for current MT technology. For instance, MT is successfully used to aid human translation. Church and Hovy (1993) analyzed what requirements a good niche application for MT should meet.

1.4 Our Proposal

This work focuses on Fully Automatic Machine Translation of written Natural Language. By fully automatic we emphasize the fact that very light human interaction is required. We also distinguish

text translation from speech translation.

We suggest to exploit current NLP technology and knowledge with the intent to improve an *empirical MT* system. Among empirical approaches we focus on Statistical Machine Translation (SMT). This approach belongs to the family of direct translation. SMT is based on ideas borrowed from the work by Weaver (1955), and is very well fundamented from a theoretical viewpoint. Besides, with the availability of most of the components, SMT systems can be relatively rapidly developed (Knight et al., 1999). Moreover, once built, it is easy to train the system for a new language pair, provided the corresponding parallel corpus. But the main reason for selecting SMT is that it allows for obtaining very competitive results without using any linguistic information further than lexical. So, there seems to be plenty of room for improvement. Our golden assumption is that a system working with richer linguistic knowledge would be able to make better decisions.

However, to our knowledge, no significant improvement has been reported so far (Och et al., 2003). Some authors claim that linguistic knowledge is not useful at all for SMT. Others argue that evaluation metrics are not well suited to capture improvements attained, and they are right to a certain extent. But we, like many others, still believe that by doing things in a different manner linguistically rich knowledge models should significantly boost performance and quality of SMT systems.

Figure 1.3 shows the enrichments we suggest in the architecture of an empirical MT system. Again we would start from a parallel corpus produced by a set of human translators. Linguistic Processors would be used to annotate the corpus. This linguistically enriched corpus would be used to train knowledge models. At translation time, these models would suggest a set of possible translations for a given (linguistically) pre-processed input. Discriminative learning models could help the system to find the best translation. Finally, output could be (linguistically) post-processed. Additional external knowledge sources could be used at any stage.

In order to deploy such an architecture, first, we develop a number of NLP tools based on Machine Learning (ML) technology, such as part-of-speech taggers and shallow syntactic and semantic parsers. Second, we collect resources such as parallel corpora, dictionaries and lexical databases. These tools and resources² are integrated in a state-of-the-art phrase-based SMT system.

Initial results suggest that significant improvements may be attained by utilizing rich linguistic knowledge. Details are presented in Chapter 3.

Overview of the Document

The rest of the report is organized as follows. In Chapter 2 the fundamentals of SMT are described. We introduce phrase-based SMT and comment the state-of-the-art. We also open a discussion on automatic MT evaluation metrics. In Chapter 3 we suggest and deploy some linguistically motivated improvements. In Chapter 4 we sketch thesis research project, presenting work done so far and outlining further work. Finally, in Chapter 5 related publications are presented.

²See details of NLP resources and tools under development in Section 4.1.

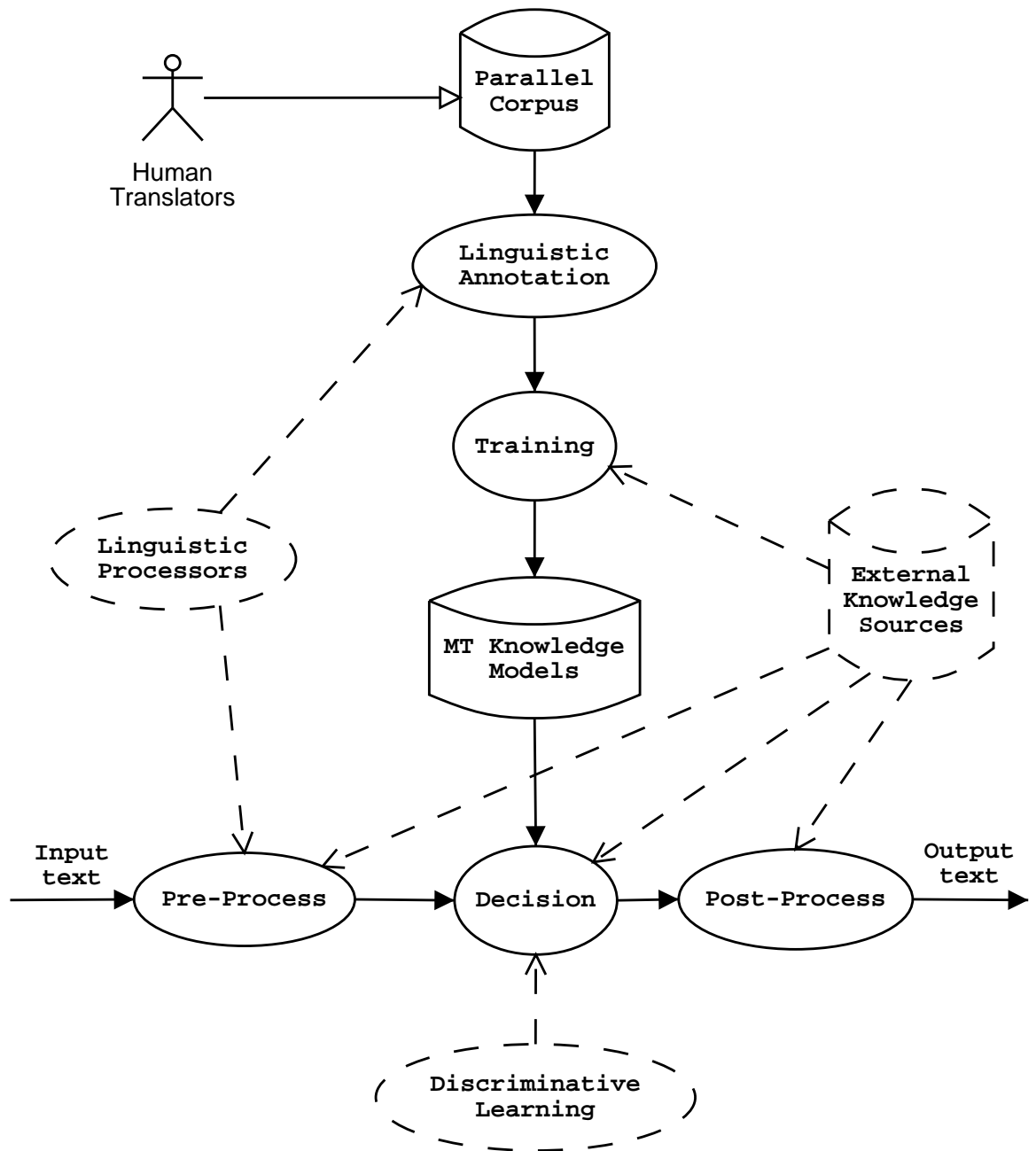


Figure 1.3: Architecture of a linguistically-aided empirical MT system.

Chapter 2

Statistical Machine Translation

Statistical Machine Translation is today a very promising approach to Machine Translation. SMT systems can be developed very quickly (Knight et al., 1999). Besides, SMT is fully automatic and results are also very competitive.

2.1 Fundamentals

Current state-of-the-art SMT systems are based on ideas borrowed from the Communication Theory field (Weaver, 1955). Brown et al. (1988; 1990) suggested that MT could be statistically approximated to the transmission of information through a *noisy channel*. Given a sentence $f = f_1..f_n$ (distorted signal), it is possible to approximate the sentence $e = e_1..e_m$ (original signal) which produced f . We need to estimate $P(e|f)$, the probability that a translator produces f as a translation of e .

By applying Bayes' rule we decompose it:

$$P(e|f) = \frac{P(f|e) * P(e)}{P(f)} \quad (2.1)$$

To obtain the string e which maximizes the translation probability for f , a search in the probability space must be performed. Because the denominator is independent of e , we can ignore it for the purpose of the search:

$$e = \operatorname{argmax}_e P(f|e) * P(e) \quad (2.2)$$

Equation 2.2 devises three components in a SMT. First, a *language model* that estimates $P(e)$. Second, a *translation model* representing $P(f|e)$. Last, a *decoder* responsible for performing the search.

The language model is typically estimated from a large monolingual corpus. The translation model is built out from a multilingual parallel corpus. In both cases, the corpus domain must be as closer as possible to the target domain. Also, the larger the corpora the greater statistical significance.

Regarding the search, performing an optimal decoding can be extremely costly because the search space is polynomial in the length of the input (Knight, 1999). Hence, most decoders perform a suboptimal search usually by introducing reordering constraints or by heuristically pruning the search space. MT decoding is a very active research topic. Among current approaches to decoding, we may find A* search (Och et al., 2001), integer programming (Germann et al., 2001), based on parsing (Yamada & Knight, 2002), greedy search (Germann, 2003), and beam search (Koehn, 2004).

An excellent and very detailed report on the mathematics of Machine Translation may be found at (Brown et al., 1993).

2.2 Phrase-based SMT

Word-based models as described by Brown et al. (1993) exhibit a main deficiency. The modeling of the context in which the words occur is very weak. This problem was considerably alleviated by phrase-based¹ models (Och, 2002). These models represent nowadays the state-of-the-art in SMT.

Fox (2002) showed that there are many regularities in phrasal movement. Words inside a phrase tend to stay together during translation. Phrase-based models outperform word-based ones because they capture phrasal cohesion in a very natural way, i.e. they allow *many-to-many* translations, thus taking local context into account. Besides, phrase-based models allow translation of non-compositional phrases.

Several approaches to phrase-based MT appeared in the last decade. Usually, phrase-based models are based on conditional probabilities, as detailed in Section 2.1. In contrast, Marcu and Wong (2002) proposed a phrase-based joint probability model. Their approach does not try to capture how source phrases are mapped into target phrases, but rather how source and target phrases can be generated simultaneously out from a bag of concepts.

Wang and Waibel (1998) were first to demonstrate the intuition shared with many other researchers that word-based alignment is a major cause of errors in MT. They proposed a new alignment model based on shallow phrase structures automatically acquired from a parallel corpus. At the same time, Alshawi et al. (1998) suggested a method for fully automatic learning of hierarchical finite state translation models. Phrases are modeled by the topology of the transducers. A third approach was suggested by Och et al. (1999). Their work on *alignment templates* may be reframed as a phrase-based translation system. They use phrases rather than single words as the basis for the alignment models. A group of adjacent words in the source sentence may be aligned to a group of adjacent words in the target. As a result, the local context has a greater influence.

All these models make an implicit use of syntactic knowledge. With the intent to make a explicit use of syntax, Yamada and Knight (2001) introduced a syntax-based statistical translation model in which a source-language parse tree is transformed in a target-language string through a series of stochastic operations at each node.

In our work, far from full syntactic complexity, we suggest to go back to the simpler alignment methods first described by Brown et al. (1993). We follow the approach by Koehn et al. (2003) in which phrase pairs are automatically induced from word alignments. In order to build phrase-based

¹The term '*phrase*' used hereafter refers to a sequence of words not necessarily syntactically motivated.

translation models, a phrase extraction must be performed on a word-aligned parallel corpus. To our knowledge, no hand-built word-aligned parallel corpora exist. However, a parallel corpus may be automatically aligned using word translation models ($P(f|e)$) estimated from it. Effective dynamic programming techniques, i.e. Viterbi algorithm, are often used to compute the word alignment which maximizes the global probability for each sentence.

We applied the phrase-extract algorithm described by Och (2002). This algorithm takes as input a word alignment matrix and outputs a set of phrase pairs that is *consistent* with it. A phrase pair is said to be consistent with the word alignment if all the words within the source phrase are only aligned to words within the target phrase, and viceversa.

There are several ways of assigning a probability to a phrase pair. In our case phrase pairs are scored by relative frequency. Let ph_f be a phrase in the source language (f) and ph_e a phrase in the target language (e). We define a function $count(ph_f, ph_e)$ which counts the number of times the phrase ph_f has been seen aligned to phrase ph_e in the training data. The conditional probability that ph_f maps into ph_e is estimated as:

$$P(ph_f|ph_e) = \frac{count(ph_f, ph_e)}{\sum_{ph_f} count(ph_f, ph_e)} \quad (2.3)$$

No smoothing is performed.

