SisHiTra* : A hybrid machine translation system from Spanish to Catalan

José R. Navarro[†], Jorge González[†], David Picó[†], Francisco Casacuberta[†], Joan M. de Val[‡], Ferran Fabregat[‡], Ferran Pla[+], and Jesús Tomás[*]

[†]Instituto Tecnológico de Informática
Universidad Politécnica de Valencia
{jonacer, jgonza, dpico, fcn}@iti.upv.es

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

[+]Departamento de Sistemas Informáticos y Computación
Universidad Politécnica de Valencia
fpla@dsic.upv.es

[*]Departamento de Comunicaciones
Universidad Politécnica de Valencia
jtomas@dcom.upv.es

Abstract. In the current European scenario, characterized by the coexistence of communities writing and speaking a great variety of languages, machine translation has become a technology of capital importance. In areas of Spain and of other countries, coofficiality of several languages implies producing several versions of public information. Machine translation between all the languages of the Iberian Peninsula and from them into English will allow for a better integration of Iberian linguistic communities among them and inside Europe. The purpose of this paper is to show a machine translation system from Spanish to Catalan that deals with text input. In our approach, both deductive (linguistic) and inductive (corpus-based) methodologies are combined in an homogeneous and efficient framework: finite-state transducers. Some preliminary results show the interest of the proposed architecture.

1 INTRODUCTION

Machine translation and natural computer interaction are questions that engineers and scientists have been interested in for decades. In addition to their importance for the study of human speech characteristics, these applications have social and economic interests because their development would allow for a reduction of the linguistic barriers that prevent us to make with confidence activities as, for example, travelling to other countries or the access to some computer science services (foreign websites and so on).

* Work partially supported by the Spanish CICYT under grants TIC 2000-1599-C02-01 and TIC 2003-08681-C02-02

Machine translation has received an increasing attention in the last decades due to its commercial interest and to the availability of large linguistic and computational resources. These resources are allowing machine translation systems to leave the academic scope to become more useful tools for professionals and general users.

Nevertheless, natural language complexity creates too many difficulties to develop high quality systems. This opens multiple investigation lines in which researchers hardly work to improve translation results. The three most important machine translation problems are:

- PoS [1] tagging, whose objective is to identify the lexical category that a word has in a sentence [1,2,3].
- Semantic disambiguation, that decides which is the right sense of a word in a text [4,5].
- Reordering, which can appear quite often when translating between different family languages.

The approaches that have been traditionally used to face these problems can be classified into two big families: *knowledge-based* and *corpus-based* methods. Knowledge-based techniques formalize expert linguistic knowledge, in 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.

SisHiTra (Hybrid Translation System) project tries to combine knowledge-based and corpus-based techniques to produce a Spanish-to-Catalan machine translation system with no semantic constraints. Spanish and Catalan are languages belonging to the Romance language family and have a lot of characteristics in common. SisHiTra makes use of their similarities to simplify the translation process. A SisHiTra future perspective is the extension to other language pairs (Portuguese, French, Italian, etc.).

Knowledge-based techniques are classical approaches to tackle general scope machine translation systems. Nevertheless, inductive methods have shown competitive results dealing with semantically constrained tasks.

Moreover, finite-state transducers [6,7,8] have been successfully used to implement both rule-based and corpus-based machine translation systems. Techniques based on finite-state models have also allowed for the development of useful tools for natural language processing [9,10,11,3] that are interesting because of their simplicity and their adequate temporal complexity.

With the experience acquired in InterNOSTRUM [12] and TAVAL [13], *SisHiTra* project was proposed. SisHiTra system is able to deal with both eastern and western Catalan dialectal varieties, because the dictionary, which is its main database, establishes explicitely such distinction.

SisHiTra prototype has been thought to be a serial process where every module performs a specific task. In the next section we will explain the different parts in which SisHiTra system is divided.

[1] Parts of Speech.

2 IMPLEMENTATION

The methodologies that are going to be used to represent the different knowledge sources (dictionary, module interfaces, etc.) are based on finite-state machines: Hidden Markov Models (HMM) are applied in disambiguation modules [13], and stochastic transducers are used as data structures for dictionary requests as well as for inter-module communication. Reasons for using finite-state methodology are as following:

- Finite-state machines are easily represented in a computer, which facilitates their exploitation, visualization and transference.
- There are algorithms that allow for their manipulation in an efficient way (Viterbi, beam search, etc.).
- There are algorithms for their automatic inference (both their topology and their associated probability distributions) from examples.
- Linguistic knowledge incorporation can be adequately carried out.
- It allows for both serial or integrated use of the different knowledge sources.
- More powerful models can be used, such as context-free grammars, by means of a finite-state approach.

