2

Sentence Alignment by Means of Cross-Language Information Retrieval

Marta R. Costa-jussà[1] and Rafael E. Banchs[2]
[1]*Barcelona Media Innovation Center, Spain*
[2] *Institute for Infocomm Research, Singapore*

1. Introduction

In this chapter, we focus on the specific problem of sentence alignment given two comparable corpora. This task is essential to some specific applications such as parallel corpora compilation Utiyama & Tanimura (2007) and cross-language plagiarism detection Potthast et al. (2009).

We address this problem by means of a cross-language information retrieval (CLIR) system. CLIR deals with the problem of finding relevant documents in a language different from the one used in the query. Different strategies are used, from ontology based Soerfel (2002) to statistical tools. Latent Semantic Analysis can be used to get a list of parallel words Codina et al. (2008). Multidimensional Scaling projections Banchs & Costa-jussà (2009) can also be used in order to find similar documents in a cross-lingual environment. Other techniques are based on machine translation, where the search is performed over translated texts Kishida (2005). Within this framework, two basic components should be distinguished: a translation model, and a retrieval model that may work as in the monolingual case. The translation can be faced either in the query, or in the document. In the case of document translation, statistical machine translation systems can be used for translating document collections into the original query language. In the case of query translation, the challenges of deciding how a term might be written in another language, which of the possible translations should be retained, and how to weight the importance of translation alternatives when more than one translation is retained should be considered.

Here, we use the query translation approach. Then, a segment of text in a given source language is used as query for recovering a similar or equivalent segment of text in a different target language. Given that we are using complete sentences which provide a certain context for the terms to be translated, we do not have the disadvantages mentioned in the above lines. Particularly, when using the query translation approach, we investigate if using either a rule-based or a statitical-based machine translation system influence the final quality of the sentence alignment. Additionally, we test if standard automatic MT metrics are correlated with the standards metrics of the sentence alignment.

Rule-based machine translation (RBMT) systems were the first commercial machine translation systems. Much more complex than translating word to word, these systems develop linguistic rules that allow the words to be put in different places, to have different meaning depending on context, etc. RBMT technology applies a set of linguistic rules in three

different phases: analysis, transfer and generation. Therefore, a rule-based system requires: syntax analysis, semantic analysis, syntax generation and semantic generation.
Statistical Machine Translation (SMT), a corpus-based approach, is a more complicated form of word translation, where statistical weights are used to decide the most likely translation of a word. Modern SMT systems are phrase-based rather than word-based, and assemble translations using the overlap in phrases.

2. Organization of the chapter

The rest of this chapter is structured as follows. Next section describes several sentence alignment approaches. Section 4 reports the motivation of our CLIR approach. Section 5 describes in detail how our sentence alignment system works. Section 6 describes the two machine translation approaches that are used and compared in this chapter: rule-based and statistically-based. Next, experimental framework and the proposed methodology are illustrated by performing cross-language text matching at the sentence level on a tetra-lingual document collection. Also, within this section, the performance quality of the implemented systems is compared, showing that in this application the statistical system provides better results than the rule-based system. Section 8 reports the translation quality of both translation systems and reports the correlation among translation quality and cross-language sentence matching quality. Finally, in section 9, most relevant conclusions derived from the experimental results are presented.

3. Related work

Sentence alignment has been approached from different perspectives. In the following subsections we briefly describe some well-known methods.

- Gale & Church (1993) proposed a sentence alginer provided a probability score for each sentence pair based on sentence-length (number of characters). Their method use dynamic programming to find maximum likelihood alignment.
- The Bilingual Sentence Aligner Moore (2002) combines sentence length based method with word correspondence. It makes a first pass based on sentence length and a second pass based on IBM Model-1. The former is based on the distribution of length variable and the latter is trained during runtime and uses alignments obtained from the first pass. The larger corpus size, the more effective (better model of distribution of word length variable and word correspondence).
- Hunalign Varga et al. (2005) uses the diagonal of the alignment matrix, plus a bias of 10%. The weights are a combination of length-based and dictionary-based similarity. If there is no dictionary, they do length-based, estimate dictionary from result and reiterate once. The main problems is that it is not designed to handle corpora of over 20k sentences, it copes by splitting larger corpora and this causes worse dictionary estimates.
- Gargantua Braune & Fraser (2010) is an alignment model similar to Moore (2002), but it introduces differences in pruning and search strategy.
- Bleualign Senrich & Volk (2010) is based on automatic translation of source text. It uses dynamic programming to find path that maximizes BLEU scorePapineni et al. (2001) between target text and translation of source text.

