Automatic and human evaluation study of a rule-based and a statistical Catalan-Spanish machine translation systems

Marta R. Costa-jussà*, **Mireia Farrús** †, **José B. Mariño** †, **José A. R. Fonollosa** †

* Barcelona Media Research Center,
Av. Diagonal 177, 08018 Barcelona, Spain
marta.ruiz@barcelonamedia.org
†Universitat Politècnica de Catalunya – TALP Research Center,
Jordi Girona 1-3, 08034 Barcelona, Spain
{mfarrus, canton, adrian}@gps.tsc.upc.edu

Abstract

Machine translation systems can be classified into rule-based and corpus-based approaches, in terms of their core technology. Since both paradigms have largely been used during the last years, one of the aims in the research community is to know how these systems differ in terms of translation quality. To this end, this paper reports a study and comparison of a rule-based and a corpus-based (particularly, statistical) Catalan-Spanish machine translation systems, both of them freely available in the web. The translation quality analysis is performed under two different domains: journalistic and medical. The systems are evaluated by using standard automatic measures, as well as by native human evaluators. Automatic results show that the statistical system performs better than the rule-based system. Human judgements show that in the Spanish-to-Catalan direction the statistical system also performs better than the rule-based system, while in the Catalan-to-Spanish direction is the other way round. Although the statistical system obtains the best automatic scores, its errors tend to be more penalized by human judgements than the errors of the rule-based system. This can be explained because statistical errors are usually unexpected and they do not follow any pattern.

1. Introduction

Machine Translation (MT) is a subfield of computational linguistics that investigates the use of computer software to translate text from one given source language to another target language. Since natural languages are highly complex, MT becomes a difficult task. Many words have multiple meanings, sentences may have various readings, and certain grammatical relations in one language might not exist in another language. Moreover, there are non-linguistic factors such as the need of having a world knowledge to perform a translation. In order to face the MT challenge, many dependencies have to be taken into account. These are often weak and vague, which makes it rarely possible to describe simple and relevant rules that hold without exception for different language pairs.

Increasing computational power picked the current interest in MT. As a consequence, available machine translation systems in the web are becoming more and more popular. Nowadays, the most widely used MT systems use the rule-based and the statistical approaches. Moreover, there have been several research works which combine both technologies. Our study is intended to reinforce the system combination research works (Matusov et al., 2008) by further analysing both the structure of the two technologies. In the specific case of Catalan-Spanish translation, there are available systems that use either a rule-based or a statistical MT system. Section 2. provides a comparison of the rule-based and the statistical-based systems at the level of core technology. Also it describes the general advantages and disadvantages of each approach. Section 3. reports a brief description of two Catalan-Spanish MT freely-available systems: Translendium as rule-based system and UPC as statistical system. Section 4. describes the experimental framework used to compare the cited systems. Section 5. and 6. provide a deep comparison by using automatic and human evaluation, respectively. Finally, Section 7. presents the conclusions.

2. Rule-based vs. Statistical-based machine translation

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. 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 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. Speaking in general terms, RBMT generates the target text given a source text following steps shown in Figure 1.

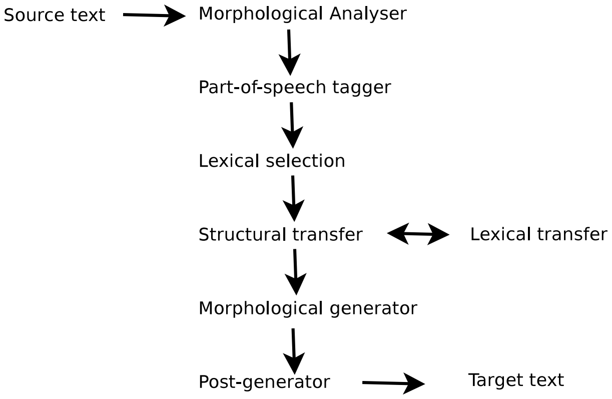


Figure 1: *Architecture of the RBMT approach.*

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 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. As follows, it takes place the structural and lexical transfer. 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 technology may be the only way to build an MT system, given that SMT requires massive amounts of sentence-aligned parallel text (*is there such a resource for Icelandic?*). Additionally, the use of linguists may be a good choice. RBMT may use linguistic data elicited by speakers without access to existing machine-readable resources and it is more transparent: errors are easier to diagnose and debug.