2.1 System architecture

The system developed in SisHiTra project translates from Spanish to Catalan. It is a general scope translator with a wide vocabulary coverture, so it is able to deal with all kind of sentences.

As previously commented, translation prototype modules are based on finite-state machines. This provides an homogeneous and efficient framework. Engine modules process input text in Spanish by means of a cascade of finite-state models that represent both linguistic and statistical knowledge. For example, two finite-state machines are needed to do PoS tagging of input sentences: first of them represents a knowledge-based dictionary and the second one defines a corpus-based disambiguation model. Finite-state models are also used to represent partial information during translation stages (e.g. lexically ambiguous sentences).

SisHiTra and lots of other systems need, somehow, to semantically disambiguate words before turning them into target language items. Semantic disambiguation methods try to find out the implicit meaning of a word in a surrounding context.

SisHiTra is designed to make semantic disambiguation in two steps: first, a rule-based module solves some ambiguities according to certain well-known linguistic information and, afterwards, a second module ends the job by means of corpus-based inductive methods. Statistical models are receiving more interest every day for several reasons. The most important one 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 free of effort. Models for semantic disambiguation in SisHiTra need parallel corpora, that is, corpora where text segments (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 system is structured in the following modules:

- **Fragmenter module:** It divides the original text into sentences.
- **Labeler module:** A dictionary request produces a syntactic graph that represents all the possible analysis over the input sentence.
- **Syntactic disambiguation module:** By means of statistical models, it finds the most likely syntactic analysis between all those that labeler module produces.
- **Nominal phrase agreement module:** Every phrase element must agree in gender and number with each other.
- **Localizer module:** Another dictionary request produces a lemma-graph that represents all the possible translations for the previously analyzed and disambiguated sentence.
- **Semantic disambiguation module:** Here, a prototype in which disambiguation is carried out according only to the dictionary is presented, but we are testing some beta-systems that consider statistical models to make this decision, yet for us one researching open line.
- **Inflection module:** Lemmas are turned into their corresponding inflected words from the morphological information previously analyzed.
- **Formatting module:** Contraction and apostrophization are applied in order to respect the Catalan ortographic rules.
- **Integration module:** Compilation of translations, according to the original text format, is finally done.

In the following section, examples are used in order to show the functionality of each module in a more concrete way.

3 MODULES

3.1 Fragmenter module

Input text must be segmented into sentences so that translation can be carried out. By means of rules, this module is able to do it.

Input: Input to this module is Spanish text to be translated.

La estudiante atendió.

Output: Output from this module expresses the whole text in a *xml* format, in which upper characters have been lowered and where every paragraph, sentence and translation unit (*ut*) has been detected.

<doc> <o> <ut ort=”M”>la</ut> <ut>estudiante</ut> <ut>atendió</ut> </o> </doc>

3.2 Labeler module

This module outputs a graph that represents all the syntactic analysis possibilities for the input sentence. The applied method consists of a full search of translation units (words or compound expressions) through a finite-state network representing the dictionary.

Input: Input to this module are fragmented sentences.

<doc> <o> <ut ort="M">la</ut> <ut>estudiante</ut> <ut>atendió</ut> </o> </doc>

Output: Output from this module is a finite-state transducer in which each edge associates translation units and lexical categories[2] according to the dictionary. Note that each translation unit, represented as a conection between states, can be referred to both a word or a compound expression, since TAVAL dictionary stores lexical labels for single words as well as for phrases.

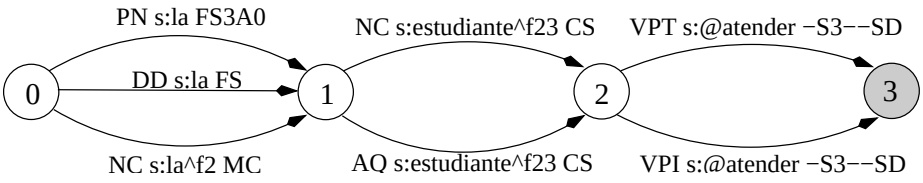

Fig. 1. Labeler's output

Fig. 1 shows all the possible PoS-tags for the example sentence, together with some linguistic information. In concrete, word *la* can be a pronoun, an article or a noun. Word *estudiante* can be an adjective or a noun; it is a singular word and its gender depends on some issues that are implemented in Nominal phrase agreement module. PoS-tags for word *atendió* are VPT and VPI, both corresponding to a third person singular from simple past.

3.3 Syntactic disambiguation module

Syntactic disambiguation aims to decide the lexical category that a word has in a context. To do this, both rule-based and corpus-based techniques are applied.