However, these phrase-based models have some limitations. First, non-contiguous phrases are not allowed. Second, long distance dependencies are not modeled. Third, syntactic transformations are not captured. We are bound to overcome these limitations by supplying linguistic knowledge to a phrase-based SMT system.

2.3 Use of Linguistic Knowledge in SMT

In the last years, many efforts have been devoted to including linguistic information in the parameter estimation of translation models.

In a first attempt, Och and Ney (2000) revised and improved the parameter estimation of word alignment models proposed by Brown et al. (1993) by introducing word classes automatically trained from parallel corpora. Although not really linguistically motivated, these word classes group together words that are realized in similar contexts.

Koehn and Knight (2002) proposed in their *ChunkMT* system to integrate two linguistic concepts, morphosyntactic analysis (part-of-speech tags) and shallow parsing (chunks). They obtained promising results. However, the applicability of their work was limited to very short sentences. They later abandoned this approach and focused on the particular case of noun phrase translation. They developed special modeling and features which integrated into a phrase-based SMT system (Koehn, 2003b).

Schafer and Yarowsky (2003) suggested a combination of models based on shallow syntactic analysis (part-of-speech tagging, lemmatization and phrase chunking). They followed a very interesting backoff strategy in the application of their models. Decoding was based on Finite State Automata. Although no significant improvement in MT quality was reported, results were promising taking into account the short time spent in the development of the linguistic tools utilized.

In the work by Koehn et al. (2003) syntactic information was used so as to limit the phrases in the translation model only to those syntactically motivated. But this actually proved to be harmful to the system performance.

Moving onto full parsing, Wu (1997) presented a novel stochastic inversion transduction grammar formalism for bilingual language modeling of sentence-pairs. They introduced the concept of bilingual parsing and applied it, among other tasks, to phrasal alignment. Recently, Melamed (2004) suggested a similar approach based on multitext grammars. MT is seen as a particular case of synchronous parsing in which the input can have fewer dimensions than the grammar. However, their approach is based on the use of multitreebanks² which are very expensive to build.

Yamada and Knight (2001) presented a syntax based tree-to-string probability model in which tree constituents are aligned to strings. Further details may be found in (Yamada, 2002). Gildea (2003) followed and improved this same idea by working on tree-to-tree alignments. In spite of their degree of sophistication these models do not achieve significant improvements on standard evaluation metrics. However, Charniak et al. (2003) presented a syntax-based language model based upon that described by Charniak (2001), which combined with the syntax based translation model described by Yamada and Knight (2001), achieved a notable improvement in grammaticality. They measured this improvement by means of human evaluation, though.

Zhang and Gildea (2004) made a direct comparison between syntactically supervised and unsupervised alignment models. They compared the model by Wu (1997) (unsupervised) to the model by Yamada and Knight (2001) (supervised). They concluded that automatically derived trees resulted in better agreement with human-annotated word-level alignments for unseen test data.

Gildea (2004) tried also working with dependency trees instead of constituents. They found out constituent trees to perform better. Lin (2004) proposed a path-based transfer model using dependency trees. They suggested a training algorithm that extracts a set of rules that transform a path in the source dependency tree into a fragment in the target dependency tree. Decoding was formulated as a graph-theoretic problem of finding the minimum path covering the source dependency tree. Results were under the performance of not syntactically motivated phrase-based models.

Finally, Shen et al. (2004) applied discriminative learning techniques to MT. They defined a *reranking* approach in which a system generated a series of n-best candidates which were then reranked according to a collection of linguistic features. The top ranked translation was selected as the system output. Och et al. (2004) suggested a smorgasbord of syntactically motivated features for reranking. They used more than 450 different feature functions to rerank 1000-best lists of candidates. However, only a small improvement was reported. They argued that linguistic processors introduce many errors and that the BLEU score is not specially sensitive to the grammaticality of MT output. Further details of this work may be found in (Och et al., 2003).

All in all, no significant improvements in MT quality have been reported so far. However, we still believe that linguistically guided phrase-based translation models should significantly outperform linguistically blind ones. This is the big challenge of our research project.

²A multitreebank is basically a multilingual parsed parallel corpus in which constituents are aligned.

2.4 MT Evaluation Metrics

Because human evaluation is very costly, MT researchers have developed several automatic evaluation metrics. The commonly accepted criterion that defines a plausible evaluation metric is that it must correlate well with human evaluators.

In the last decade, most successful evaluation metrics have been word error rate (WER), position independent word error rate (PER), bilingual evaluation understudy (BLEU) (Papineni et al., 2001), an improved version of BLEU by National Institute of Standards and Technology (NIST) (Lin & Hovy, 2002), the F-measure provide by the General Text Matcher (GTM) (Melamed et al., 2003), and the ROUGE metrics (Lin & Och, 2004a).

All these metrics work by rewarding lexical similarity (n-gram matches) among the system output and a set of reference translations. Therefore, they do not explicitly take into account linguistic criteria. As explained in Section 1.3, natural languages allow for many different ways of expressing the same idea. In order to capture this flexibility a very large number of translations would be required. Unfortunately often, as it is our case, only one reference translation is available.

input text	la casa verde estaba situada justo delante del lago .			
reference	the green house was right in front of the lake .			
		BLEU	GTM	NIST
output A	the green house was by the lake shore .	0.2954	0.7000	2.2940
output B	the green potato right in front of the lake was right .	0.5156	0.8696	2.8980
output C	a green house was by the lake shore .	0.0000	0.6000	1.9579

Table 2.1: Example of some deficiencies of automatic evaluation metrics.

Table 2.1 shows an example of bad translations getting higher scores than good translations. Highest scores are obtained by output *B*, which is wrong and totally absurd. Output *A*, which conveys most of the meaning of the input, attains much lower scores. As to output *C*, in which only the first word is changed with respect to output *A*, it scores a dramatic 0 BLEU score. Thus, in our opinion, metrics based on lexical similarity do not allow for a fair evaluation. However, building automatic MT evaluation metrics that account for Natural Language flexibility results again in an NLP-complete problem which is as hard as MT itself.

In our work we have chosen three different evaluation metrics, namely the GTM F-measure, the BLEU score, and the NIST score, which have proved to correlate well with both human adequacy and fluency. BLEU has become a ‘de facto’ standard nowadays in MT. Therefore, we discuss our results based on the BLEU score. However, it has several deficiencies that turn it impractical for error analysis (Turian et al., 2003). First, BLEU does not have a clear interpretation. Second, BLEU is not adequate to work at the segment³ level but only at the document level. Third, in order to punish candidate translations that are too long/short, BLEU computes a heuristically motivated word penalty factor.

In contrast, the GTM F-measure has an intuitive interpretation in the context of a bitext grid. It represents the fraction of the grid covered by aligned blocks. It also, by definition, works well

³A segment is the minimal unit of parallel text. It is usually the size of a sentence. It can be smaller (a word, a phrase) or bigger (a couple of sentences, a paragraph), though.

at the segment level and punishes translations too divergent in length. Therefore, we also analyze individual cases based on the GTM F-measure.

Evaluation of MT systems is a problem under permanent discussion. Recently, Babych and Hartley (2004) suggested extending BLEU with frequency weightings so as to account for the relevance of lexical items. In a very interesting work, Lin and Och (2004b) presented a method, called *ORANGE*, for evaluating automatic MT evaluation metrics.

We plan to perform several experiments regarding MT evaluation in order to capture subtle improvements introduced by linguistically-aided SMT. We also consider the idea of conducting very modest human evaluations.

Chapter 3

Our Approach

In this section we discuss the current state of our research. The whole research project is depicted in Chapter 4.

In the following, we describe the implementation of a state-of-the-art Phrase-based SMT system and discuss some of its deficiencies. We study the architecture of the system and select several points where to start working on improvements. We suggest and deploy some novel ideas.

In Section 3.2 we explain the work describing the SMT system with which we participate at the Shared Task 2: “Exploiting Parallel Texts for Statistical Machine Translation” of the Workshop¹ on “Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond”. This workshop is a satellite event of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005). We present our system to the task of translating parliament proceedings from Spanish into English. In our contribution we suggest to introduce linguistic information, other than lexical units, to the process of building word and phrase alignments. We successfully integrate shallow parsing NLP tools in the construction of phrase-based translation models for SMT.

In Section 3.3 we deploy the work submitted at the “Cross-Language Knowledge Induction Workshop” which takes place in the frame of the EuroLan 2005 Summer School. One of the main criticisms against SMT is that it is domain oriented. Since parameters are estimated from a parallel corpus in a specific domain, the performance of the system on a different domain is often much worse. We have studied the behaviour of our system when moving to a new domain. Our SMT system is applied to the translation of dictionary definitions from English into Spanish. We use additional external knowledge sources such as new corpora, dictionaries and lexical databases, so as to facilitate the adaptation.

We still leave a number of issues for further work. First, we are willing to explore the possibility of building *semantic models* as a new component of a SMT system. Second, we consider the possibility of applying Discriminative Learning techniques to post-process the output of our SMT system. Third, we suspect that hybrid solutions among rule-based and empirical approaches may lead to a notable gain in MT quality. Further work is outlined in Section 4.2.

¹Visit the Workshop website at <http://www.statmt.org/wpt05/>.

3.1 System Description

As explained in Section 2.1, a SMT system consists of three main components:

- Language Model ($P(e)$)
- Translation Model ($P(f|e)$)
- Search Algorithm

Fortunately, we can count on a number of freely available tools to build most of the components of such a system. We utilize these tools to build the first prototype of our SMT system. However, we do not discard to implement some of these components ourselves in the future if they become a limitation to the deployment of our ideas.

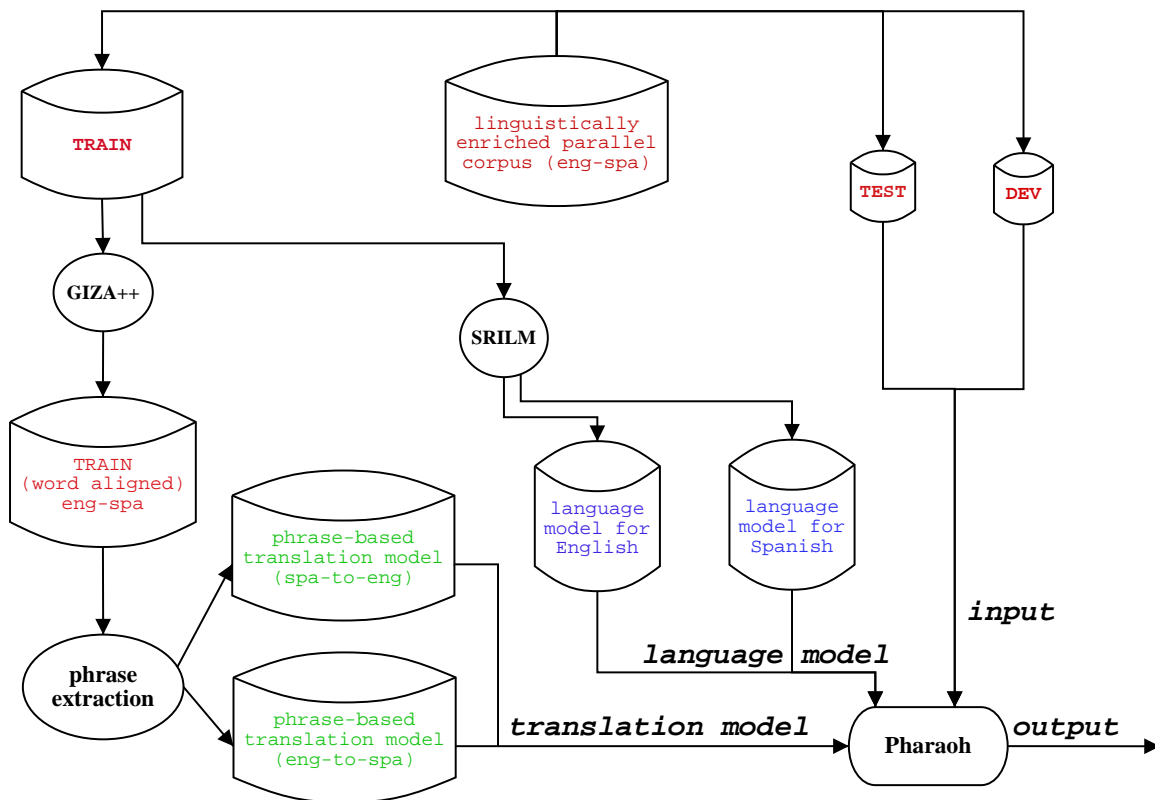


Figure 3.1: Architecture of our bidirectional Phrase-based SMT system.

