SisHiTra: A Spanish-to-Catalan hybrid machine translation system

J. González*, A. L. Lagarda*, J. R. Navarro†, L. Eliodoro†, A. Giménez*, F. Casacuberta†, J. M. de Val‡, F. Fabregat‡

*Departament de Sistemes Informàtics i Computació
Universitat Politècnica de València
{jgonzalez, alagarda, agimenez}@dsic.upv.es

† Institut Tecnològic d'Informàtica
Universitat Politècnica de València
{jonacer, leliodoro, fcn}@iti.upv.es

‡ Servei de Normalització Lingüística
Universitat de València
{Joan.M.Val, Ferran.Fabregat}@uv.es

Abstract

In this paper, we describe how deductive and inductive techniques can be successfully combined in the framework of SisHiTra, a machine translation system from Spanish to Catalan with no semantic constraints. The translation process is based on finite-state machines and statistical models. Linguistic knowledge is appropriately incorporated as a database. Our results are compared with other systems.

1. Introduction

Machine translation (*MT*) is a challenging topic that engineers and scientists have been interested in for years. In addition to its importance for the study of human speech characteristics, MT is of social and economic interest because its development would allow for the preservation of the use of minority languages, such as Catalan, Basque or Galician in Spain. It could thus mean a reduction of linguistic barriers. This is particulary important in the access to some computer services. Another feature of Catalan is that it is a language that is spoken by more people than some official European languages.

Spanish and Catalan are languages that belong to the Romance language family, although they are from different linguistic branches: Catalan is a Gallo-Romance language, whereas Spanish is an Ibero-Romance one. Nonetheless, their resemblance is quite notable, since both of them are inflectively and morphosyntactically similar languages.

The approaches that have been traditionally used for MT can be classified into two families: *knowledge-based* and *corpus-based* methods. Knowledge-based techniques formalize expert linguistic knowledge in the form of rules, dictionaries, etc., in a computable way. Corpus-based methods use statistical pattern recognition techniques to automatically infer models from text samples without necessarily using a-priori linguistic knowledge.

Knowledge-based techniques are classical approaches for dealing with general scope MT systems. Nevertheless, inductive methods have achieved competitive results with semantically constrained tasks. On the other hand, finite-state transducers (Karttunen, 1993; Roche, 1999; Roche and Schabes, 1997) have been successfully used to implement both rule-based and corpus-based MT systems. Techniques based on finite-state models have also allowed for the development of useful tools for natural language processing (Mehryar, 1997; Mohri et al., 2000; Oflazer, 1996; Roche and Schabes, 1995; Casacuberta et al., 2004), which are interesting because of their simplicity and their adequate temporal complexity. SisHiTra makes use of finite-state models to combine knowledge-based and corpus-based techniques so as to produce a Spanish-to-Catalan MT system with no semantic constraints. Some other finite-state approaches to Spanish↔Catalan translation, such as *interNOSTRUM* (Forcada et al., 2001), confirm their adequateness to MT between these two languages.

SisHiTra's main aim is the achievement of high quality translations from Spanish to Catalan (and vice versa) for dissemination purposes. Of course, this is an ideal objective for any MT system; however, in our case, it is an especially important issue that has been taken into account in the design of each stage. Thus, we consider that *perfect* translations would be those that did not seem to be the result of a translation process, but that seemed as if they had been produced directly in the target language. This is not a problem for a human translator, but it is crucial for MT systems. For instance, semantic ambiguity is easily solved by a human speaker or reader, but it is usually a significant problem in MT. As a consequence of that, the evaluation of SisHiTra's performance is in terms of how far hypotheses are from a set of translation references, which experts have considered to be linguistically optimal.

The SisHiTra prototype is designed to be a serial process where every module performs a specific task. There is an online version running on the Internet[1] that is able to translate plain text, web pages, and LaTEX files.

Future versions of SisHiTra would be extended to other language pairs (Portuguese, French, Italian, etc.). In the fol-

This work was partially supported by the "Agència Valenciana de Ciència y Tecnologia" under grant GRUPOS03/031 and by the Spanish project ITEFTE (TIC2003-08681-C02-02).

[1] http://prhltdemos.iti.upv.es/~taval/

lowing section, we will explain SisHiTra's architecture.