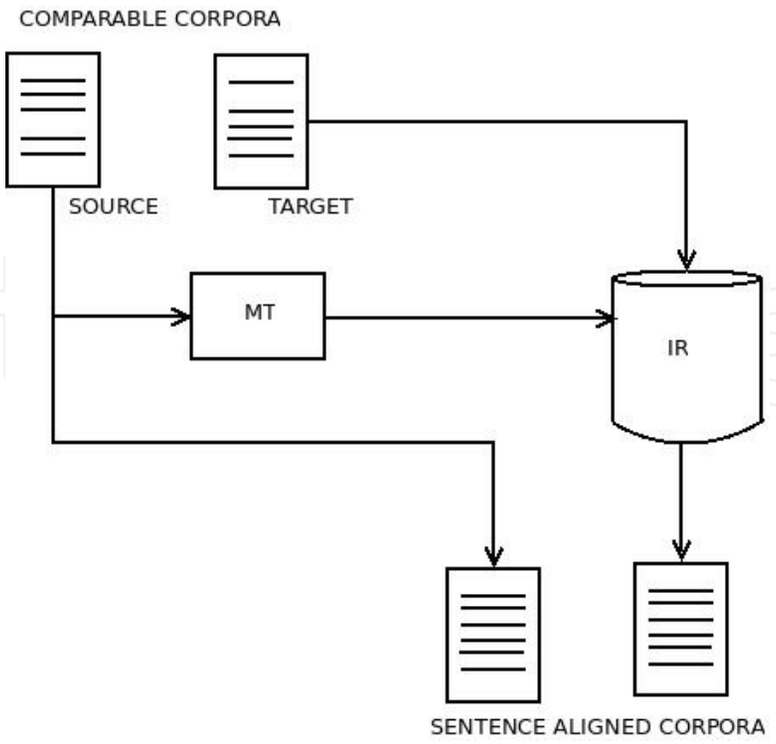

Fig. 1. Block Diagram of the CLIR approach for Sentence Alignment.

4. Motivation

CLIR systems are becoming more and more accurate due to the improvement in machine translation and information retrieval quality. As fas as we are concerned, CLIR have never been used before for sentence alignment. However, with this study, we are demonstrating that it is a nice shot to try. Building a CLIR system is relatively easy if using available tools.
In addition to testing a new methodology for sentence alignment, we want to experiment with different machine translation systems. Particularly, we want to compare two translation systems from different core technologies: rule-based and statistical. This two types of MT commit different types of errors, which may have different effects on the sentence alignment challenge. Although it is not objective of this work, we also report the correlation between translation quality in terms of BLEU and sentence alignment quality.

5. Sentence alignment based on cross-language information retrieval

A cross-language information retrieval (CLIR) system can be used for sentence alignment. The idea is to use a sentence as a query and search for the indexed sentence that matches best. One of the most popular systems in CLIR is the query translation approach which consists of concatenating a machine translation system and a monolingual information retrieval system. See the block diagram in Figure 1.
Basically, an information retrieval (IR) system uses a query to find objects that are indexed in a database. Several documents may match the same query but with different degrees of relevance. In order to make information retrieval efficient, the queries and documents are typically transformed into a suitable representation. One of the most popular representations is the vector space model where documents and queries are represented as vectors, each

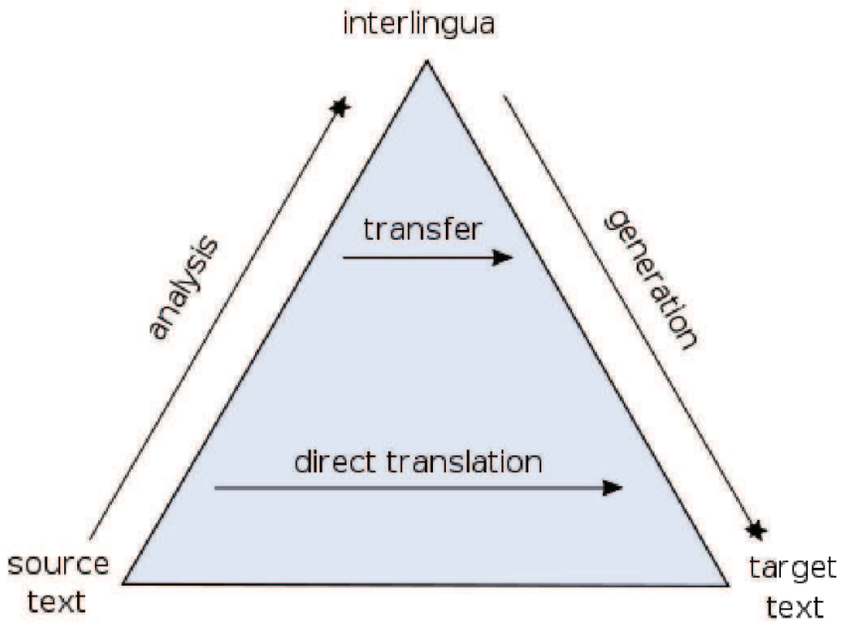

Fig. 2. Machine translation approaches.

dimension corresponding to a separate term. Usually, terms are weighted with the term frequency and inverse document frecuency (tf-idf) scheme.
The main challenge in CLIR with respect to IR is that the query language is different from the document language. We approach the problem of sentence aligning by operating a machine-translation-based CLIR system at the sentence level over a bilingual comparable corpus. In this context, we are comparing the performance of two machine translation systems with different core technologies: rule-based and statistical.