See the architecture of our bidirectional² Spanish-English Phrase-based SMT system in Figure 3.1. First, a parallel corpus is linguistically enriched.³ This corpus is split into training, validation and test sets.

²Our SMT system works currently for translation from Spanish into English and from English into Spanish.

³See further details of NLP tools utilized to annotate the corpus in Subsection 4.1.2.

We use the GIZA++ SMT Toolkit⁴ (Och & Ney, 2003) to generate word alignments ($P(f|e)$). Phrase extraction is performed on top of these alignments as detailed in Section 2.2. Phrase-based translations models are built for both directions (Spanish-to-English and English-to-Spanish).

We utilize the *SRI Language Modeling Toolkit*⁵ (SRILM) (Stolcke, 2002) to build language models ($P(e)$). SRILM supports creation and evaluation of a variety of language model types based on N-gram statistics, as well as several related tasks, such as statistical tagging and manipulation of N-best lists and word lattices. Two language models have been built, one for English and one for Spanish.

For the search, we use the *Pharaoh*⁶ beam search decoder (Koehn, 2004). *Pharaoh* is an implementation of an efficient dynamic programming search algorithm with lattice generation and XML markup for external components. Because decoding allowing arbitrary reordering is an NP-complete problem (Knight, 1999), *Pharaoh* performs a suboptimal (beam) search by pruning the search space according to certain heuristics based on the translation cost.

3.2 Combining Linguistic Data Views

Our approach explores the possibility of working with alignments at two different levels of granularity, lexical (words) and shallow parsing (chunks). We try to exploit phrasal cohesion between Spanish and English to build chunk-based phrase alignments.

Apart from redefining the scope of the alignment unit, we may use different degrees of linguistic annotation. We introduce the general concept of *data view*, which is defined as any possible representation of the information contained in a bitext. We enrich data view tokens⁷ with features further than lexical such as *PoS*, *lemma*, and *chunk label*.

As an example of the applicability of data views, suppose the case of the word ‘*plays*’ being seen in the training data acting as a verb. Representing this information as ‘*plays_{V BZ}*’ would allow us to distinguish it from its homograph ‘*plays_{N N S}*’ for ‘*plays*’ as a noun.

Of course, there is a natural trade-off between the use of data views and data sparsity. Fortunately, we have data enough so that statistical parameter estimation remains reliable.

3.2.1 Building Linguistic Data Views

In order to build data views, training data are first linguistically annotated for the two languages involved. Segments are then PoS-tagged using the *SVMTool*, lemmatized using *Freeling*, and chunked using *Phreco*. See details of NLP tools utilized in Subsection 4.1.2. No additional tokenization or pre-processing steps other than case lowering have been performed. 10 different data views have been built. Notice that it is not necessary that the two parallel counterparts of a bitext share the same data view, as long as they share the same granularity. However, in all our experiments we have annotated both sides with the same linguistic information. See token descriptions: (W) word, (WL) word and lemma, (WP) word and PoS, (WC) word and chunk label, (WPC) word, PoS and chunk

⁴The GIZA++ SMT Toolkit may be freely downloaded at <http://www.fjoch.com/GIZA++.html>.

⁵The SRI Language Modeling Toolkit may be freely downloaded at <http://www.speech.sri.com/projects/srilm/download.html>.

⁶The Pharaoh beam search decoder may be freely downloaded at <http://www.isi.edu/licensed-sw/pharaoh/>.

⁷In order to avoid confusion so forth we will talk about *tokens* instead of *words* as the minimal alignment unit.

label, (Cw) chunk of words (Cwl), chunk of words and lemmas, (Cwp) chunk of words and PoS (Cwc) chunk of words and chunk labels (Cwpc) chunk of words, PoS and chunk labels. By chunk label we refer to the IOB label associated to every word inside a chunk, e.g. $'I_{B-NP} declare_{B-VP} resumed_{I-VP} the_{B-NP} session_{I-NP} of_{B-PP} the_{B-NP} European_{I-NP} Parliament_{I-NP} .O'$. We build chunk tokens by explicitly connecting words in the same chunk, e.g. $'(I)_{NP} (declare_resumed)_{VP} (the_session)_{NP} (of)_{PP} (the_European_Parliament)_{NP}'$. See examples of some of these data views in Table 3.1 and Table 3.2.

3.2.2 Building Combined Translation Models

Running *GIZA++*⁸, we obtain token alignments for each of the data views. Combined phrase-based translation models are built on top of the Viterbi alignments output by *GIZA++*. Because data views capture different, possibly complementary, aspects of the translation process it seems reasonable to combine them. We consider two different ways of building such combo-models:

LPHEX Local phrase extraction. To build a separate phrase-based translation model for each data view alignment, and then combine them. There are two ways of combining translation models:

MRG Merging translation models. We work on a weighted linear interpolation of models. These weights may be tuned, although a uniform weight selection yields good results. Additionally, phrase-pairs may be filtered out by setting a score threshold.

noMRG Passing translation models directly to the Pharaoh decoder. However, we encountered many problems with phrase-pairs that were not seen in all single models. This obliged us to apply arbitrary smoothing values to score these pairs.

GPHEX Global phrase extraction. To build a single phrased-based translation model from the union of alignments from several data views.

In its turn, any MRG operation performed on a combo-model results again in a valid combo-model. *Combo-models* must be then post-processed in order to remove the additional linguistic annotation and split chunks back into words, so they fit the format required by *Pharaoh*.

In any case, phrase extraction is performed as depicted by (Och, 2002). We always work with the union of alignments, no heuristic refinement, and phrases up to 5 tokens. Phrase pairs appearing only once have been discarded. Scoring is performed by relative frequency. No smoothing is applied.

3.2.3 Using the MCR

Moreover, we have used the Multilingual Central Repository (MCR), a multilingual lexical-semantic database (Atserias et al., 2004), to build a word-based translation model. We back-off to this model in the case of unknown words, with the goal of improving system recall. See further details in Subsection 3.3.4.

⁸We use the *GIZA++* default configuration (5 iterations for model 1, 4 iterations for model 3, 3 iterations for model 4, and 5 iterations for HMM model).

W	<p>It would appear that a speech made at the weekend by Mr Fischler indicates a change of his position .</p> <p>Fischler pronunció un discurso este fin de semana en el que parecía haber cambiado de actitud .</p>
WL	<p>It/It would/would appear/appear that/that a/a speech/speech made/make at/at the/the weekend/weekend by/by Mr/Mr Fischler/Fischler indicates/indicate a/a change/change of/of his/his position/position ./.</p> <p>Fischler/Fischler pronunció/pronunciar un/uno discurso/discurso este/este fin/fin de/de semana/semana en/en el/el que/que parecía/parecer haber/haber cambiado/cambiar de/de actitud/actitud ./.</p>
WP	<p>It_{PRP} would_{MD} appear_{VB} that_{IN} a_{DT} speech_{NN} made_{VBN} at_{IN} the_{DT} weekend_{NN} by_{IN} Mr_{NNP} Fischler_{NNP} indicates_{VBZ} a_{DT} change_{NN} of_{IN} his_{PRP} position_{NN} .</p> <p>Fischler_{VMN} pronunció_{VMI} un_{DI} discurso_{NC} este_{DD} fin_{NC} de_{SP} semana_{NC} en_{SP} el_{DA} que_{PRO} parecía_{VMI} haber_{VAN} cambiado_{VMP} de_{SP} actitud_{NC} ._{FP}</p>
WC	<p>It_{B-NP} would_{B-VP} appear_{I-VP} that_{B-SBAR} a_{B-NP} speech_{I-NP} made_{B-VP} at_{B-PP} the_{B-NP} weekend_{I-NP} by_{B-PP} Mr_{B-NP} Fischler_{I-NP} indicates_{B-VP} a_{B-NP} change_{I-NP} of_{B-PP} his_{B-NP} position_{I-NP} ._O</p> <p>Fischler_{B-VP} pronunció_{B-VP} un_{B-NP} discurso_{I-NP} este_{B-NP} fin_{I-NP} de_{B-PP} semana_{B-NP} en_{B-PP} el_{B-SBAR} que_{I-SBAR} parecía_{B-VP} haber_{I-VP} cambiado_{I-VP} de_{B-PP} actitud_{B-NP} ._O</p>
WPC	<p>It_[PRP:B-NP] would_[MD:B-VP] appear_[VB:I-VP] that_[IN:B-SBAR] a_[DT:B-NP] speech_[NN:I-NP] made_[VBN:B-VP] at_[IN:B-PP] the_[DT:B-NP] weekend_[NN:I-NP] by_[IN:B-PP] Mr_[NNP:B-NP] Fischler_[NNP:I-NP] indicates_[VBZ:B-VP] a_[DT:B-NP] change_[NN:I-NP] of_[IN:B-PP] his_[PRP\$:B-NP] position_[NN:I-NP] ._[.O]</p> <p>Fischler_[VMN:B-VP] pronunció_[VMI:B-VP] un_[DI:B-NP] discurso_[NC:I-NP] este_[DD:B-NP] fin_[NC:I-NP] de_[SP:B-PP] semana_[NC:B-NP] en_[SP:B-PP] el_[DA:B-SBAR] que_[PRO:I-SBAR] parecía_[VMI:B-VP] haber_[VAN:I-VP] cambiado_[VMP:I-VP] de_[SP:B-PP] actitud_[NC:B-NP] ._[FP:O]</p>

Table 3.1: Rich linguistic representation of a sentence pair, including 5 different word granularity data views: (W) word (WL) word and lemma (WP) word and PoS (WC) word and chunk label (WPC) word, PoS and chunk label.