Statistical Machine Translation (SMT) 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 MT is the translation of a text given in some source language into a target language. A source string $s_1^J = s_1 \dots s_j \dots s_J$ is to be translated into a target string $t_1^I = t_1 \dots t_i \dots t_I$. In SMT, 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 of the target and source sentence, 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 overall architecture of this statistical translation approach is summarised in Figure 2.

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 intuition of phrase-based statistical MT is to use phrases as well as single words as the fundamental units of translation. Phrases are

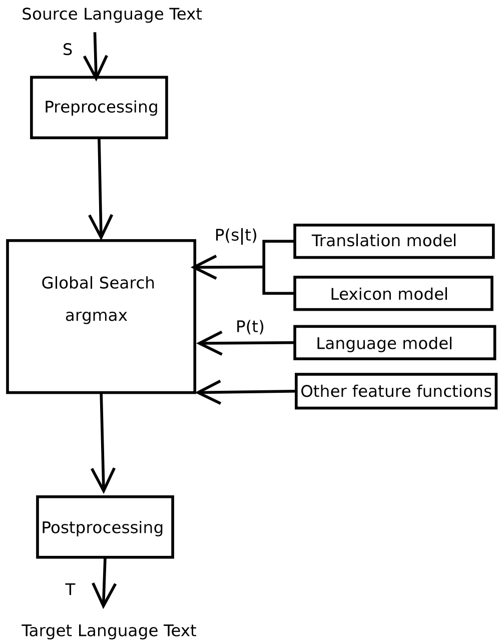


Figure 2: *Architecture of the SMT approach based on the log-linear framework approximation.*

estimated from multiple segmentation of the aligned bilingual corpora 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 (Vidal, 1997; Bangalore and Riccardi, 2000; Casacuberta, 2001). The Ngram-based system implements a translation model based on this finite-state perspective (de Gispert and 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 and Goodman, 1998). Statistical MT systems make use of the same n-gram language models as do speech recognition and other applications. 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 of the unit. 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.

Different approaches of MT have complementary pros and cons. Main core advantages and disadvantages of both approaches are shown in Figure 3. At this point, no comparison is made at the level of performance, which will be studied later in the paper.

3. Freely available Catalan-Spanish MT systems: brief description

Two Catalan-Spanish MT systems available in the web are introduced in this section. Obviously, there are more translation systems available in the web. However, we decided to take two systems (one of each core technology) to make a first comparison study. As RBMT system we include Translendium which is the official Catalan Government machine translation system. As statistical SMT system, we include the UPC system which was developed in the Universitat Politècnica de Catalunya. Next, a brief description of each system and its corporation is presented.

- **Translendium** (*http://www.translendium.com*)

  Translendium S.L., located in Barcelona, develops the Lucy Translator (LT) machine translation system, previously called Comprendium. Translendium is a Catalan company, subsidiary of the European group Lucy Software. The Translendium team is composed by linguists, lexicographers and computer scientists with more than 15 years experience in the machine translation field.

  The system consists of a translation engine, with a modular structure of computational grammars and lexicons that makes possible to carry out a morphosyntactic analysis of the source text and then transfers it into the target language. This engine can be connected to translation memory modules and to a professional lexicon editor. Additionally, it can be accessed through a multi-user task distribution server either from a web client or from a professional single user client.

- **UPC** (*http://www.n-ii.org/*)

  This system has been developed at the Universitat Politècnica de Catalunya (UPC) and the work has been funded by the European Union under the integrated project TC-STAR: Technology and Corpora for Speech to Speech Translation (IST-2002-FP6-506738), and the Spanish Government under the project AVIVAVOZ: Technologies for Speech-to-Speech Translation (TEC2006-13694-C03-01).

  The machine translation engine is based on an N-gram translation model integrated in an opti-

Advantages	Disadvantages
RBMT	
Based on linguistic theories	Requires linguistic rules and dictionaries
Adequate for languages with limited resources	Human Language Inconsistency (i.e. exception to the rules)
Does not require many computational resources	Disambiguation problems
Easy to perform error analysis	Local translations, Language dependent
	Expensive to maintain and extend
SMT	
No linguistic knowledge required	Requires parallel text
Reduces the human cost	Requires high computational resources
Easy to build	Difficult to perform error analysis
Easy to maintain (if data available)	Problems with pairs of languages with different morphology/order
Trained with human translations (extract knowledge from corpus)	No linguistic background
Independent from the the pair of languages	

Figure 3: *Brief comparison of the advantages and disadvantages of the RBMT and SMT systems.*

mised log-linear combination of additional features. Thus the system is mainly statistical; however, a series of additional linguistic rules is included in order to solve some errors caused by the statistical translation, such as the ambiguity in adjective and possessive pronouns, orthographic errors or time expressions, among others.