2. System architecture

SisHiTra is a general scope Spanish-to-Catalan translator with a wide vocabulary recall, so it is able to deal with all kinds of sentences. A previous version of the SisHiTra system can be found in (Navarro et al., 2004).

The methodologies to be used in the representation of the different knowledge sources are based on finite-state machines: on the one hand, stochastic transducers, which are employed as data structures for dictionary requests as well as for inter-module communication; on the other hand, Hidden Markov Models (HMM), which are applied in disambiguation processes (Sanchis et al., 2001). Finite states have proven to be adequate models for translation purposes. They can be easily inferred from corpora, and there are efficient algorithms for their manipulation (Viterbi, beam search, etc.). In addition, linguistic knowledge can be properly incorporated.

As previously stated, translation prototype modules are based on finite-state machines, providing a homogeneous and efficient framework. Engine modules process input text by means of a cascade of finite-state models that represent both linguistic and statistical knowledge. Finite-state models are also used to represent partial information during translation stages.

The SisHiTra system is structured in the following modules:

- **The preprocess module:** It divides the original text into sentences, thus allowing the translation process to be applied to each individual sentence. Let us introduce a simple example in order to better understand how SisHiTra performs. Figure 1 shows some Spanish text to be translated.

La estudiante atendió.

Figure 1: Translation text

Moreover, sentences are split up as a sequence of translation units, where every translation unit is then identified and classified into one of the following groups: punctuation marks, numbers, abbreviations, proper names, or general words. Output is expressed in a *xml* format, in which every paragraph, sentence, translation unit, and case information, has been detected (see Figure 2).

```
<doc>
 <p>
  <o>
   <ut ort="M">la</ut> <ut>estudiante</ut>
   <ut>atendió</ut> <ut uti="signo">.</ut>
  </o>
 </p>
</doc>
```

Figure 2: Preprocess

As it can be deduced from Figure 1, the translation example is composed of only one sentence. Figure 2 shows how preprocess has segmented the whole text, thus identifying the most significant components. Xml tags stand for as follows:

- <[/]doc> labels refer to the whole document.
- <[/]p> labels point paragraphs out.
- <[/]o> labels show sentence beginning/ending.
- <[/]ut> labels identify translation units.

Note that punctuation marks must be isolated from words in order to properly detect the right translation units. In the example, full stop is separated from last word *atendió*.

In addition, upper case characters are identified, then lowered so as to be able to perform case-independent dictionary requests. Once the translation process has been carried out, case information can be restored to their original format. Figure 2 shows how uppercasing (in the example, initial word *La*) is handled by means of a translation unit feature, ort. Possible values for ort are 'M' (initial character), 'T' (all the characters) or 'U', which in conjunction with another feature, $mask$, take into account some particular configuration, just as it happens with *SisHiTra*.

- **The generation module:** A dictionary request produces a syntactic graph that represents all the possible analyses over the input sentence together with all their possible translations. For the proposed example, Figure 3 shows the result of this stage.

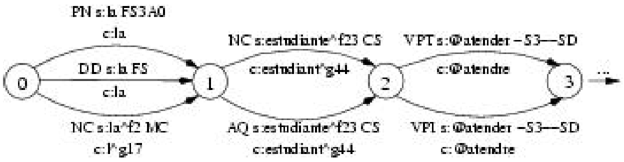

Figure 3: Dictionary access

Every edge represents a dictionary answer, which is directly related to, at least, one translation unit (note that set phrases would be modelled as a transition from state x to y, where $y - x > 1$). Therefore, transitions that share the same source/target states refer to the different readings that an input segment has, according to the dictionary. Information on edges is composed of a lexical category, together with a one-to-many relationship of Spanish and Catalan database entries. Because of space limitations, only one translation per edge appears, although it must be observed as if there could be several ones.

- **The disambiguation module:** Syntactic and semantic disambiguation is performed using statistical models. First, morphological and syntactic disambiguation is carried out through a tagging process, finding the most likely path over the analysis graph. This implies a segmentation of the the input sentence into tokens, selecting one lexical category per token. In a second step, semantic disambiguation is also performed by means of context-dependent methods. That means choosing one of the provided translations for every surviving edge from the previous morpho-syntactic disambiguation stage. This is accomplished through the very same statistical approach, namely the one that is based on HMMs, using a Catalan language model and a stochastic dictionary. Figure 4 shows the result of the disambiguation module over the proposed example.