Cw	(It) (would_appear) (that) (a_speech) (made) (at) (the_weekend) (by) (Mr_Fischler) (indicates) (a_change) (of) (his_position) (.) (Fischler) (pronunció) (un_discurso) (este_fin) (de) (semana) (en) (el_que) (parecía_haber_cambiado) (de) (actitud) (.)
Cwl	(It/It) (would/would_appear/appear) (that/that) (a/a_speech/speech) (made/make) (at/at) (the/the_weekend/weekend) (by/by) (Mr/Mr_Fischler/Fischler) (indicates/indicate) (a/a_change/change) (of/of) (his/his_position/position) (./.) (Fischler/Fischler) (pronunció/pronunciar) (un/uno_discurso/discurso) (este/este_fin/fin) (de/de semana/semana) (en/en) (el/el_que/que) (parecía/parecer_haber/haber_cambiado/cambiar) (de/de) (actitud/actitud) (./.)
Cwp	(It _{PRP}) (would _{MD} -appear _{VB}) (that _{IN}) (a _{DT} -speech _{NN}) (made _{VB} _N) (at _{IN}) (the _{DT} -weekend _{NN}) (by _{IN}) (Mr _{NNP} -Fischler _{NNP}) (indicates _{VBZ}) (a _{DT} -change _{NN}) (of _{IN}) (his _{PRP} -position _{NN}) (..) (Fischler _{VMN}) (pronunció _{VMI}) (un _{DI} -discurso _{NC}) (este _{DD} -fin _{NC}) (de _{SP}) (semana _{NC}) (en _{SP}) (el _{DA} -que _{PRO}) (parecía _{VMI} -haber _{VAN} -cambiado _{VMP}) (de _{SP}) (actitud _{NC}) (.Fp)
Cwc	(It _{PRP:B-NP}) (would _{B-VP} -appear _{I-VP}) (that _{B-SBAR}) (a _{B-NP} -speech _{I-NP}) (made _{B-VP}) (at _{B-PP}) (the _{B-NP} -weekend _{I-NP}) (by _{B-PP}) (Mr _{B-NP} -Fischler _{I-NP}) (indicates _{B-VP}) (a _{B-NP} -change _{I-NP}) (of _{B-PP}) (his _{B-NP} -position _{I-NP}) (.O) (Fischler _{B-VP}) (pronunció _{B-VP}) (un _{B-NP} -discurso _{I-NP}) (este _{B-NP} -fin _{I-NP}) (de _{B-PP}) (semana _{B-NP}) (en _{B-PP}) (el _{B-SBAR} -que _{I-SBAR}) (parecía _{B-VP} -haber _{I-VP} -cambiado _{I-VP}) (de _{B-PP}) (actitud _{B-NP}) (.O)
Cwpc	(It _{PRP:B-NP}) (would _[MD:B-VP] -appear _[VB:I-VP]) (that _[IN:B-SBAR]) (a _[DT:B-NP] -speech _[NN:I-NP]) (made _[VBN:B-VP]) (at _[IN:B-PP]) (the _[DT:B-NP] -weekend _[NN:I-NP]) (by _[IN:B-PP]) (Mr _[NNP:B-NP] -Fischler _[NNP:I-NP]) (indicates _[VBZ:B-VP]) (a _[DT:B-NP] -change _[NN:I-NP]) (of _[IN:B-PP]) (his _[PRP\$:B-NP] -position _[NN:I-NP]) (.[.:O]) (Fischler _[VMN:B-VP]) (pronunció _[VMI:B-VP]) (un _[DI:B-NP] -discurso _[NC:I-NP]) (este _[DD:B-NP] -fin _[NC:I-NP]) (de _[SP:B-PP]) (semana _[NC:B-NP]) (en _[SP:B-PP]) (el _[DA:B-SBAR] -que _[PRO:I-SBAR]) (parecía _[VMI:B-VP] -haber _[VAN:I-VP] -cambiado _[VMP:I-VP]) (de _[SP:B-PP]) (actitud _[NC:B-NP]) (.[Fp:O])

Table 3.2: Rich linguistic representation of a sentence pair, including 5 different chunk granularity data views: (Cw) chunk of words (Cwl) chunk of words and lemmas (Cwp) chunk of words and PoS (Cwc) chunk of words and chunk labels (Cwpc) chunk of words, PoS and chunk labels.

3.2.4 Experimenting with Linguistic Data Views

Results using several phrase-based translation models obtained from linguistic data views suggested in Subsection 3.2.1 are reported. We have used the data⁹ sets and language model provided by the organizers of the shared task. No extra training or development data were used in our experiments.

Table 3.3 presents MT results for the 10 elementary data views devised. Default parameters are used to control the importance of the translation model (λ_{tm}), the language model (λ_{lm}), and word penalty (λ_w) during decoding. Default decoder parameters were used. A -ttable-limit of 20, a -beam-threshold of 0.00001, maximum beam size -stack of 100, no limit one -ttable-threshold neither on -distortion-limit. We did not work on a model for reordering. No tuning has been performed. As expected, word-based views obtain significantly higher results than chunk-based. All data views at the same level of granularity obtain comparable results.

data view	GTM-1	GTM-2	BLEU	NIST
W	0.6108	0.2609	25.92	7.1576
WL	0.6110	0.2601	25.77	7.1496
WP	0.6096	0.2600	25.74	7.1415
WC	0.6124	0.2600	25.98	7.1852
WPC	0.6107	0.2587	25.79	7.1595
Cw	0.5749	0.2384	22.73	6.6149
Cwl	0.5756	0.2385	22.73	6.6204
Cwp	0.5771	0.2395	23.06	6.6403
Cwc	0.5759	0.2390	22.86	6.6207
Cwpc	0.5744	0.2379	22.77	6.5949

Table 3.3: MT Results for the 10 elementary data views on the development set. GTM-1 and GTM-2 show the GTM F_1 -measure for different values of e ($e=1.0$, $e=2.0$, respectively). BLEU shows the accumulated BLEU score for 4-grams. Finally, NIST shows the accumulated NIST score for 5-grams.

In Table 3.4 MT results for different data view combinations are showed. Merged model weights are set equiprobable, and no phrase-pair score filtering is performed. We refer to the W model as our baseline. In this view, only words are used. The 5W-MRG and 5W-GPHEX models use a combination of the 5 word-based data views, as in MRG and GPHEX, respectively. The 5C-MRG and 5C-GPHEX system use a combination of the 5 chunk based data views, as in MRG and GPHEX, respectively. The 10-MRG system uses all 10 data views combined as in MRG. The 10-GPHEX/MRG system uses the 5 word based views combined as in GPHEX, the 5 chunk based views combined as in GPHEX, and then a combination of these two combo-models as in MRG.

It can be seen that results improve by combining several data views. Furthermore, global phrase extraction (GPHEX) seems to work much finer than local phrase extraction (LPHEX).

Table 3.5 shows MT results after optimizing λ_{tm} , λ_{lm} , λ_w , and the weights for the MRG operation, by means of the *Downhill Simplex Method in Multidimensions* (William H. Press & Flannery, 2002). Observe that tuning the system improves the performance considerably. The λ_w parameter is particularly sensitive to tuning.

⁹These data sets are based on the Europarl corpus (see Subsection 4.1.1).

data view	GTM-1	GTM-2	BLEU	NIST
W	0.6108	0.2609	25.92	7.1576
5W-MRG	0.6134	0.2631	26.25	7.2122
5W-GPHEX	0.6172	0.2615	26.95	7.2823
5C-MRG	0.5786	0.2407	23.18	6.6754
5C-GPHEX	0.5739	0.2368	22.80	6.5714
10-MRG	0.6130	0.2624	26.24	7.2196
10-GPHEX/MRG	0.6142	0.2600	26.58	7.2542

Table 3.4: MT Results without tuning, for some data view combinations on the development set. GTM-1 and GTM-2 show the GTM F_1 -measure for different values of e ($e=1.0$, $e=2.0$, respectively). BLEU shows the accumulated BLEU score for 4-grams. Finally, NIST shows the accumulated NIST score for 5-grams.

Even though the performance of chunk-based models is poor, the best results are obtained by combining the two levels of abstraction, thus proving that syntactically motivated phrases may help. 10-MRG and 10-GPHEX models achieve a similar performance. The *10-MRG-best_{WN}* system corresponds to the 10-MRG model using WordNet. The *10-MRG-sub_{WN}* system is this same system at the time of submission. Results using WordNet, taking into account that the number of unknown¹⁰ words in the development set was very small, are very promising.

data view	GTM-1	GTM-2	BLEU	NIST
W	0.6174	0.2583	28.13	7.1540
5W-MRG	0.6206	0.2605	28.50	7.2076
5W-GPHEX	0.6207	0.2603	28.38	7.1992
5C-MRG	0.5882	0.2426	25.06	6.6773
5C-GPHEX	0.5816	0.2387	24.40	6.5595
10-MRG	0.6218	0.2623	28.88	7.2491
10-GPHEX/MRG	0.6229	0.2622	28.82	7.2414
<i>10-MRG_{WN}</i>	0.6228	0.2625	28.90	7.2583
<i>10-MRG-sub_{WN}</i>	0.6228	0.2622	28.79	7.2528

Table 3.5: MT Results for some data view combinations after tuning on the development set. GTM-1 and GTM-2 show the GTM F_1 -measure for different values of e ($e=1.0$, $e=2.0$, respectively). BLEU shows the accumulated BLEU score for 4-grams. Finally, NIST shows the accumulated NIST score for 5-grams.

At competition time, our *LDV-COMBO* system obtained a BLEU of 28.13 on the test set, thus scoring sixth on nine systems which presented to the Spanish-English shared task.

3.2.5 Conclusions

We have showed that it is possible to obtain better phrase-based translation models by utilizing alignments built on top of different linguistic data views. These models can be robustly combined, significantly outperforming all of their components in isolation. Further steps to be taken are outlined in Section 4.2.

¹⁰Translation for 349 unknown words was found in the MCR.

3.3 Automatic Translation of WordNet Glosses

We study the possibility of translating the glosses in the English WordNet (Fellbaum, 1998). WordNet glosses are a very useful resource. Mihalcea and Moldovan (1999) suggested an automatic method for generating sense tagged corpora which uses WordNet glosses. Hovy et al. (2001) used WordNet glosses as external knowledge to improve their Webclopedia Question Answering (QA) system.

However most of the wordnets in the *Multilingual Central Repository* (MCR) (Atserias et al., 2004) contain very few glosses. For instance, in the current version of the Spanish WordNet fewer than 10% of the synsets have a gloss. Conversely, since version 1.6 every synset in the English WordNet has a gloss. We believe that a method to rapidly obtain glosses for all wordNets in the MCR may be helpful. These glosses could serve as a starting point for a further step of revision and post-editing. Furthermore, from a conceptual point of view, the idea of enriching the MCR using the MCR itself results very attractive.

In the absence of a parallel corpus of definitions, we decided to build phrase-based translation models on the Europarl corpus (see Subsection 4.1.1). However, the language of definitions is very specific and different to that of parliament proceedings. This is particularly harmful to the system recall, because many unknown words will be processed.

In order to adapt the system to the new domain we study two separate lines. First, we use electronic dictionaries in order to build more adequate language models. Second, we work with domain independent word-based translation models extracted from the MCR. Other authors have previously applied information extracted from aligned wordnets to other NLP problems like Word Sense Disambiguation (WSD) (Tufis et al., 2004). We suggest to use these models as a complement to phrase-based models. These two proposals together with a good tuning of the system parameters lead to a notable improvement of results. In our experiments, we focus on translation from English into Spanish. A relative increase of 64% in BLEU measure is achieved by limiting the use of the MCR-based model to the case of unknown words .

3.3.1 Experimental Setting

In the following we deploy experimental work. We tokenized and case lowered the Europarl corpus. A set of 327,368 parallel segments of length between five and twenty was selected for training. The Spanish side consisted of 4,243,610 tokens, whereas the English side consisted of 4,197,836 tokens.

We built a trigram language model from the Spanish side of the Europarl corpus selection. Linear interpolation was applied for smoothing.

We used GIZA++¹¹ to estimate the word alignment probabilities. In the phrase extraction we worked with the union of source-to-target and target-to-source alignments, with no heuristic refinement. Only phrases up to length five were considered. Also, phrase pairs in which the source/target phrase was more than three times longer than the target/source phrase were ignored. Finally, phrase pairs appearing only once were discarded, too.

¹¹We use the GIZA++ default configuration (5 iterations for model 1, 4 iterations for model 3, 3 iterations for model 4, and 5 iterations for HMM model).

By means of the MCR we obtained a set of 6503 parallel glosses. These definitions correspond to 5684 nouns, 87 verbs, and 732 adjectives. Examples and parenthesized texts were removed. Gloss average length was 8,03 words for English and 7,83 for Spanish. Parallel glosses were tokenized and case lowered, and randomly split into development (3295 gloss pairs) and test (3208 gloss pairs) sets.