6. Machine translation core technologies

As mentioned, there are different core tecnologies in machine translation. Corpus-based approaches (such as Statistical) use a direct translation and rule-based approaches use a transfer translation. See Figure 2 [1]. As follows we briefly describe the two technologies.

6.1 Rule-based machine translation

Rule-based machine translation (RBMT) systems develop linguistic rules that allow the words to be put in different places, to have different meaning depending on context, etc. The Georgetown-IBM experiment in 1954 was one of the first rule-based machine translation systems and Systran was one of the first companies to develop RBMT systems.
RBMT methodology applies a set of linguistic rules in three different phases: analysis, transfer and generation. Therefore, a rule-based system requires: syntax analysis, semantic analysis, syntax generation and semantic generation. In general terms, RBMT generates the target text given a source text following the next steps.
Given a source text, the first step is to segment it, for instance, by expanding elisions or marking set phrases. These segments are then looked up in a dictionary. This search returns

[1] http://en.wikipedia.org/wiki/Machine_translation

the base form and tags for all matches (morphological analyser). Afterwards, the task is to resolve ambiguous segments, i.e. source terms that have more than one match, by choosing only one (part of speech tagger). Additionally, a RBMT system may add a lexical selection to choose between alternative meanings. After the module taking care of the lexical selection, two modules follow, namely the structural and the lexical transfers. The former consists of looking up disambiguated source-language base work to find the target-language equivalent. The latter consists in: (1) flagging grammatical divergences between source language and target language, e.g. gender or number agreement; (2) creating a sequence of chunks; (3) reordering or modifying chunk sequences; and (4) substituting fully-tagged target-language forms into the chunks. Then, tags are used to deliver the correct target language surface form (morphological generator). Finally, the last step is to make any necessary orthographic change (post-generator).
One of the main problems of translation is choosing the correct meaning, which involves a classification or disambiguation problem. In order to improve the accuracy, it is possible to apply a method to disambiguate meanings of a single word. Machine learning techniques automatically extract the context features that are useful for disambiguating a word.
RBMT systems have a big drawback: the construction of such systems demands a great amount of time and linguistic resources, thus resulting very expensive. Moreover, in order to improve the quality of a RBMT it is necessary to modify rules, which requires more linguistic knowledge. The modification of one rule cannot guarantee that the overall accuracy will be better. However, using rule-based methodology may be the only way to build an MT system when dealing with minor languages, given that SMT requires massive amounts of sentence-aligned parallel text. RBMT may use linguistic data without access to existing machine-readable resources. Moreover, it is more transparent: errors are easier to diagnose and debug.

6.2 Statistical machine translation

Statistical Machine Translation (SMT), which started with the CANDIDE system Berger et al. (1994), is, at its most basic, a more complicated form of word translation, where statistical weights are used to decide the most likely translation of a word. Modern SMT systems are phrase-based rather than word-based, and assemble translations using the overlap in phrases. The main goal of SMT is the translation of a text given in some source language into a target language by maximizing the conditional propability of the translated sentence given the source one. A source string $s_1^J = s_1 \dots s_j \dots s_J$ is translated into a target string $t_1^I = t_1 \dots t_i \dots t_I$. Among all possible target strings, the goal is to choose the string with the highest probability:

$$\tilde{t_1^I} = \underset{t_1^I}{argmax}\, P(t_1^I | s_1^J)$$

where I and J are the number of words in the target and source sentences, respectively.
The first SMT systems were reformulated using Bayes' rule. In recent systems, such an approach has been expanded to a more general maximum entropy approach in which a log-linear combination of multiple feature functions is implemented (Och, 2003). This approach leads to maximising a linear combination of feature functions:

$$\tilde{t} = \underset{t}{argmax} \left\{ \sum_{m=1}^{M} \lambda_m h_m(t, s) \right\}.$$

The job of the translation model, given a target sentence and a foreign sentence, is to assign a probability that t_1^I generates s_1^J. While these probabilities can be estimated by thinking about how each individual word is translated, modern statistical MT is based on the intuition that a better way to compute these probabilities is by considering the behavior of phrases (sequences of words). The phrase-based statistical MT uses phrases as well as single words as the fundamental units of translation. Phrases are extracted from multiple segmentations of the aligned bilingual corpora and their probabilities are estimated by using relative frequencies.
The translation problem has also been approached from the finite-state perspective as the most natural way for integrating speech recognition and machine translation into a speech-to-speech translation system (Bangalore & Riccardi, 2000; Casacuberta, 2001; Vidal, 1997). The Ngram-based system implements a translation model based on this finite-state perspective (de Gispert & Mariño, 2002) which is used along with a log-linear combination of additional feature functions (Mariño et al., 2006).
In addition to the translation model, SMT systems use the language model, which is usually formulated as a probability distribution over strings that attempts to reflect how likely a string occurs inside a language (Chen & Goodman, 1998). Statistical MT systems make use of the same *n*-gram language models as speech recognition and other applications do. The language model component is monolingual, so acquiring training data is relatively easy.
The lexical models allow the SMT systems to compute another probability to the translation units based on the probability of translating word per word. The probability estimated by lexical models tends to be in some situations less sparse than the probability given directly by the translation model. Many additional feature functions can also be introduced in the SMT framework to improve the translation, like the word or the phrase bonus.