Statistical disambiguation can be defined as a maximization problem. Let $\mathcal{W} = \{w_1, w_2, \ldots, w_N\}$ be the source language vocabulary and let $\mathcal{C} = \{c_1, c_2, \ldots, c_m\}$ be all the possible categories. Given an input sentence $w = w_1, \ldots, w_L$, the process can be accomplished by searching the category sequence $\widetilde{c} = c_1, \ldots, c_L$ that maximizes:

[2] *PN*:Pronoun, *DD*:Article, *NC*:Noun, *AQ*:Adjective, *VPT*:Transitive verb, *VPI*:Intransitive verb, etc.

$$\widetilde{c} = \arg\max_{c \in \mathcal{C}^L} P(c|w) \tag{1}$$

Using Bayes rule and given that the maximization process is independent on the input sentence w, equation (1) can be rewritten as:

$$\widetilde{c} = \arg\max_{c \in \mathcal{C}^L} P(c)P(w|c) \tag{2}$$

In this equation, contextual (or language model) probabilities, $P(c)$, represent all the possible category sequences, whereas emission (or lexical model) probabilities, $P(w|c)$, establish the relationship between words and categories.

To solve this equation, certain Markov assumptions can be accepted to simplify the problem. First, contextual probabilities for one determined category are assumed to only depend on the immediately previous *n* categories. The second constraint consists of assuming that emission probabilities only depend on the category itself.

For 1st order Markov models (bigrams), problem is reduced to solve next equation:

$$\widetilde{c} = \arg\max_{c_1,\ldots,c_L} \left(\prod_{1\ldots L} P(c_i|c_{i-1})P(w_i|c_i) \right) \tag{3}$$

Parameters from this equation can be represented as a Hidden Markov Model (HMM) in which states and categories are one-to-one associated. Contextual probabilities, $P(c_i|c_{i-1})$, are transition probabilities between states, and lexical model probabilities, $P(w_i|c_i)$, can be seen as word-category probability distributions. Viterbi algorithm [14] has been used to find, for a given input sentence, its associated category sequence.

Input: Input to this module is labeler's output, which is represented in Fig. 1, and models all the possible syntactic analysis for the input sentence.

Output: Output from this module is the linear graph corresponding to the most likely path through the input graph, according to the category-based models described before.

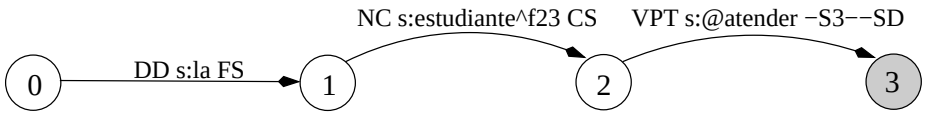

Fig. 2. Syntactic disambiguation output

Actually, some rules are used so as to reduce ambiguity, then the statistical disambiguation model presented here is applied.

3.4 Nominal phrase agreement module

Due to the fact that training corpus for syntactic disambiguation does not include information about word gender or number, it is necessary to perform a subsequent process making agree all the words in each nominal phrase.

The followed method consists of nominal phrase localization inside a sentence by means of a knowledge-based nominal phrase codification in terms of category sequences [15].

Once a nominal phrase has been located, it is possible to make agree gender and number words inside it through the application of some hierarchical rules that depend on the kind of phrase detected.

Input: Input to this module is Syntactic disambiguation module's output. As Fig. 2 shows, it consists of a linear graph containing PoS-tag labelling.

Output: Output from this module offers sentences in which gender and number agreement has been made at a nominal phrase level. In Fig. 3, it is possible to see a nominal phrase detection and, as a result, noun's gender has been agreed with article's.

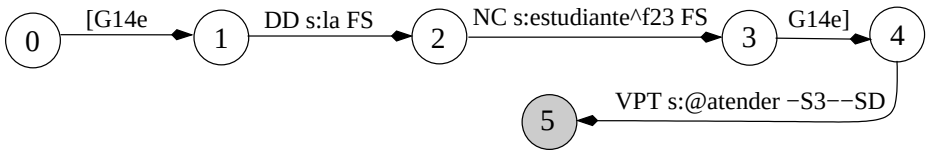

Fig. 3. Nominal phrase agreement output

3.5 Localizer module

This module is dedicated to expand each *ut* into all its possible translations according to the dictionary.

Input: Input to this module is agreement module's output, where nominal phrase marks have been deleted.

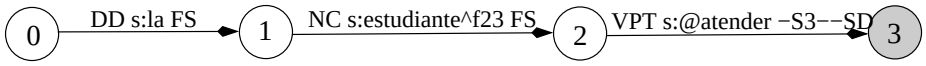

Fig. 4. Nominal phrase agreement output without phrase marks

Output: Output from this module is a lemma graph including every possible translation to the input graph, according to the dictionary.