3.3.2 Results for the Baseline System

MT results for the baseline system on the new domain are very low in comparison to the performance on a set of 8490 unseen sentences from the European Parliament Proceedings. See Table 3.6.

system	GTM-1	GTM-2	BLEU	NIST
EU-baseline-dev	0.3091	0.2196	0.0730	3.0953
EU-baseline-test	0.3028	0.2155	0.0657	3.0274
EU-europarl	0.5885	0.3567	0.2725	7.2477

Table 3.6: Preliminary MT Results on development (dev) and test (test) sets, and on a Europarl test set. GTM-1 and GTM-2 show the GTM F_1 -measure for different values of e ($e = 1$, $e = 2$, respectively). BLEU shows the accumulated BLEU score for 4-grams. Finally, NIST shows the accumulated NIST score for 5-grams.

Some cases of good and bad translations are shown in Table 3.7 and Table 3.8, respectively. Only 28 glosses obtain an F_1 over 0.9. Most of them are too short, less than 5 words (e.g. 1807 and 2917). 10% of the glosses (320) obtain an F_1 over 0.5. Interestingly, many of them are somehow related to the domain of politics (e.g. 193, 293, 1414, 1674 and 1721) or economy (e.g. 362). On the other hand, 18% of the glosses obtain an F_1 below 0.1. In many cases this is due to unknown vocabulary (e.g. 34, 508, 2263 and 2612). However, we found many translations unfairly scoring too low due to strong divergences between source and reference. We call this phenomenon ‘*quasi-parallelism*’ (e.g. 7, 1606, and 2985).

3.3.3 Improved Language Modeling

We report some improvements by working on specialized language models. In order to build these additional language models, we introduced two large monolingual Spanish electronic dictionaries, consisting of 142,892 definitions (2,112,592 tokens) and 168,779 definitons (1,553,674 tokens), respectively.

We tried different language model configurations. See Table 3.9. We refer to the baseline system, which uses the Europarl language model only, as ‘*EU*’. In ‘*D1*’ and ‘*D2*’ we replaced the language model with those obtained from dictionaries D1 and D2, respectively. ‘*D1-D2*’ combines the two dictionaries with equal probability. ‘*D1-D2-EU*’ combines all three language models with equal probability.

As expected, the language models built out from dictionaries work much better than the one built from the Europarl corpus. Results improve still slightly further by combining the two dictionaries.

case	synset-ili	Source	Target	Reference
<i>'good' translations</i>				
193	00392749#n	the office and function of president	el cargo y función de presidente	cargo y función de presidente
293	00630513#n	the action of attacking the enemy	acción de atacar al enemigo	acción y efecto de atacar al enemigo
345	00785108#n	the act of giving hope or support to someone	la acción de dar esperanza o apoyo a alguien	acción de dar esperanza o apoyo a alguien
362	00804210#n	the combination of two or more commercial companies	la combinación de dos o más comerciales compañías	combinación de dos o más empresas
1414	05359169#n	the act of presenting a proposal	el acto de presentar una propuesta	acto de presentar una propuesta
1674	06089036#n	a military unit that is part of an army	unidad militar que forma parte de un ejército	unidad militar que forma parte de un ejército
1721	06213619#n	a group of representatives or delegates	grupo de representantes o delegados	grupo de representantes o delegados
1807	06365607#n	a safe place	lugar seguro	lugar seguro
2917	01612822#v	perform an action	realizar una acción	realizar una acción

Table 3.7: MT examples of good translations output by the baseline system. 'Source' and 'Target' refer to the input and output of the system, respectively. 'Reference' corresponds to the expected output.

case	synset-ili	Source	Target	Reference
<i>'bad' translations</i>				
7	00012865#n	a feature of the mental life of a living organism	una característica de la vida mental de un organismo vivo	rasgo psicológico
34	00029442#n	the act of departing politely	el acto <i>departing politely</i>	acción de marcharse de forma educada
508	02581431#n	a kitchen appliance for disposing of garbage	<i>kitchen appliance</i> para <i>disposing</i> de <i>garbage</i>	cubo donde se depositan los residuos
1606	05961082#n	people in general	gente en general	grupo de gente que constituye la mayoría de la población y que define y mantiene la cultura popular y las tradiciones
2263	07548871#n	a painter of theatrical scenery	una <i>painter</i> de <i>theatrical scenery</i>	persona especializada en escenografía
2612	10069279#n	rowdy behavior	<i>rowdy behavior</i>	comportamiento escandaloso
2985	00490201#a	without reservation	sin reservas	movido por una devoción o un compromiso entusiasta y decidido

Table 3.8: MT examples of bad translations output by the baseline system. ‘Source’ and ‘Target’ refer to the input and output of the system, respectively. ‘Reference’ corresponds to the expected output.

language model	GTM-1	GTM-2	BLEU	NIST
EU	0.3091	0.2196	0.0730	3.0953
D1	0.3361	0.2409	0.0905	3.4881
D2	0.3374	0.2419	0.0890	3.4719
D1-D2	0.3422	0.2457	0.0940	3.5515
D1-D2-EU	0.3428	0.2456	0.0949	3.5655

Table 3.9: MT Results on the development set for different language model configurations. GTM-1 and GTM-2 show the GTM F_1 -measure for different values of e ($e = 1$, $e = 2$, respectively). BLEU shows the accumulated BLEU score for 4-grams. Finally, NIST shows the accumulated NIST score for 5-grams.

A relative increase of 30% in BLEU score is reported. Adding the EU language model does not report any significant improvement.

3.3.4 Using the MCR

We also exploit the MCR to build domain independent translation models. Outer knowledge may be supplied to the *Pharaoh* decoder by annotating the input with alternative translation options via XML-markup. In the default setting we enrich all content words (nouns, verbs, adjectives and adverbs) by looking up possible translations for all their meanings in the MCR.

Translation pairs are heuristically scored by relative frequency according to the number of senses which may lexicalize in the same manner (Equation 3.1). Let w_f, p_f be the source word and PoS, and w_e be the target word, we define a function $Scount(w_f, p_f)$ which returns the number of senses for (w_f, p_f) . We define also a function $Scount(w_f, p_f, w_e)$ which counts the number of senses for (w_f, p_f) which may lexicalize as w_e .

$$score(w_f, p_f | w_e) = \frac{Scount(w_f, p_f, w_e)}{Scount(w_f, p_f)} \quad (3.1)$$

In our experiments we work on a normalized version of this heuristic.

$$score(w_f, p_f | w_e) = \frac{Scount(w_f, p_f, w_e)}{\sum_{(w_f, p_f)} Scount(w_f, p_f, w_e)} \quad (3.2)$$

In WordNet all word forms related to the same concept are grouped and represented by their lemma and part-of-speech (PoS). Therefore, input word forms must be lemmatized and PoS-tagged. WordNet takes care of the lemmatization step. For PoS-tagging we utilize the *svmTool*. Similarly, at the output, the MCR provides us with lemmas instead of word forms as translation candidates. A lemma extension must be performed. We utilize components from the *Freeling* package for this step. Details of NLP tools utilized may be found in Subsection 4.1.2. See an example of enriched input in Table 3.10.

Then, we proceeded applying the MCR-based model. Several strategies were tried. See results in Table 3.11. In any case, we allowed the decoder to bypass the MCR-based model when a better solution was found using the phrase-based model alone.

```

<NN english="consecuciones|consecución|logro|logros|
realizaciones|realización" prob="0.1666|0.1666|0.1666|
0.1666|0.1666|0.1666">accomplishment</NN> of an objective

an organism such as an <NN english="insecto|insectos"
prob="0.5|0.5">insect</NN>that habitually shares the
<NN english="madriguera|madrigueras|nido|nidos"
prob="0.25|0.25|0.25|0.25"> nest</NN> of a species of
<NN english="hormiga|hormigas" prob="0.5|0.5">
ant</NN>

the part of the human <NN english="pierna|piernas"
prob="0.5|0.5">leg</NN> between the
<NN english="rodilla|rodillas" prob="0.5|0.5">knee</NN>
and the <NN english="tobillo|tobillos" prob="0.5|0.5"> ankle</NN>

a <JJ english="casada|casadas|casado|casados"
prob="0.25|0.25|0.25|0.25">married</JJ>man

a football game in which two teams of 11 players try to
<VB english="chuta|chutaba|chutabais|chutaban|chutabas|
chutad|chutada|chutadas|chutado|chutados|chutamos|chutan|
chutando|chutar|chutara|chutarais|chutaran|chutaras|
chutare|chutareis|chutaremos|chutaren|chutares|chutaron|
chutará|chutarán|chutarás|chutaré|chutaréis|chutaría|
chutaríais|chutaríamos|chutarían|chutarías|chutas|chutase|
chutaseis|chutasen|chutases|chutaste|chutasteis|chute|
chutemos|chuten|chutes|chuto|chutábamos|chutáis|chutáramos|
chutáremos|chutásemos|chuté|chutéis|chutó"
prob="0.0185|0.0185|0.0185|0.0185|0.0185|0.0185|0.0185|
0.0185|0.0185|0.0185|0.0185|0.0185|0.0185|0.0185|0.0185|
0.0185|0.0185|0.0185|0.0185|0.0185|0.0185|0.0185|0.0185|
0.0185|0.0185|0.0185|0.0185|0.0185|0.0185|0.0185|0.0185|
0.0185|0.0185|0.0185|0.0185|0.0185|0.0185|0.0185|0.0185|
0.0185|0.0185|0.0185|0.0185|0.0185|0.0185|0.0185|0.0185">
kick</VB> or head a ball into the opponents' goal

strengthening the concentration by removing <JJ english="irrelevante|
irrelevantes" prob="0.5|0.5">extraneous</JJ> material

```

Table 3.10: A sample of enriched input, scored as detailed in Equation 3.2.

We defined as new baseline the system which combines the three language models (no-MCR). In a first experiment, we enriched all content words in the validation set with all possible translation candidates (ALL). No improvement is achieved. By inspecting input data, apart from some PoS-tagging errors, we found that the number of translation options generated via MCR grows too fast for words with too many senses, particularly for verbs. In order to reduce the degree of polysemy in the enriched input we tried limiting to words with 1, 2, 3, 4 and 5 different senses at most (S1, S2, S3, S4 and S5). Results improve slightly.

Ideally, one would wish to work with accurately word sense disambiguated input. We tried restricting translation candidates to those generated by the most frequent sense only (ALL-mfs). There is no significant variation in results.

We also studied the behavior of the model applied separately to nouns (N-mfs), verbs (V-mfs), adjectives (A-mfs) and adverbs (R-mfs). The system works worst for nouns, and seems to work a little better for adjectives than for verbs and adverbs.

All in all, we did not find an adequate manner to have the two translation models, to cooperate properly. Therefore we decided to use the MCR-based model only for those words unknown¹² to the phrase-based model (UNK-mfs). A significant relative improvement of 9% in BLEU score is achieved.

This result seems to evidence that recall is more important than precision for SMT. The ‘UNK-mfs’ model is intended to improve recall whereas the ‘ALL-mfs’ model is intended to improve precision.

Finally, we tried translating only those words that were both unknown and monosemous (UNK-and-S1), and those that were either unknown or monosemous (UNK-or-S1). Results did not improve.

3.3.5 Tuning the System

Another path we explored is the tuning of the *Pharaoh* parameters that control the importance of the different probabilities that govern the search.

In general, there are 4 important parameters to adjust: the language model probability (λ_{lm}), the translation model probability (λ_{ϕ}), the distortion probability (λ_d) and the word penalty factor (λ_w).

We utilize a software based on the *Downhill Simplex Method in Multidimensions* (William H. Press & Flannery, 2002). Parameters were tuned for the ‘no-MCR’ and ‘UNK-mfs’ strategies on the development set. A further relative gain of 9% in BLEU score is reported. See Table 3.12.

Recall, for instance, the difference in length between source and target references. Tuning the λ_w parameter leads to better results. Also, a proper tuning of the probabilities of the three language models yields a significant improvement.

We analyzed results by the ‘UNK-mfs’ and ‘ALL-mfs’ strategies based on the GTM F-measure ($e = 2$). Table 3.13 shows some cases where MCR-based models prove their usefulness (e.g. 29, 35, 194, 268, 351, 377 and 965) and some cases where they cause the system to make a mistake (e.g. 1001, 1125 and 2570).