6.3 Challenges of RBMT and SMT

State-of-the-art rule-based MT approaches have the following challenges:

- *Semantic*. RBMT approaches concentrate on a local translation. Usually, this translation tends to be literal and it lacks of fluency. Additionally, words may have different meanings depending on their grammatical and semantic references.
- *Lexical*. Words which are not included in the dictionary will have no translation. When keeping the system updated, new language words have to be introduced in the dictionary.

State-of-the-art statistical MT approaches have the following challenges:

- *Syntactic*. The main challenge in this category is word reordering, which can be of two natures: long reordering, as when translating between languages with different structures (SVO versus VSO), and short reorderings, as such involving relative locations of modifiers and nouns Costa-jussà & Fonollosa (2009); Tillmann & Ney (2003); Zhang et al. (2007).
- *Morphological*. Here there are chanllenges as gender and number agreement. For instance, keeping number agreement when translating from English to Spanish in structures such as*Noun + Adjective* de Gispert et al. (2006); Nießen & Ney (2004).
- *Lexical*. Here there are the Out-of-Vocabulary words which can not be translated. The main causes of out of vocabulary words is the dependency with the training data. In most SMT approaches, the limitation of training data, domain changes and morphology are not taken into account. Approaches such as the one from Langlais & Patry (2007) try to deal with these challenges.

The semantic and lexical problems may affect more to a CLIR system than the syntactic and morphological errors, taking into account that IT systems work with bag-of-words and use words and stems.

7. Experiments

As already mentioned in the introduction, in this work, we focus on the problem of sentence alignment given two comparable corpora. In this particular task, a segment of text in a given source language is used as query for recovering an equivalent segment of text in a different target language. In this section, we evaluate a conventional query translation approach first described by Chen & Bao (2009) which considers a cascade combination of a machine translation system and a monolingual IR system. We use two machine translation systems with different core technologies: a rule-based and a statistical-based machine translation systems.

7.1 Multilingual sentence dataset

The dataset considered for the experiments is a multilingual sentence collection that was extracted from the Spanish Constitution, which is available for downloading at the Spanish government's main web portal: *www.la-moncloa.es*. In this website, all constitutional texts are available in five different languages, including the four official languages of Spain: Spanish, Catalan, Galego and Euskera, as well as English. Given that the MT systems used do not provide Euskera translation, we limited the experiments to four languages. The texts are organized in 169 articles plus some additional regulatory dispositions. All texts were segmented into sentences and the resulting collection was filtered according to sentence length. More specifically, sentences having less than five words were discarded aiming at eliminating titles and some other non-relevant information. Moreover, we had to perform a manual postprocessing to correct some errors in the sentence alignment. Table 1 summarizes the main statistics for both the overall collection.Table 2 shows a sentence example.

Collection	English	Spanish	Catalan	Gallego
Sentences	611	611	611	611
Running words	15285	14807	15423	13760
Vocabulary	2080	2516	2523	2667
Average sent. length	25.01	24.23	25.24	22.52

Table 1. Corpus statistics.

7.2 Evaluation of the methodology

The system to be considered implements a query translation strategy followed by a standard monolingual information retrieval approach.
For the query translation step, we used the following MT systems:

1. A rule-based system implemented with the Opentrad platform[2]. This system Ramírez-Sánchez et al. (2006) constitutes a state-of-the-art machine translation service that provides automatic translation among several language pairs including the four Spanish languages plus English, Portuguese and French. See Figure 3. Besides, Opentrad is

[2] http://www.opentrad.com/

Language	Sentence example
English	The entire wealth of the country in its different forms, irrespective of ownership, shall be subordinated to the general interest.
Spanish	Toda la riqueza del país en sus distintas formas y sea cual fuere su titularidad está subordinada al interés general.
Catalan	Tota la riquesa del país en les seves diverses formes, i sigui quina sigui la titularitat, resta subordinada a l'interès general.
Gallego	Toda a riqueza do país nas súas distintas formas e calquera que sexa a súa titularida de está subordinada ó interese xeral.

Table 2. Sentence example from the Spanish Constitution.

designed to be adapted and configured according to user needs, allowing its integration with other systems. Opentrad's design allows for its customization and personalization both from a linguistic point of view, adopting the style book of an organization, and from a technical point of view, allowing its integration into IP networks or a full integration with other systems.