Since time expressions differ largely in both languages, a detection-translation-generation module is added. The same procedure is used in the numbers, since many of them were were not included in the training corpus. Other unknown words apart from numbers were solved by including a Spanish-Catalan dictionary as a post-process after the translation, and by a spell checker in order to avoid wrong-written —and thus unknown— words as input. The system is continuously actualised by adding new corpora and the feedback of the users.

In the following sections, these systems are compared and they are all used with their respective versions date of *1st of April 2009*.

4. Experimental Framework

The aim of this section is to define an experimental framework in which the systems presented above can be compared both manually and automatically. The idea is to report the main differences in performance terms between the state-of-the-art rule-based and statistical systems. In the following sections, the results are reported through two different evaluation types: automatic and manual.

Two test sets are defined in order to perform the evaluation. The first one is a compilation of journalistic material. The Spanish source test corpus consists of 711 sentences extracted from *El País* and *La Vanguardia* newspapers, and the Catalan source test corpus consists of 813 sentences extracted from the *Avui* newspaper and transcriptions from the TV program *Àgora*. For each set and each direction of translation, two manual references are provided. Table 1 shows the number of sentences, words and vocabulary used for each language.

	Spanish	Catalan
Sentences	711	813
Words	15974	17099
Vocabulary	5702	5540

Table 1: *Corpus statistics for the journalistic Catalan-Spanish test set.*

A second test corpus is provided within the medicine domain. This medical corpus was kindly provided by the UniversalDoctor Project company, which focuses on facilitating communication between health-care providers and patients from various origins (*http://www.universaldoctor.com*). The medical corpus consists of 554 parallel sentences and only one manual reference for each direction of translation was available. Table 2 shows the number of sentences, words and vocabulary used for each language.

5. Automatic evaluation

Automatic evaluation is one of the most crucial issues in the development stage of a MT system, given that other types of evaluation are usually expensive. Error rate is typically measured by comparing the system output against a set of human references, according to an evaluation metric at choice. In this paper, we

	Spanish	Catalan
Sentences	554	554
Words	3127	3117
Vocabulary	920	913

Table 2: *Corpus statistics for the medical Catalan-Spanish test set.*

present the results with the well-known BLEU (Papineni et al., 2002) and TER (Snover et al., 2006).

Tables 3 and 4 show the results obtained through automatic evaluations in the journalistic and medical corpora, respectively. It can be clearly seen that, in the journalistic translations, the UPC statistical system perform better in terms of automatic evaluation than the Translendium rule-based system. These results may be explained by the fact that the UPC statistical system is trained with journalistic corpora, since these kind of corpora can be easily collected.

	Es2Ca		Ca2Es	
	BLEU	TER	BLEU	TER
Transl	85.97	11.04	87.81	7.83
UPC	**86.54**	**10.76**	**88.58**	**7.80**

Table 3: *Automatic evaluation using the journalistic corpus.*

	Es2Ca		Ca2Es	
	BLEU	TER	BLEU	TER
Transl	56.30	**32.60**	53.31	37.71
UPC	**56.66**	32.18	**55.34**	**34.26**

Table 4: *Automatic evaluation using the medical corpus.*

In the Spanish-to-Catalan medical translation results, Translendium provides the best translation in terms of TER but not in terms of BLEU. However, in the Catalan-to-Spanish medical translation, UPC offers the best system.

Although results tend to show coherence in all automatic measures (except in the Spanish-to-Catalan medical translation) —which gives more consistency to the evaluation— all measures present several deficiencies that cast serious doubts on the coherence with human criteria and on its usefulness, both for sentence-level error analysis and for system-level comparison (Callison-Burch et al., 2006). The following section presents a human evaluation.

6. Human evaluation

The comparison between different translation system outputs was performed by 10 different human evaluators. All the evaluators were bilingual in Catalan and Spanish, therefore, no reference of translation was shown to them, in order to avoid any bias in their evaluation.

Each judge was asked to make a system-to-system (pairwise) comparison. Each annotator evaluated 200 randomly extracted translation pairs, and assessed in each case whether one system produced a better translation than the other one, or whether the two outputs were equivalent. Figure 4 shows an example of the screenshot shown to the annotator. Each judge evaluated 200 hundred sentences in each direction and test set. Therefore, a total number of 4000 judgements was collected for each journalistic and medical test sets. Table 5 and 6 show average results in percentage for the journalistic corpus and for the medical corpus, respectively.