¹²7.87% of the words in the development set are unknown.

strategy	GTM-1	GTM-2	BLEU	NIST
no-MCR	0.3428	0.2456	0.0949	3.5655
ALL	0.3382	0.2439	0.0949	3.4980
ALL-mfs	0.3367	0.2434	0.0951	3.4720
S1	0.3432	0.2469	0.0961	3.5774
S2	0.3424	0.2464	0.0963	3.5686
S3	0.3414	0.2459	0.0963	3.5512
S4	0.3412	0.2458	0.0966	3.5441
S5	0.3403	0.2451	0.0962	3.5286
N-mfs	0.3361	0.2428	0.0944	3.4588
V-mfs	0.3428	0.2456	0.0945	3.5649
A-mfs	0.3433	0.2462	0.0959	3.5776
R-mfs	0.3428	0.2456	0.0949	3.5655
UNK-mfs	0.3538	0.2535	0.1035	3.7580
UNK-and-S1	0.3463	0.2484	0.0977	3.6313
UNK-or-S1	0.3507	0.2523	0.1026	3.7104

Table 3.11: MT Results on the development set, using the MCR-based model. GTM-1 and GTM-2 show the GTM F_1 -measure for different values of e ($e = 1$, $e = 2$, respectively). BLEU shows the accumulated BLEU score for 4-grams. Finally, NIST shows the accumulated NIST score for 5-grams.

3.3.6 Conclusions

By working with specialized language models and MCR-based translation models we achieved a relative gain of 63.62% in BLEU score (0.0657 vs 0.1075) when porting the system to a new domain. Further work is outlined in Section 4.2.

strategy	GTM-1	GTM-2	BLEU	NIST
baseline-dev	0.3091	0.2196	0.0730	3.0953
baseline-test	0.3028	0.2155	0.0657	3.0274
no-MCR-dev	0.3428	0.2456	0.0949	3.5655
no-MCR-test	0.3352	0.2420	0.0915	3.4802
ALL-mfs-dev	0.3367	0.2434	0.0951	3.4720
ALL-mfs-test	0.3323	0.2422	0.0937	3.4425
UNK-mfs-dev	0.3538	0.2535	0.1035	3.7580
UNK-mfs-test	0.3478	0.2500	0.0991	3.6946
no-MCR-dev-T	0.3492	0.2496	0.1026	3.5352
noMCR-test-T	0.3431	0.2450	0.0965	3.4628
UNK-mfs-dev-T	0.3599	0.2582	0.1124	3.7609
UNK-mfs-test-T	0.3554	0.2546	0.1075	3.7079
ALL-mfs-dev-T	0.3395	0.2462	0.0983	3.4891
ALL-mfs-test-T	0.3340	0.2428	0.0939	3.4282

Table 3.12: MT Results for the baseline, ‘no-MCR’, ‘ALL-mfs’, and ‘UNK-mfs’ strategies, before and after tuning (T) on development (dev) and test (test) sets. GTM-1 and GTM-2 show the GTM F_1 -measure for different values of e ($e = 1$, $e = 2$, respectively). BLEU shows the accumulated BLEU score for 4-grams. Finally, NIST shows the accumulated NIST score for 5-grams.

case	synset-ili	Source	Target-base	Target-MCR	Reference
UNK-mfs					
29	00025788#n	accomplishment of an objective	accomplishment de un objetivo	consecución de un objetivo	consecución de un objetivo
194	00393890#n	the position of secretary	situación de secretary	el cargo de secretario	posición de secretario
268	00579072#n	the activity of making portraits	actividad de hacer portraits	actividad de hacer retratos	actividad de hacer retratos
377	00913742#n	an organism such as an insect that habitually shares the nest of a species of ant	un organismo como un insect que habitually comparte el nest de una especie de ant	un organismo como un insecto que habitually comparte el nido de una especie de hormiga	organismo que comparte el nido de una especie de hormigas
965	04309478#n	the part of the human leg between the knee and the ankle	parte de la persona leg entre los knee y el ankle	parte de la persona pierna entre la rodilla y el tobillo	parte de la pierna humana comprendida entre la rodilla y el tobillo
ALL-mfs					
35	00029961#n	the act of withdrawing	el acto de retirar	el acto de retirarse	acción de retirarse
351	00790504#n	a favorable judgment	una sentencia favorable	una opinión favorable	opinión favorable
1001	04395081#n	source of difficulty	fuelle de dificultad	fuelle de problemas	fuelle de dificultad
1125	04634158#n	the branch of biology that studies plants	rama de la biología que estudios plantas	rama de la biología que estudia factoría	rama de la biología que estudia las plantas
2570	10015334#n	balance among the parts of something	equilibrio entre las partes de algo	equilibrio entre las partes de entidades	equilibrio entre las partes de algo

Table 3.13: MT examples of the ‘ALL-mfs’ and ‘UNK-mfs’ strategies. ‘Source’ refers to the raw input. ‘Target-base’ and ‘Target-MCR’ correspond to the output of the baseline and MCR helped systems, respectively. ‘Reference’ corresponds to the expected output.

Chapter 4

Work Plan

This research project, which began two years ago, is planned in three stages which can be clearly separated in time:

1. Development of Resources
2. Construction and Improvement of a Phrase-based SMT System
3. Expecting a Breakthrough

The first year was devoted to developing linguistic resources, basically corpora and NLP tools. Details are reported in Section 4.1. The second year has been devoted to the construction of a state-of-the-art phrase-based SMT system, as described in Chapter 3. We introduced shallow syntactic information and used external knowledge sources, achieving significant improvements.

We are currently involved in the final phase, in which we maintain a crusade in favor of linguistically ruled MT. In Section 4.2 we sketch some further steps.

4.1 Development of Resources

4.1.1 Corpora Collection

As detailed in Chapter 2, SMT relies on the availability of large parallel corpora. We explored existing corpora and developed new ones.

New Corpora

In the framework of the Lc-star project¹ an oral database of almost 60 hours of conversations in the tourist domain was developed.

These conversations were transcribed and translated so as to form the TALP-tourist corpus (Aranz et al., 2003), a middle size parallel corpus for Catalan, Spanish and US-English. We selected

¹<http://www.lc-star.com>.

words from this corpus to build lexica. We also mined the web for touristic sites, from which we built new lexica, too. Finally, an XML representation was designed to encode linguistic information.

Speech corpora present many grammatical disfluences such as repetitions, corrections or false starts. Because our research project is focused on text translation we later abandoned this resource.

Existing Corpora

In our search for parallel texts, we sought specially for widely used and well-known corpora. That would allow us to contrast our results with results by many other groups working in SMT. We also preferred large corpora over small ones:

Hansards Corpus Official records (Hansards) of the 36th Canadian Parliament. It is available² for French and English.

United Nations Corpus The documents come from the Office of Conference Services at the UN in New York and are drawn from archives that span the period between 1988 and 1993. It is available³ for English, French and Spanish.

Europarl Corpus European Parliament Proceedings Parallel Corpus 1996-2003 (Koehn, 2003a). It is available⁴ for 11 European languages.

In the framework of the Shared Task 2: “Exploiting Parallel Texts for Statistical Machine Translation” of the ACL-2005 Workshop on “Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond”, we utilized the Europarl Corpus, as provided by the organizers.

4.1.2 Development of NLP Tools

As explained along this report, the main challenge of our project is to integrate linguistic information in the construction of a MT system. In order to achieve robustness we aim to use same tools for all the languages.

Tools

PoS Tagging We developed the SVMTool⁵ which is a simple, flexible, effective and efficient part-of-speech tagger based on Support Vector Machines. The SVMTool offers a fairly good balance among these properties which make it really practical for current NLP applications. This research has allowed for two publications in international conferences (Giménez & Màrquez, 2003; Giménez & Màrquez, 2004).

Lemmatization We use the *Freeling*⁶ package (Carreras et al., 2004).

²The Hansards corpus may be freely downloaded at <http://www.isi.edu/natural-language/download/hansard/>.

³The UN corpus is sold by LDC (LDC94T4A).

⁴The Europarl corpus may be freely downloaded at <http://people.csail.mit.edu/people/koehn/publications/europarl/>.

⁵The SVMTool may be freely downloaded at <http://www.lsi.upc.es/~nlp/SVMTool>.

⁶Freeling Suite of Language Analyzers may be downloaded at <http://www.lsi.upc.es/~nlp/freeling/>.

Shallow Parsing We use the *Phreco* software by (Carreras et al., 2005). In order to build chunkers for Spanish and Catalan we performed a deep study of the syntactic annotation of the 3LB corpus.

Clause Splitting Currently under development, (Carreras et al., 2005).

Semantic Role Labeling Currently under development. We presented at the Ninth Conference on Computational Natural Language Learning (CoNLL 2005) shared task on Semantic Role Labeling for English. Our system resulted in third position (Márquez et al., 2005).

Word Sense Disambiguation currently under development by group fellows, (Villarejo et al., 2004).

Some of these linguistic processors are already available (V), some are under development (~), and some are not yet being constructed (X). See Table 4.1.

NLP Tool	Catalan	English	Spanish
PoS tagging	V	V	V
Lemmatization	V	V	V
Shallow Parsing	V	V	V
Clause Splitting	~	V	~
Semantic Role Labeling	X	V	X
Word Sense Disambiguation	X	V	X

Table 4.1: Availability of Linguistic Processors for Catalan, English and Spanish.

Training Data

The development of these NLP tools is based on annotated corpora:

Penn Treebank We use it to build Pos taggers, chunkers and clause splitters for English (Marcus et al., 1994).

3LB We use it to build Pos taggers, chunkers and clause splitters for Spanish (Navarro et al., 2003).

Propbank We use it to build Semantic Role Labelers for English (Kingsbury et al., 2002).

SemCor We use it to build WSD systems for English (Fellbaum, 1997)⁷.

4.2 Further Work

There are several open lines we plan to investigate. Some refer to minor improvements and some devise important research prospectives.

⁷Different versions of the SemCor corpus may be downloaded from <http://www.cs.unt.edu/rada/downloads.html#semcor>.

4.2.1 Minor Improvements

Pre-processing

Special treatment of named entities, dates, numbers, currency, etcetera, should be considered so as to further enhance the system.

Phrase extraction

We believe that linguistic annotation may be used to select better phrase-pairs during phrase-extraction. We will consider PoS-tagging, shallow parsing, clause splitting, and semantic role labeling. Because words may lexicalize differently according to their PoS, meaning, syntactic role, semantic role, etcetera.

Phrase scoring

We have scored phrase pairs based on their relative frequency. There are additional scoring techniques which we have not tried yet.

Moreover, linguistic information may be taken into account when scoring phrase pairs in the translation model. We could define a series of linguistic constraints and promote those phrase pairs that comply with constraints and demote those which do not.

Errors in PoS-tagging and Lemmatization

WordNet glosses are not typically sentences. In most cases, as seen in Section 3.3, glosses are arguments instead. That is surely causing many tagging errors because the Pos-tagger has been trained on a corpus of well-formed sentences. For instance, in the gloss “*charge anew*” for the verb synset “00361278#v (*recharge*)”, the word *charge* may be either a noun or a verb. Without further context it is very hard to tag it correctly. We believe these errors could be minimized by fitting glosses in sentence templates, e.g. “**to recharge is to charge anew**”

We also noticed that results for adverbs are equal to the baseline because in WordNet the lemma for adverbs is, in contrast to our lemmatizer, an adjective, thus causing our system not to find any match. Words unknown to the lemmatizer are also systematically missed where a couple of rules would suffice in a number of cases.

Integration of MCR-based models

There is a strong limitation in the way we integrate MCR-based models into our SMT system. When we markup the input to Pharaoh we are somehow forcing the decoder to choose between a word-to-word translation and a phrase-to-phrase translation. It has been demonstrated that in SMT phrase-based models outperform word-based ones. A better way to integrate MCR-based models with phrase-based models should be investigated.