Fig. 3. Opentrad screenshot

2. A statistical-based system implemented with the Google API translation[3]. See Figure 4. Google's research group has developed its statistical translation system for the language pairs now available on Google Translate. Their system, in brief, feeds the computer with billions of words of text, both monolingual text in the target language, and aligned text consisting of examples of human translations between the languages. Then, they apply statistical learning techniques to build a translation model.

[3] http://code.google.com/apis/ajaxlanguage/

The *detect language* option automatically determines the language of the text the user is translating. The accuracy of the automatic language detection increases with the amount of text entered. Google is constantly working to support more languages and introduce them as soon as the automatic translation meets their standards. In order to develop new systems, they need large amounts of bilingual texts.

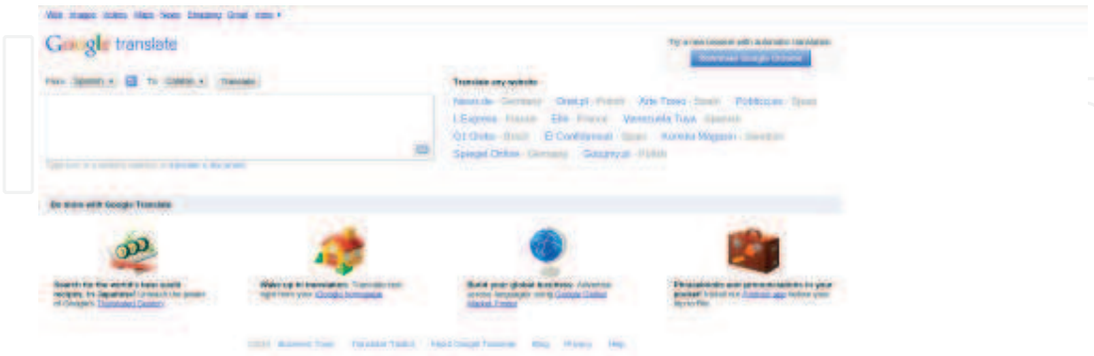

Fig. 4. Google Translate screenshot

The monolingual information retrieval step was implemented by using Solr, which is an XML-based open-source search server based on the Apache-Lucene search library [4]. See Figure 5. Particularly, Solr is the popular, blazing fast open source enterprise search platform from the Apache Lucene project. Its major features include powerful full-text search, hit highlighting, faceted search, dynamic clustering, database integration, and rich document (e.g., Word, PDF) handling. Solr is highly scalable, providing distributed search and index replication, and it powers the search and navigation features of many of the world's largest internet sites.

Solr is written in Java and runs as a standalone full-text search server within a servlet container such as Tomcat. Solr uses the Lucene Java search library at its core lecelfor full-text indexing and search, and has REST-like HTTP/XML and JSON APIs that make it easy to use from virtually any programming language. Solr's powerful external configuration allows it to be tailored to almost any type of application without Java coding, and it has an extensive plugin architecture when more advanced customization is required.

Table 3 summarizes the results obtained from the comparative evaluation between the two contrastive systems. We measure the quality of the system in terms of accuracy. We show top-1 and top-5 results. The former reports the percentage of times that the correct result coincides with the top-ranked sentence retrieved by the system and the latter reports the percentage of times that the correct result is within the top-five ranked sentences retrieved by the system.

The query translation system using statistical translation performs slightly better than the rule-based system. It is worth noticing the high quality of cross-language sentence matching using the query translation approach. This high quality is mainly due to the quality of translation.

Figure 6 shows some examples of the system performance.