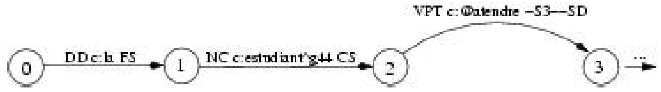

Figure 4: Statistical disambiguation

- **The postprocess module:** Here, several rule-based conversions are applied in order to transform output (that is not yet in natural language) into correct sentences from the target language. This includes noun phrase agreement, inflection, spelling, and format. Figure 5 shows the final output.

L'estudiant va atendre.

Figure 5: Translated text

MT needs to somehow semantically disambiguate source words before turning them into target language items. Semantic disambiguation methods try to determine the implicit meaning of a word in a surrounding context. SisHiTra makes use of statistical models for doing such a task.

Statistical models are becoming popular for several reasons. The most important reason is that they are cheaper and faster to generate than knowledge-based systems. Statistical techniques learn automatically from corpora, without the process of producing linguistic knowledge. Of course, obtaining corpora for model training is not a task that is free of effort.

Models for semantic disambiguation in SisHiTra need parallel corpora, that is, corpora where text segments (such as sentences or paragraphs) in a language are matched with their corresponding translations in the other language. These corpora have been obtained from different bilingual electronic publications (newspapers, official texts, etc.) and they have been paralleled through different alignment algorithms.

SisHiTra's modules were implemented following two different schemes: pre- and postprocess modules were coded in *flex* language, which is a very useful tool for generating programs that perform pattern-matching of regular expressions in text; and, the remaining modules were all written in *C*, following a Viterbi(Viterbi, 1967) searching strategy.

In the next section, we will explain in detail SisHiTra's cornerstone: its dictionary.

3. Linguistic knowledge as a database

One of the advantages of our system is its convenient incorporation of linguistic knowledge, which seems to be essential to achieve *natural* translations. This knowledge is represented as a database where the main table registers refer to common dictionary entries, that is, words or set phrases from the source language (generally known as *tokens*).

Each possible translation is considered for every source token by means of a one-to-many relationship. Thus, in order to decide which translation is the most suitable, some extra information must be taken into account, such as surrounding context, the dialectal variety that is chosen, or the target language structure. All this information is reflected in our dictionary fields. These fields are:

- Spanish literal: the token's citation form. That is, masculine, if it has only gender inflection; singular masculine, if it also has number inflection; infinitive, if it is a verb, etc.

- Label of the Spanish token. It indicates the token's root and the way it is inflected.

- Morphotactics. Grammatical category of the Spanish token (noun, adjective, verb, adverb, etc.) together with some syntactic and lexical information (if a verb is transitive, intransitive, auxiliar, etc.; or if a conjunction is correlative, subordinating, etc.; and so on).

- Remission (only applied to verbs). Unlike nouns or adjectives, all the verb forms from the same infinitive do not necessarily have to share the same root. Therefore, we need to explicitely introduce every different pair (root, paradigm) as a separate entry. However, translation information is not cloned for every register but only stored in the one corresponding to the infinitive. This field links all these different entries towards their common translation information.

- Morphological unit class. Tokens are classified into one of the following 4 classes: common, infinitive verbs, prefixes, and suffixes. The usefulness of the first two groups has already been explained. Prefixes and suffixes are special tokens that are needed in order to be able to parse (and translate) unknown words, which are truly composed of:

  a) a well known prefix plus a dictionary entry

  or

  b) a dictionary entry plus a well known suffix

  Therefore, we can identify and successfully translate any compound or derivate word not explicitely included in the database.

- Abbreviation. It shows if tokens are abbreviations or acronyms.

- Case. It specifies if tokens are always written in upper case.

- Nominal inflection. Description of all the Spanish and Catalan inflectional paradigms.

- Extra. Additional information (literal or figurative sense, usual context, etc.).