Candidate selection heuristics for MCR-based models

We tried selecting *all* possible candidates, or candidates for the most frequent sense. Better results should be obtained by working with word sense disambiguated text. We believe that coarse-grained WSD is sufficient for the purpose of MT. We could utilize disambiguated glosses, either from eXtended WordNet (Mihalcea & Moldovan, 2001) or from the system winner in the Senseval-3 workshop (Castillo et al., 2004). We could also use information regarding domain, semantic file, top ontology, sumo, etc. to select lexicalizations belonging to a specific domain.

Scoring heuristics for MCR-based models

We have used fairly simple heuristics. Our assumption is that all lexicalizations for a given sense are equally probable. We only favor those target words that are valid lexicalizations for several senses of the source word. More sophisticated heuristics should be considered for scoring MCR-based translation candidates. For instance, we could favor some lexicalizations according to the domain, semantic file, top ontology, sumo, etc. to favour lexicalizations belonging to a given domain. We could also work on a coarse granularity setting.

4.2.2 Research Prospectives

Further Linguistic Analysis

We believe that by providing further detailed linguistic information a gain in MT quality may be achieved. So far we have performed a shallow syntactic analysis, providing information regarding part-of-speech tagging, lemmatization and chunking.

Ideally, one would wish to have still deeper information, moving through syntax onto semantics, such as *word senses*. We could use this information to build richer linguistic data views as explained in Section 3.2. Therefore, it would be possible to distinguish for instance between two realizations of ‘plays’ with different meanings: ‘*he_{PRP} plays_{VBG} guitar_{NN}*’ and ‘*he_{PRP} plays_{VBG} basketball_{NN}*’.

Also, we could elaborate a semantic analysis based on *semantic roles*. We would know *who* the player is, *what* is being played, *when*, *where*, etc., and thus gain insight of the meaning of the source sentence. We could impose constraints on the target sentence so as to guarantee that it conveys the same meaning as the source.

We consider clause splitting as well. Due to the recursive nature of Natural Language, we may find very long and complicated sentences which are hard and costly to translate. Splitting these sentences into a set of smaller clauses could report many benefits.

Semantic models

With the release of the MultiSemCor⁸ Italian-English parallel corpus (Bentivogli et al., 2005) a new set of possibilities is devised. MultiSemCor is semantically annotated with WordNet senses for both Italian and English. Moreover, it is word-aligned. Most interestingly, the domain of the corpus is closer to the domain of parliament proceedings than dictionary definitions are.

⁸The MultiSemCor Corpus is available at <http://multisemcor.itc.it/index.php>

We consider the possibility of introducing a semantic model that would account for the probability of each of the different senses of a given word in its context.

A first possibility is to replace the default translation model with a semantic model. We must estimate $P(e|f)$, the probability that a translator produces f as a translation of e taking senses into account. Therefore we define $P(f|e, s_e)$ as the probability that a translator produces the word f as a translation of the word e having the meaning s_e :

$$P(e|f) = \frac{P(f|e, s_e) * P(e, s_e)}{P(f)} \quad (4.1)$$

where s_e is the sense of the word e according to a given sense repository.

Analogously, the search:

$$e = \operatorname{argmax}_e P(f|e, s_e) * P(e, s_e) \quad (4.2)$$

The advantage of this alternative is that the same decoder will work. However, word sense disambiguation is still at a premature stage. Therefore experimental work will have to deal with the noise cause by erroneous sense tagging.

A second alternative is to use the semantic model as a separate component:

$$P(e|f) = \frac{P(f|e) * P(e) * P(s_f|s_e) * P(s_e)}{P(f)} \quad (4.3)$$

where s_e is the sense of the word e according to a given sense repository.

Analogously, the search:

$$e = \operatorname{argmax}_e P(f|e) * P(e) * P(s_f|s_e) * P(s_e) \quad (4.4)$$

This alternative may imply working on a new decoder because the search space is slightly different. We are bound to do so if necessary, but first we must study the possibility of adapting the existing decoder to the new problem.

Discriminative Learning

SMT is based on Generative Models. We consider the possibility of applying Discriminative Models borrowed from the Machine Learning community, such as Support Vector Machines, Perceptrons or Decision Trees.

A first option is to try the reranking approach as suggested by Shen et al. (2004) (Section 2.3). The reranking approach is based on the idea that during the search good translations tend to get separated from bad translations. Therefore it is possible to build classifiers that learn to distinguish good and bad translations (Och et al., 2003).

Hybrid Solutions

Although high-quality MT is still a utopia, we believe that there is room for improving the fluency. Linguistic information may be used to automatically post-edit SMT results with the intent to improve the level of grammaticality. A set of language dependent rules could be applied. Habash

and Dorr (2002) tried the opposite, using statistics to improve a system based on interlingua. Our proposal moves toward a hybrid solution between empirical and rule-based systems.

MT Evaluation

Because current automatic MT evaluation metrics are not yet well suited to capture linguistic improvements, we do not discard to work on this issue. Otherwise, at some point it may be probably necessary to perform some kind of human evaluation.

New Language Pairs

In the future we will move to other language pairs. With that intent we have already developed a number of NLP tools and resources for Catalan. Particularly, we will consider Catalan-Spanish and Catalan-English.

4.2.3 Project Scheduling

This thesis work is expected to conclude in two years from now. The first six months will be mainly devoted to the minor improvements suggested in Subsection 4.2.1. At the same time, we will also study and deploy some of the research prospectives sketched in Subsection 4.2.2. The time dedicated to the different alternatives will depend on results attained in the following six months. We are particularly optimistic about *semantic models* and *hybrid solutions*. Simultaneously, and during approximately 18 months, because we are continuously developing new linguistic processors, experimental work based on *further linguistic analysis* will be carried out as new tools are becoming available. We will continue to participate in MT international events, such as shared tasks, competitions and conferences. The last six months will be spent in the writing of the doctoral thesis.

Chapter 5

Publications

In the following, related publications¹ authored or coauthored by us are cited and briefly commented.

5.1 Our approach to MT

Related work is detailed in Chapter 3.

- Jesús Giménez and Lluís Márquez.
Combining Linguistic Data Views for Phrase-based SMT.
In proceedings of the ACL Workshop on “Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond”. Ann Arbor, Michigan, US, 2005.
Abstract: We describe the Spanish-to-English *LDV-COMBO* system for the Shared Task 2: “Exploiting Parallel Texts for Statistical Machine Translation” of the ACL-2005 Workshop on “Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond”. See Section 3.2.
- Jesús Giménez, Lluís Márquez and German Rigau.
Automatic Translation of WordNet Glosses.
Eurolan Cross-Language Knowledge Induction Workshop. Cluj-Napoca, Romania, 2005.
Abstract: We approach the task of automatically translating the glosses in the English WordNet. See Section 3.3.

5.2 Developing NLP Tools

Related work is detailed in Subsection 4.1.2.

- Lluís Márquez, Pere Comas, Jesús Giménez and Neus Català.
Semantic Role Labeling as Sequential Tagging.

¹Listed publications are available at <http://www.lsi.upc.edu/~jgimenez/pubs.html>

Ninth Conference on Computational Natural Language Learning (CoNLL), 2005.

Abstract: We describe the Semantic Role Labeling system presented to the CoNLL 2005 shared task.

- Jesús Giménez and Lluís Márquez.
SVMTool: A general POS tagger generator based on Support Vector Machines.
 In Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04), vol. I, pages 43 - 46. Lisbon, Portugal. 2004 . (ISBN 2-9517408-1-6)
Abstract: This paper presents the SVMTool, a simple, flexible, effective and efficient part-of-speech tagger based on Support Vector Machines. The SVMTool offers a fairly good balance among these properties which make it really practical for current NLP applications. The SVMTool may be freely downloaded at <http://www.lsi.upc.es/~nlp/SVMTool>. A number of NLP researchers have already tried it reporting satisfactory feedback.
- Jesús Giménez and Lluís Márquez.
SVMTool: A general POS tagger generator based on Support Vector Machines (Technical Manual).
 Departament Research Report (LSI-04-34-R), Technical University of Catalonia, 2004.
Abstract: This report is a detailed technical manual for the SVMTool.
- Jesús Giménez and Lluís Márquez.
Fast and Accurate Part-of-Speech Tagging: The SVM Approach Revisited.
 In Proceedings of the International Conference RANLP - 2003 (Recent Advances in Natural Language Processing), pages 158 - 165. September, 10-12, 2003. Borovets, Bulgaria. (ISBN 954-90906-6-3). Selected as a chapter in volume 260 of CILT series (Current Issues in Linguistic Theory). John Benjamins Publishers, Amsterdam.
Abstract: In this paper we present a very simple and effective part-of-speech tagger based on Support Vector Machines (SVM). Simplicity and efficiency are achieved by working with linear separators in the primal formulation of SVM, and by using a greedy left-to-right tagging scheme.

5.3 Generation of Resources

Related work is detailed in Subsection 4.1.1.

- Folkert de Vriend, Núria Castell, Jesús Giménez and Giulio Maltese.
LC-STAR: XML-coded Phonetic Lexica and Bilingual Corpora for Speech-to-Speech Translation.
 In Proceedings of the Papillon Workshop on Multilingual Lexical Databases. Grenoble, France. 2004.
Abstract: This paper describes XML encoding of lexica and multilingual corpora and their validation in the framework of the LC-STAR project.
- Victoria Arranz, Núria Castell, Josep Maria Crego, Jesús Giménez, Adrià de Gispert and Patrik Lambert.

Bilingual Connections for Trilingual Corpora: An XML Approach.

In Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04), vol. IV, pages 1459 - 1462. Lisbon, Portugal. 2004 . (ISBN 2-9517408-1-6)

Abstract: An XML representation for a trilingual spontaneous speech corpus for statistical speech-to-speech translation is suggested.

- Victoria Arranz, Núria Castell i Jesús Giménez.
Creació de recursos lingüístics per a la traducció automàtica.
2n Congrés d'Enginyeria en Llengua Catalana. (CELC'04). Andorra, 2004.
(presented also in III Jornadas en Tecnología del Habla. Valencia, Spain. 2004.)
Abstract: Creation of lexica and corpora for Catalan, Spanish and US-English is described.
- Victoria Arranz, Núria Castell and Jesús Giménez.
Development of Language Resources for Speech-to-Speech Translation.
In Proceedings of the International Conference RANLP - 2003 (Recent Advances in Natural Language Processing), pages 26 - 30. September, 10-12, 2003. Borovets, Bulgaria.
Abstract: This paper describes the design and development of a trilingual spontaneous speech corpus for statistical speech-to-speech translation.
- David Conejero, Jesús Giménez, Victoria Arranz, Antonio Bonafonte, Neus Pascual, Núria Castell and Asunción Moreno.
Lexica and Corpora for Speech-to-Speech translation: A Trilingual Approach.
In Proceedings of the 8th European Conference on Speech Communication and Technology (EuroSpeech 2003). September, 1-4, 2003. Geneva, Switzerland. (ISSN 1018-4074)
Abstract: Creation of lexica and corpora for Catalan, Spanish and US-English is described.
- Victoria Arranz, Núria Castell, Jesús Giménez, Hermann Ney and Nicola Ueffing.
Description of language resources used for experiments.
Technical Report Deliverable D4.2, LC-STAR project by the European Community (IST project ref. No. 2001-32216), 2003.
Abstract: This documents describes the language resources used in the first experiments as well as the experiments themselves, in the frame of the LC-STAR project. These experiments are described in detail, providing information on both acquisition and expansion of already existing language resources.
- Victoria Arranz, Núria Castell, Jesús Giménez and Asunción Moreno.
Description of raw corpora.
Technical Report Deliverable 5.3, LC-STAR project by the European Community (IST project ref. No. 2001-32216), 2003.
Abstract: Creation of lexica and corpora for Catalan, Spanish and US-English is described.
- Victoria Arranz, Núria Castell and Jesús Giménez.
Speech Corpora Creation for Tourist Domain.
LSI Department Technical Report (LSI-03-2-T), Technical University of Catalonia, 2003.
Abstract: Creation of lexica and corpora for Catalan, Spanish and US-English is described.