[4] http://lucene.apache.org/solr/tutorial.html

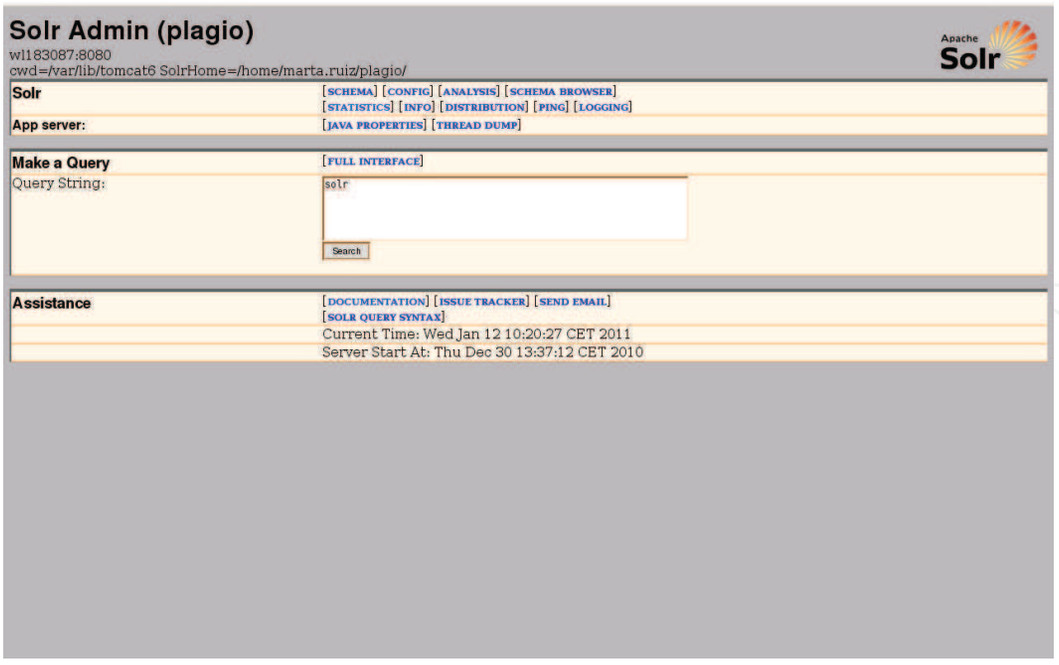

Fig. 5. SOLR screenshot

Source language	System	Target language							
		English		Spanish		Catalan		Gallego	
		top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5
English	rule-based	100	100	95.0	99.5	92.0	96.0	93.0	96.0
	statistical	100	100	**100**	**100**	**100**	**100**	**97**	**100**
Spanish	rule-based	96.0	99.0	100	100	100	100	**99.0**	**100**
	statistical	**97.5**	**100**	100	100	100	100	96	99
Catalan	rule-based	95.5	99.0	100	100	100	100	93.5	97.0
	statistical	**99**	**99.5**	100	100	100	100	**96**	**99**
Gallego	rule-based	93.5	97.5	**99.5**	**99.5**	83.5	90.5	100	100
	statistical	**97**	**98.5**	97	99	**97.5**	**99**	100	100

Table 3. Comparative results.

8. Correlation between machine translation quality and sentence matching performance

We evaluate the quality of the translation in terms of BLEUPapineni et al. (2001) and PER, see table 4. BLEU stands for Bilingual Evaluation Understudy. It is a quality metric and it is defined in a range between 0 and 1 (or in percentage between 0 and 100), 0 meaning the worst translation (where the translation does not match the reference in any word), and 1 the perfect translation. BLEU computes lexical matching accumulated precision for n-grams up to length four Papineni et al. (2001).

PER stands for Position-Independent Error Rate (PER) and it is computed on a sentence-by-sentence basis. The main difference with WER (Word error rate) is that it does not penalise the wrong order in the translation. WER (McCowan et al., 2004) is a standard speech recognition evaluation metric. A general difficulty of measuring performance lies in the fact that the translated word sequence can have a different length from the reference

Source: Si la moción de censura no fuere aprobada por el Congreso, sus signatarios no podrán presentar otra durante el mismo período de sesiones.
Translation-Google: Si la moció de censura no fos aprovada pel Congrés, els signataris no podran presentar cap més durant el mateix període de sessions.
Retrieval: Si la moció de censura no fos aprovada pel Congrés, els signataris no en podran presentar cap més durant el mateix període de sessions.
Translation-Opentrad: Si la moció de censura no anàs aprovada pel Congrés, els seus signataris no podrán presentar una altra durant el mateix període de sessions.
Retrieval: Si la moció de censura no fos aprovada pel Congrés, els signataris no en podran presentar cap més durant el mateix període de sessions.
Reference: Si la moció de censura no fos aprovada pel Congrés, els signataris no en podran presentar cap més durant el mateix període de sessions.

Source:The Congress may require political responsibility from the Government by adopting a motion of censure by overall majority of its Members.
Translation-Google: O Congreso pode esixir responsabilidade política do Goberno, aprobando unha moción de censura por maioría absoluta dos seus membros.
Retrieval: O Congreso dos Deputados pode esixi-la responsabilidade política do Goberno mediante a adopción por maioría absoluta da moción de censura.
Translation-Opentrad: O Congreso pode requirir responsabilidade política desde o Goberno por adoptar unha moción de censure por maioría total dos seus Membros.
Retrieval: O Congreso dos Deputados pode esixi-la responsabilidade política do Goberno mediante a adopción por maioría absoluta da moción de censura.
Reference: O Congreso dos Deputados pode esixi-la responsabilidade política do Goberno mediante a adopción por maioría absoluta da moción de censura.

Source:O Pleno poderá, con todo, avocar en calquera momento o debate e votación de calquera proxecto ou proposición de lei que xa fora obxecto desta delegación.
Translation-Google: The Chamber may, however, take over at any moment the debate and vote on any project or proposed law that had already been the subject of this delegation.
Retrieval: However, the Plenary sitting may at any time demand that any Government or non--governmental bill that has been so delegated be debated and voted upon by the Plenary itself.
Translation-Opentrad: The Plenary will be able to, however, avocar in any moment the debate and vote of any project or proposición of law that already was object of this delegation.
Retrieval: However, the Plenary sitting may at any time demand that any Government or non--governmental bill that has been so delegated be debated and voted upon by the Plenary itself.
Reference: However, the Plenary sitting may at any time demand that any Government or non--governmental bill that has been so delegated be debated and voted upon by the Plenary itself.