Es2Ca	UPC	Ca2Es	UPC
Transl	48%	Transl	56%

Figure 5: *Human judgements after the system-to-system comparison using the journalistic corpus. Results show in which percentage the system in the left column was marked as better than the system in the upper row.*

Es2Ca	UPC	Ca2Es	UPC
Transl	48%	Transl	69%

Figure 6: *Human judgements after the system-to-system comparison using the medical corpus. Results show in which percentage the system in the left column was marked as better than the system in the upper row.*

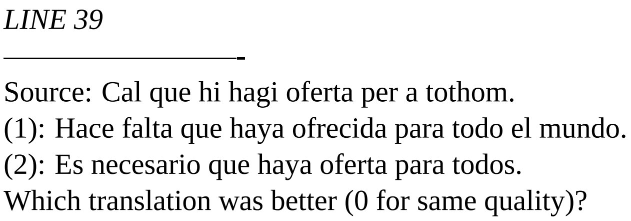
In Catalan-to-Spanish translations, Translendium shows the best performance, while in the Spanish-to-Catalan translations the best performance is provided by UPC system.

When comparing automatic measures and human judgements, results are coherent in the Spanish-to-Catalan direction. Automatic measures and human judgements correlate better when comparing systems of the same core technology, which is not the case of this work.

7. Conclusions

The aim of this work was to analyse the main differences in terms of performance between a rule-based and a statistical machine translation systems in the specific case of Catalan-Spanish pair. In this paper, two different kinds of evaluation have been carried out in order to compare them.

Human and automatic evaluation seem to correlate in one translation direction (Spanish-to-Catalan) and they do not seem to correlate in the opposite (Catalan-to-Spanish). In the automatic evaluation, the statistical system performs better than the rule-based system. In the human evaluation and in the Spanish-to-Catalan direction the statistical system also performs better than the rule-based system, while in the Catalan-to-Spanish direction is the other way round.



LINE 39

Source: Cal que hi hagi oferta per a tothom.
(1): Hace falta que haya ofrecida para todo el mundo.
(2): Es necesario que haya oferta para todos.
Which translation was better (0 for same quality)?

Figure 4: *Screenshot of the human evaluation when comparing system-to-system.*

The statistical system errors tend to be more penalized by human judgements than the errors of the rule-based system. This can be explained because statistical errors are usually unexpected and they do not follow any pattern.

When considering both evaluations together, we are able to conclude that UPC tends to perform better than Translendium in the journalistic domain. Moreover, the Translendium rule-based system tends to perform better than UPC in the medical domain. This may be explained because the statistical system is trained with journalistic corpora. Further system combination research work may take into account these results.

In the future, it will be interesting to perform a deeper manual evaluation comparing more systems in order to find out: which are the kind of errors committed by each system and how does the type of error affect the human evaluation.

8. Acknowledgements

This work has been partially funded by the Spanish Department of Education and Science through the *Juan de la Cierva* fellowship program and the BUCEADOR project (TEC2009-14094-C04-01). The authors also wants to thank the Barcelona Media Innovation Centre for its support and permission to publish this research and the UniversalDoctor Project for providing the medical test material.

9. References

S. Bangalore and G. Riccardi. 2000. Finite-state models for lexical reordering in spoken language translation. In *Proc. of the 6th Int. Conf. on Spoken Language Processing, ICSLP'02,* volume 4, pages 422–425, Beijing, October.

C. Callison-Burch, M. Osborne, and P. Koehn. 2006. Re-evaluating the role of BLEU in machine translation research. In *Proc. of the 11th Conf. of the European Chapter of the Association for Computational Linguistics,* pages 249–256, Trento, Italy, March.

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

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

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

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

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

E. Matusov, G. Leusch, R. E. Banchs, N. Bertoldi, D. Dechelotte, M. Federico, M. Kolss, Y. Lee, J. B. Marino, M. Paulik, S. Roukos, H. Schwenk, and H. Ney. 2008. System combination for machine translation of spoken and written language. *IEEE Transactions on Audio, Speech and Language Processing,* 16(7):1222–1237, September.

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

K. Papineni, S. Roukos, T. Ward, and W-J. Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics,* pages 311–318, Philadelphia, PA, July.

M. Snover, B. E. Dorr, R. Schwartz, L. Micciulla, and J. Makhou. 2006. A study of translation edit rate with targeted human annotation. In *Proc. of the 5th Conf. of the Association for Machine Translation in the Americas, AMTA'06,* Boston,USA.

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