- Number of meanings of the Spanish token.
- Senses. A set of different meanings for the current Spanish token. Each of them has the following information:
  - Thematic marks. Different topics where tokens might appear (technology, biology, sports, shows, chemistry, business, etc.). This field helps to do sense disambiguation in order to choose the right meaning according to the terminological sphere which tokens are referring to.
  - Semantic marks. Knowledge-based logical and semantic information.
  - Sense order. Priority over the set of meanings of the Spanish token. In addition, each sense has a set of Catalan equivalences, that is, there is a translation (or a set of synonymous ones) for each meaning of a Spanish token. Equivalences that belong to the first sense are preferred to the ones from the second sense; in turn, these are preferred to the ones from the third sense, and so on. These values have been manually established according to some linguistic criteria, taking into account both frequency of use and linguistic expressiveness.
  - Equivalences. A set of synonyms for a given sense of the Spanish token. Each equivalence has the following information:
    - ∗ Priority of a particular translation over a set of synonyms. As previously explained, a source token can have several meanings, and once one is chosen, there are several equivalent translations. We also set a range of priorities over these synonyms.
    - ∗ Catalan literal. See Spanish literal.
    - ∗ Catalan label. See Spanish label.
    - ∗ Catalan grammatical category. In general, given the similarity between both languages, Catalan tokens inherit their corresponding grammatical category of the Spanish token.
    - ∗ Preference. It refers to the linguistic dialect that is more likely to produce the Catalan item. Two of these Catalan dialects are taken into account: eastern and western. This distinction allows us to produce adequate expressions according to the user's linguistic area. In future reverse versions (from Catalan to Spanish), it will be essential to consider the whole Catalan vocabulary.

Although only a very superficial description of our dictionary has been presented here, it provides a general framework for building new dictionaries for other language pairs.

4. Evaluation

Several corpora were collected to assess the translation quality achieved. A comparison between SisHiTra and some other Spanish-to-Catalan systems has been made.

4.1. Corpora

In order to be able to make a statistical estimation of the different models used in the implemented version of the prototype, several corpora were collected.

Specific tools were developed to look for information through the web. The *LexEsp* corpus (Carmona et al., 1998), with nearly 90.000 running words, was used to estimate *syntactic disambiguation* model parameters. A label, from a set of approximately 70 categories, was manually assigned to each word.

Two other corpora (*El Periódico de Catalunya* and *Diari oficial de la Generalitat Valenciana*) were obtained by means of web tools. These corpora will be used in some system improvements such as training models for *semantic disambiguation*. These corpora consist of parallel texts that are aligned at the sentence level in a Spanish-to-Catalan translation framework without semantic constraints.

An evaluation corpus was created to perform the system assessment. This corpus is composed of 240 sentence pairs (4389 running words), which were extracted from different sources and published in both languages. Needless to say, they are not included in any training corpus.

- 120 sentence pairs from *El Periódico de Catalunya*, with no semantic constraints.
- 50 pairs from *Diari Oficial de la Generalitat Valenciana*, an official publication from the Valencian Community government.
- 50 pairs from technical software manuals.
- 20 pairs from websites (Valencia Polytechnical University, Valencia city council, etc.).

4.2. Results

Word error rate (WER[2]) is a translation quality measure that computes the edition distance between translation hypotheses and a predefined reference translation. The edition distance calculates the number of substitutions, insertions, and deletions that are needed to turned a translation hypothesis into the reference translation. The accumulated number of errors for all the test sentences is then divided by the number of running words, and the resulting percentage shows the average number of incorrect words. Since it can be automatically computed, it has become a very popular measure. The WER results for the SisHiTra system are similar to the ones achieved by other non-commercial systems (*interNOSTRUM*[3] and *SALT*[4]) as shown in Table 1.

interNOSTRUM is a realtime MT system that provides approximate translations from Spanish to Catalan. Texts can be processed in any of the following formats: ANSI, HTML, and RTF. *SALT* is a completely knowledge-based MT system that performs an interactive method that minimizes mistakes, thus providing naturalness to translations.

[2] Also known as Translation WER (TWER)

[3] See http://www.internostrum.com

[4] See http://www.cult.gva.es/salt/salt_programes_salt2.htm

Table 1: WER comparison between some MT systems

System	WER
interNOSTRUM	12.6
SisHiTra	12.5
SALT 3.0	12.2

A disadvantage of WER is that it only compares the translation hypothesis with a fixed reference translation. This does not offer any margin to possibly correct translations that are expressed in a different writing style. Therefore, to avoid this problem, we used the WER with multireferences (MWER[5]) to evaluate the prototype. MWER considers several reference translations for the same test sentence, then computes the edition distance with all of them, returning the minimum value as the error corresponding to that sentence. MWER offers a more realistic measure than WER because it allows for more variability in translation style. Other two more references were created by expert linguists, making variations to the original reference sentence. The MWER results for the SisHiTra system are the best in the three tested systems, as shown in Table 2.