Fig. 6. Examples of the system performance.

word sequence (supposedly the correct one). WER is derived from the Levenshtein distance, working at the word level.
We see that Google translator is better than Opentrad in most translation pairs. It may be possible that Google has part of the Spanish Constitution as training material in its system. However, notice that we did not use directly the Spanish constitution that is available from the website *www.la-moncloa.es*, we had to perform a manual postprocessing to correct some errors in the sentence alignment.
After evaluating the quality of translation we computed correlation coefficients between sentence matching accuracies and translation quality metrics. We found out that some of the

Source language	System	Target language							
		English		Spanish		Catalan		Gallego	
		BLEU	PER	BLEU	PER	BLEU	PER	BLEU	PER
English	rule-based	-	-	20.80	49.14	20.02	51.66	17.49	55.34
	statistical	-	-	44.73	31.38	37.98	36.04	16.75	56.27
Spanish	rule-based	20.92	48.53	-	-	68.76	15.65	**72.57**	**14.56**
	statistical	45.57	31.44	-	-	78.55	11.05	32.90	39.78
Catalan	rule-based	20.95	50.56	70.52	14.89	-	-	**54.81**	**23.81**
	statistical	45.86	30.91	87.59	6.24	-	-	29.16	42.49
Gallego	rule-based	18.67	52.47	**75.85**	**12.60**	**57.71**	**22.31**	-	-
	statistical	30.43	41.52	53.02	26.74	43.53	32.79	-	-

Table 4. Comparative results between translation qualities of used rule-based and statistical systems.

computed correlations were quite high, see table 5. All correlations are significant ($p << 0.05$) except for the cases marked with $*$. There is a slightly high correlation between BLEU and top-1 measure in the statistical case, but it is not maintained in the rule-based case. Research in finding an MT measure which is correlated with CLIR quality or sentence alignment quality was not the objective of this work. However, it may be a nice topic for further research.

	system	top-1	top-5	BLEU
top-1	rule-based	-		
	statistical	-		
top-5	rule-based	95.82	-	
	statistical	76.28	-	
BLEU	rule-based	58.17	39.61*	-
	statistical	74.71	53.53	-
PER	rule-based	-55.24	-36.39*	-99.37
	statistical	-75.03	-50.16	-99.46

Table 5. Correlations coeficients.$*$ marks the non-significant correlations.

9. Conclusions

This chapter presented a cross-language sentence matching application. The proposed approach was a query translation cross-language information retrieval system either using a rule-based or a statistical-based translation system.

We tested the performance of rule-based and statistical systems in a multilingual collection based on the Spanish Constitution.

Results show that the statistical-based system performed slightly better than the rule-based system.

Looking at some examples we saw that the errors in sentence matching were different depending on the kind of translation system we were using, which suggests that a system combination strategy could improve the performance.

We evaluated the translation performance of the rule-based and the statistical-based translation systems. The latter performed better in 12 out of 16 translation pairs.

Finally, we saw that translation quality is correlated with the cross-language sentence matching quality, specially in terms of BLEU and top-1 measures.

10. Acknowledgements

This work has been partially funded by the Spanish Department of Science and Innovation through the *Juan de la Cierva* fellowship program.
The authors also want to thank Barcelona Media Innovation Center for its support and permission to publish this research.

11. References

Banchs, R. E. & Costa-jussà, M. (2009). Extracción crosslingue de documentos usando mapas semánticos no-lineales, *SEPLN* 43: 169–176.

Bangalore, S. & Riccardi, G. (2000). Finite-state models for lexical reordering in spoken language translation, *Proc. of the 6th Int. Conf. on Spoken Language Processing, ICSLP'02*, Vol. 4, Beijing, pp. 422–425.

Berger, A. L., Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., Gillett, J. R., Lafferty, J. D., Mercer, R. L., Printz, H. & Ureš, L. (1994). The candide system for machine translation, *HLT '94: Proceedings of the workshop on Human Language Technology*, pp. 157–162.

Braune, F. & Fraser, A. (2010). Improved sentence alignment for symmetrical and asymmetrical parallel corpora, *Coling*, pp. 81–89.

Casacuberta, F. (2001). Finite-state transducers for speech-input translation, *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU*, Trento, pp. 375–380.

Chen, J. & Bao, Y. (2009). Cross-language search: The case of google language tools, *First Monday* 14(3-2).

Chen, S. F. & Goodman, J. T. (1998). An empirical study of smoothing techniques for language modeling, *Technical report*, Harvard University.

Codina, J., Pianta, E., Vrochidis, S. & Papadoupoulos, S. (2008). Integration of semantic, metadata and image search engines with a text search engine for patent retrieval, *Proceedings of ESWC 2008*, Tenerife, Spain.