Acknowledgements

This research is being funded by the Spanish Ministry of Science and Technology (ALIADO TIC2002-04447-C02). Our research group, TALP Research Center, is recognized as a Quality Research Group (2001 SGR 00254) by DURSI, the Research Department of the Catalan Government. Authors would like to thank German Rigau for his valuable comments and suggestions. Authors are also grateful to Xavier Carreras for pointing us the availability of the Pharaoh decoder, and to Patrik Lambert for providing us with his implementation of the Simplex Method we used to tune our SMT system.

Bibliography

- ALPAC (1966). *Languages and machines: computers in translation and linguistics* (Technical Report). Automatic Language Processing Advisory Committee (ALPAC), Division of Behavioral Sciences, National Academy of Sciences, National Research Council.
- Alshawi, H., Bangalore, S., & Douglas, S. (1998). Automatic acquisition of hierarchical transduction models for machine translation. *Proceedings of COLING/ACL*.
- Arranz, V., Castell, N., & Giménez, J. (2003). Development of language resources for speech-to-speech translation. *Proceedings of the Fourth RANLP*.
- Atserias, J., Villarejo, L., Rigau, G., Agirre, E., Carroll, J., Magnini, B., & Vossen, P. (2004). The meaning multilingual central repository. *Proceedings of GWC*. Brno, Czech Republic. ISBN 80-210-3302-9.
- Babych, B., & Hartley, T. (2004). Extending the bleu mt evaluation method with frequency weightings. *Proceedings of ACL*.
- Bentivogli, L., Pianta, E., & Ranieri, M. (2005). Multisemcor: an english italian aligned corpus with a shared inventory of senses. *Proceedings of the Meaning Workshop*.
- Brown, P. F., Cocke, J., Pietra, S. A. D., Pietra, V. J. D., Jelinek, F., Lafferty, J. D., Mercer, R. L., & Roossin, P. S. (1990). A statistical approach to machine translation. *Computational Linguistics*, 16, 76–85.
- Brown, P. F., Cocke, J., Pietra, S. A. D., Pietra, V. J. D., Jelinek, F., Mercer, R. L., & Roossin, P. S. (1988). A statistical approach to language translation. *Proceedings of COLING*.
- Brown, P. F., Pietra, S. A. D., Mercer, R. L., & Pietra, V. J. D. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19, 263–311.
- Carreras, X., Chao, I., Padró, L., & Padró, M. (2004). Freeling: An open-source suite of language analyzers. *Proceedings of the 4th LREC*.
- Carreras, X., Márquez, L., & Castro, J. (2005). Filtering-ranking perceptron learning for partial parsing. *Machine Learning*, 59, 1–31.

- Castillo, M., Real, F., Atserias, J., & Rigau, G. (2004). The talp systems for disambiguating wordnet glosses. *Proceedings of SENSEVAL-3 Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, Task: Word-Sense Disambiguation of WordNet Glosses* (pp. 93–96).
- Chandioux, J., & Grimalia, A. (1996). Specialized machine translation. *Proceedings of AMTA* (pp. 206–212).
- Charniak, E. (2001). Immediate-head parsing for language models. *Proceedings of ACL*.
- Charniak, E., Knight, K., & Yamada, K. (2003). Syntax-based language models for machine translation. *Proceedings of MT SUMMIT IX*.
- Church, K. W., & Hovy, E. H. (1993). Good applications for crummy machine translation. *Machine Translation*, 8, 239–258.
- Dorr, B. J. (1994). Machine translation divergences: A formal description and proposed solution. *Computational Linguistics*, 20, 597–633.
- Fellbaum, C. (Ed.). (1997). *WordNet. An Electronic Lexical Database and Some of its Applications*. The MIT Press.
- Fellbaum, C. (Ed.). (1998). *WordNet. An Electronic Lexical Database*. The MIT Press.
- Fox, H. J. (2002). Phrasal cohesion and statistical machine translation. *Proceedings of EMNLP*.
- Germann, U. (2003). Greedy decoding for machine translation in almost linear time. *Proceedings of HLT/NAACL*.
- Germann, U., Jahr, M., Knight, K., Marcu, D., & Yamada, K. (2001). Fast decoding and optimal decoding for machine translation. *Proceedings of ACL*.
- Gildea, D. (2003). Loosely tree-based alignment for machine translation. *Proceedings of ACL*.
- Gildea, D. (2004). Dependencies vs. constituents for tree-based alignment. *Proceedings of EMNLP*.
- Giménez, J., & Màrquez, L. (2003). Fast and accurate part-of-speech tagging: The svm approach revisited. *Proceedings of the Fourth RANLP*.
- Giménez, J., & Màrquez, L. (2004). Svmtool: A general pos tagger generator based on support vector machines. *Proceedings of 4th LREC*.
- Habash, N., & Dorr, B. (2002). Handling translation divergences: Combining statistical and symbolic techniques in generation-heavy machine translation. *Proceedings of AMTA*.
- Hovy, E., Hermjakob, U., & Lin, C.-Y. (2001). The use of external knowledge of factoid qa. *Proceedings of TREC*.

- Kingsbury, P., Palmer, M., , & Marcus, M. (2002). Adding semantic annotation to the penn treebank. *Proceedings of HLT*.
- Knight, K. (1999). Decoding complexity in word-replacement translation models. *Computational Linguistics*, 25.
- Knight, K., Al-Onaizan, Y., Purdy, D., Curin, J., Jahr, M., Lafferty, J., Melamed, D., Smith, N., Och, F. J., & Yarowsky, D. (1999). *Final report of johns hopkins 1999 summer workshop on statistical machine translation* (Technical Report). Johns Hopkins University.
- Koehn, P. (2003a). *Europarl: A multilingual corpus for evaluation of machine translation* (Technical Report). <http://people.csail.mit.edu/koehn/publications/europarl/>.
- Koehn, P. (2003b). *Noun phrase translation*. Doctoral dissertation, University of Southern California.
- Koehn, P. (2004). Pharaoh: a beam search decoder for phrase-based statistical machine translation models. *Proceedings of AMTA*.
- Koehn, P., & Knight, K. (2002). Chunkmt: Statistical machine translation with richer linguistic knowledge. Draft.
- Koehn, P., Och, F. J., & Marcu, D. (2003). Statistical phrase-based translation. *Proceedings of HLT/NAACL*.
- Lin, C.-Y., & Hovy, E. (2002). *Automatic evaluation of machine translation quality using n-gram co-occurrence statistics* (Technical Report). National Institute of Standards and Technology.
- Lin, C.-Y., & Och, F. J. (2004a). Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. *Proceedings of ACL*.
- Lin, C.-Y., & Och, F. J. (2004b). Orange: a method for evaluating automatic evaluation metrics for machine translation. *Proceedings of COLING*.
- Lin, D. (2004). A path-based transfer model for machine translation. *Proceedings of COLING*.
- Marcu, D., & Wong, W. (2002). A phrase-based, joint probability model for statistical machine translation. *Proceedings of EMNLP*.
- Marcus, M. P., Santorini, B., & Marcinkiewicz, M. A. (1994). Building a large annotated corpus of english: the penn treebank. *Computational Linguistics*, 19, 313–330.
- Melamed, I. D. (2004). Statistical machine translation by parsing. *Proceedings of ACL*.
- Melamed, I. D., Green, R., & Turian, J. P. (2003). Precision and recall of machine translation. *Proceedings of HLT/NAACL*.
- Mihalcea, R., & Moldovan, D. (1999). An automatic method for generating sense tagged corpora. *Proceedings of AAAI*.

- Mihalcea, R., & Moldovan, D. (2001). extended wordnet: Progress report. *Proceedings of NAACL Workshop on WordNet and Other Lexical Resources*. Pittsburgh, PA.
- Márquez, L., Comas, P., Giménez, J., & Català, N. (2005). Semantic role labeling as sequential tagging. *Proceedings of CoNLL*.
- Navarro, B., Civit, M., Martí, M. A., Marcos, R., & Fernández, B. (2003). Syntactic, semantic and pragmatic annotation in cast3lb. *Proceedings of SProLaC*.
- Nirenburg, S., Somers, H., & Wilks, Y. (Eds.). (2003). *Readings in machine translation*. The MIT Press.
- Och, F. J. (2002). *Statistical machine translation: From single-word models to alignment templates*. Doctoral dissertation, RWTH Aachen, Germany.
- Och, F. J., Gildea, D., Khudanpur, S., Sarkar, A., Yamada, K., Fraser, A., Kumar, S., Shen, L., Smith, D., Eng, K., Jain, V., Jin, Z., & Radev, D. (2003). *Final report of johns hopkins 2003 summer workshop on syntax for statistical machine translation* (Technical Report). Johns Hopkins University.
- Och, F. J., Gildea, D., Khudanpur, S., Sarkar, A., Yamada, K., Fraser, A., Kumar, S., Shen, L., Smith, D., Eng, K., Jain, V., Jin, Z., , & Radev, D. (2004). A smorgasbord of features for statistical machine translation. *Proceedings of HLT/NAACL*.
- Och, F. J., & Ney, H. (2000). Improved statistical alignment models. *Proceedings of ACL*.
- Och, F. J., & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29, 19–51.
- Och, F. J., Tillmann, C., & Ney, H. (1999). Improved alignment models for statistical machine translation. *Proceedings of EMNLP*.
- Och, F. J., Ueffing, N., & Ney, H. (2001). An efficient a* search algorithm for statistical machine translation. *Proceedings of Data-Driven Machine Translation Workshop* (pp. 55–62).
- Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2001). *Bleu: a method for automatic evaluation of machine translation, ibm research report, rc22176* (Technical Report). IBM T.J. Watson Research Center.
- Schafer, C., & Yarowsky, D. (2003). Statistical machine translation using coercive two-level syntactic transduction. *Proceedings of EMNLP*.
- Shen, L., Sarkar, A., & Och, F. J. (2004). Discriminative reranking for machine translation. *Proceedings of HLT/NAACL*.
- Stolcke, A. (2002). Srilm - an extensible language modeling toolkit. *Proceedings of ICSLP*.
- Tufis, D., Ion, R., & Ide, N. (2004). Fine-grained word sense disambiguation based on parallel corpora, word alignment, word clustering and aligned wordnets. *Proceedings of COLING*.

- Turian, J. P., Shen, L., & Melamed, I. D. (2003). Evaluation of machine translation and its evaluation. *Proceedings of MT SUMMIT IX*.
- Villarejo, L., Màrquez, L., Agirre, E., Martínez, D., Magnini, B., Strapparava, C., McCarthy, D., Montoyo, A., & Suárez, A. (2004). The "meaning" system on the english all-words task. *Proceedings of SENSEVAL-3 Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, Task: English All Words*.
- Wang, Y.-Y., & Waibel, A. (1998). Modeling with structures in statistical machine translation. *Proceedings of COLING/ACL*.
- Weaver, W. (1955). *Translation (1949)*. Machine Translation of Languages. Cambridge, MA: MIT Press.
- William H. Press, Saul A. Teukolsky, W. T. V., & Flannery, B. P. (2002). *Numerical recipes in c++: the art of scientific computing*. Cambridge University Press.
- Wu, D. (1997). Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23, 377–404.
- Yamada, K. (2002). *A syntax-based translation model*. Doctoral dissertation, University of Southern California.
- Yamada, K., & Knight, K. (2001). A syntax-based statistical translation model. *Proceedings of ACL*.
- Yamada, K., & Knight, K. (2002). A decoder for syntax-based statistical mt. *Proceedings of ACL*.
- Zhang, H., & Gildea, D. (2004). Syntax-based alignment: Supervised or unsupervised? *Proceedings of COLING*.