Table 2: MWER comparison between some MT systems

System	MWER
interNOSTRUM	6.5
SisHiTra	**4.1**
SALT 3.0	6.1

With regard to the translation speed, SisHiTra is able to process more than 1000 words per second, which can be considered as realtime working.

5. Conclusions and future work

SisHiTra shows how deductive and inductive techniques can be successfully combined to produce a MT system with no semantic constraints from Spanish to Catalan, a *nearly* official European minority language that is spoken by an important number of the European people. The translation process is based on finite-state machines and statistical models that are automatically inferred from parallel corpora. The translation results are promising, but there are still several points that must be improved.

In addition, an appropriate representation of linguistic knowledge has been incorporated into a MT system as a database, which is essential for obtaining *natural* translations. Moreover, this database structure could be easily adapted to other language pairs.

The most relevant areas where the system could be improved are:

- Semantic disambiguation, where statistical models for ambiguous words could be trained in order to be able to choose the most appropriate context-dependent translations.
- Verb phrase agreement.

We also bear in mind a SisHiTra reversion in order to translate from Catalan to Spanish. A preliminary version of the needed linguistic dictionary can be automatically obtained from our current database.

Finally, SisHiTra's framework could be extended to other Romance languages (Portuguese, French, Italian, etc.).

6. References

J. Carmona, S. Cervell, L. Màrquez, M.A. Martí, L. Padró, R. Placer, H. Rodríguez, M. Taulé, and J. Turmo. 1998. An Environment for Morphosyntactic Processing of Unrestricted Spanish Text. In *Proceedings of the 1st International Conference on Language Resources and Evaluation, LREC*, pages 915–922, Granada, Spain, May.

F. Casacuberta, H. Ney, F. J. Och, E. Vidal, J. M. Vilar, S. Barrachina, I. Garcia-Varea, C. Martinez D. Llorens, S. Molau, F. Nevado, M. Pastor, D. Pico, and A. Sanchis. 2004. Some approaches to statistical and finite-state speech-to-speech translation. *Computer Speech and Language*, 18:25–47.

M. Forcada, A. Garrido, R. Canals, A. Iturraspe, S. Montserrat-Buendia, A. Esteve, S. Ortiz Rojas, H. Pastor, and P.M. Pérez. 2001. The spanish-catalan machine translation system internostrum. *0922-6567 - Machine Translation*, VIII:73–76.

L. Karttunen. 1993. Citation of unpublished documents. Technical report, XEROX Palo Alto Research Center.

M. Mehryar. 1997. Finite-state transducers in language and speech processing.

Mehryar Mohri, Fernando Pereira, and Michael Riley. 2000. The design principles of a weighted finite-state transducer library. *Theoretical Computer Science*, 231(1):17–32.

José R. Navarro, Jorge González, David Picó, Francisco Casacuberta, Joan M. de Val, Ferran Fabregat, Ferran Pla, and Jesús Tomás. 2004. SisHiTra : A Hybrid Machine Translation System from Spanish to Catalan. In *EsTAL*, pages 349–359.

Kemal Oflazer. 1996. Error-tolerant finite-state recognition with applications to morphological analysis and spelling correction. *Computational Linguistics*, 22(1):73–89.

Emmanuel Roche and Yves Schabes. 1995. Deterministic part-of-speech tagging with finite-state transducers. *Computational Linguistics*, 21(2):227–253.

Emmanuel Roche and Yves Schabes, editors. 1997. *Finite-State Language Processing*. Bradford Book. MIT Press, Cambridge, Massachusetts, USA.

Emmanuel Roche. 1999. Finite state transducers: parsing free and frozen sentences. pages 108–120.

A. Sanchis, D. Picó, J.M. del Val, F. Fabregat, J. Tomás, F. Casacuberta, and E. Vidal. 2001. A morphological analyser for machine translation based on finite-state transducers. In *MT Summit VIII*, pages 305–309, September.

A. Viterbi. 1967. Error bounds for convolutional codes and an asymtotically optimal decoding algorithm. *Annals of the New York Academy of Sciences*, IT-13:260–269.

[5] Multi-reference Word Error Rate