Costa-jussà, M. & Fonollosa, J. (2009). An ngram-based reordering model, *Computer Speech and Language* 23(3): 362–375.

de Gispert, A., Gupta, D., Popovic, M., Lambert, P., MariÃśo, J., Federico, M., Ney, H. & Banchs, R. (2006). Improving statistical word alignments with morpho-syntactic transformations, *Proc. of 5th Int. Conf. on Natural Language Processing (FinTAL'06)* pp. 368–379.

de Gispert, A. & Mariño, J. (2002). Using x-grams for speech-to-speech translation, *Proc. of the 7th Int. Conf. on Spoken Language Processing, ICSLP'02*, Denver, pp. 1885–1888.

Gale, W. A. & Church, K. W. (1993). A program for aligning sentences in bilingual corpora, *Computational Linguistics* 1(19): 75–102.

González, A. O., Boleda, G., Melero, M. & Badia, T. (2005). Traducción automática estadística basada en n-gramas, *Procesamiento del Lenguaje Natural, SELPN* 35: 69–76.

Kishida, K. (2005). Technical issues of cross-language information retrieval: a review, *Cross-Language Information Retrieval* 41(3): 433–455.

Langlais, P. & Patry, A. (2007). Translating unknown words by analogical learning, *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 877–886.

Mariño, J., Banchs, R., Crego, J., de Gispert, A., Lambert, P., Fonollosa, J. & Costa-jussà, M. (2006). N-gram based machine translation, *Computational Linguistics* 32(4): 527–549.

McCowan, I., Moore, D., Dines, J., Gatica-Perez, D., Flynn, M., Wellner, P. & Bourlard, H. (2004). On the use of information retrieval measures for speech recognition evaluation, *IDIAP-RR 73*, IDIAP, Martigny, Switzerland.

Moore, R. C. (2002). Fast and accurate sentence alignment of bilingual corpora, *AMTA*, pp. 135–144.

Nießen, S. & Ney, H. (2004). Statistical machine translation with scarce resources using morpho-syntactic information, *Computational Linguistics* 30(2): 181–204.

Och, F. (2003). Minimum error rate training in statistical machine translation, *Proc. of the 41th Annual Meeting of the Association for Computational Linguistics*, Sapporo, pp. 160–167.

Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. (2001). Bleu: a method for automatic evaluation of machine translation, IBM Research Report, RC22176.

Potthast, M., Stein, B., Eiselt, A., BarrÃşn, A. & Rosso, P. (2009). Overview of the 1st international competition on plagiarism detection, *Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse*.

Ramírez-Sánchez, G., Sánchez-Martínez, F., Ortiz-Rojas, S., Pérez-Ortiz, J. A. & Forcada, M. L. (2006). Opentrad apertium open-source machine translation system: an opportunity for business and research, *Proceeding of Translating and the Computer 28 Conference*.

Senrich, R. & Volk, M. (2010). Mt-based sentence alginment for ocr-generated parallel texts, *AMTA*, Colorado.

Soerfel, D. (2002). Thesauri and ontologies for digital libraries, *Proceedings of the Joint Conference on Digital Libraries*.

Tillmann, C. & Ney, H. (2003). Word reordering and a dynamic programming beam search algorithm for statistical machine translation, *Computational Linguistics* 29(1): 97–133.

Utiyama, M. & Tanimura, M. (2007). Automatic construction technology for parallel corpora, *Journal of the National Institute of Information and Communications Technology* 54(3): 25–31.

Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V. & Nagy, V. (2005). Parallel corpora for medium density languages, *RANLP*, pp. 590–596.

Vidal, E. (1997). Finite-state speech-to-speech translation, *Proc. Int. Conf. on Acoustics Speech and Signal Processing*, Munich, pp. 111–114.

Zhang, Y., Zens, R. & Ney, H. (2007). Chunk-level reordering of source language sentences with automatically learned rules for statistical machine translation, *Proc. of the Human Language Technology Conf. (HLT-NAACL'06):Proc. of the Workshop on Syntax and Structure in Statistical Translation (SSST)*, Rochester, pp. 1–8.

Speech and Language Technologies
Edited by Prof. Ivo Ipsic

ISBN 978-953-307-322-4
Hard cover, 344 pages
Publisher InTech
Published online 21, June, 2011
Published in print edition June, 2011

This book addresses state-of-the-art systems and achievements in various topics in the research field of speech and language technologies. Book chapters are organized in different sections covering diverse problems, which have to be solved in speech recognition and language understanding systems. In the first section machine translation systems based on large parallel corpora using rule-based and statistical-based translation methods are presented. The third chapter presents work on real time two way speech-to-speech translation systems. In the second section two papers explore the use of speech technologies in language learning. The third section presents a work on language modeling used for speech recognition. The chapters in section Text-to-speech systems and emotional speech describe corpus-based speech synthesis and highlight the importance of speech prosody in speech recognition. In the fifth section the problem of speaker diarization is addressed. The last section presents various topics in speech technology applications like audio-visual speech recognition and lip reading systems.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821