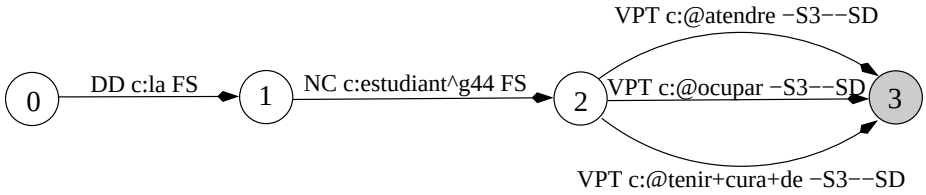

Fig. 5. Localizer's output

3.6 Semantic disambiguation module

Semantic disambiguation module tries to decide the right translation for a *ut* according to the input sentence context. In this paper, only the most likely translation for each dialectal variety is taken into account. Dictionary entries have their meanings manually scored. Therefore, for each ut, prototype chooses the best scored sense in a user-given dialectal variety. Corpus-based statistical models are planned to be working on future versions of this module.

Input: Input to this module is localizer's output. As Fig. 5 shows, every possible translation to each ut from Fig. 4 is represented.

Output: Output from this module is a linear graph which corresponds to the best scored path through the input graph.

Fig. 6. Semantic disambiguation output

3.7 Inflection module

This is a rule-based module which makes word inflection according to the Catalan inflection model.

Input: Input to this module is Semantic disambiguation module's output, which is shown in Fig. 6. It represents a Catalan lemma sentence to be inflected.

Output: Output from this module is input's inflection, that is, a sentence in which lemmas have been turned into words according to some inflection rules.

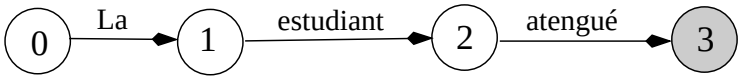

Fig. 7. Inflecter's output

3.8 Formatting module

This module is also a rule-based module and it makes some apostrophization & contraction procedures according to the Catalan grammar.

Input: Input to this module is inflection module's output, which can be seen in Fig. 7.

Output: This module finally offers well-written sentences from an ortographic point of view. In Fig. 8, it is possible to see the transformation of *La estudiant* into *L'estudiant* as well as an alternative way of expressing past tenses, which tends to be more usual.

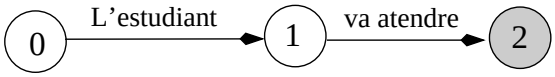

Fig. 8. Formatter's output

3.9 Integration module

This module turns finite-state linear graphs into sentences, according to the original text format.

Input: Module's input is formatter's output, which Fig. 8 shows.

Output: Output from this module (or final output) is displayed as a Catalan text, which is a translation for the Spanish input text.

L'estudiant va atendre.

4 EXPERIMENTS

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, diverse data corpora have been collected.

Specific tools were developed to look for information through the web. These tools were very useful, especially at the time of collecting the necessary corpora.

LexEsp corpus [16], 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.

Other two corpora (*El periódico de Cataluña* 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, aligned at sentence level, in a Spanish-to-Catalan translation framework without semantic constraints.

In order to perform the system assessment, a bilingual corpus was created. This corpus is composed of 240 sentence pairs, extracted from different sources and published in both languages. Of course, they are not included in any training corpus.

- 120 sentence pairs from *El Periódico de Cataluña*, with no semantic constraints.
- 50 pairs from *Diari Oficial de la Generalitat Valenciana*, a 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[3]) 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. WER results for SisHiTra system can be seen at Table 1.

A disadvantage of WER is that it only compares the translation hypothesis with a fixed reference translation. This does not offer any margin to possible right translations, expressed in a different writing style. So, to avoid this problem, we used the WER with multireferences (MWER[4]) for evaluating the prototype. MWER considers several reference translations for a same test sentence, then computes the edition distance with

[3] Also known as Translation WER (TWER)

[4] Multi-reference Word Error Rate

Table 1. WER comparison for some machine translation systems

System	WER
InterNOSTRUM	11.9
SisHiTra	10.1
SALT	9.9

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. MWER results for SisHiTra system are similar to the reached ones by other commercial systems (InterNOSTRUM[5] and SALT[6]), as it can be seen at Table 2.

Table 2. MWER comparison for some machine translation systems

System	MWER
InterNOSTRUM	8.4
SisHiTra	6.8
SALT	6.5

5 CONCLUSIONS AND FUTURE WORK

In the framework of SisHiTra project, a general scope Spanish-to-Catalan translation prototype has been developed. The translation process is based on finite-state machines and statistical models, automatically inferred from parallel corpora. Translation results are promising enough, considering that there are still a lot of things to be done.

We hope to improve results through the correction of some mistakes, accidentally produced at some of the hand-made knowledge sources (dictionary, grammatical rules, etc.), as well as to prosper in the prototype modular development, including new processes to increase translation quality.

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.
- Gender and number agreement between verbal phrases.

[5] See http://www.torsimany.ua.es

[6] See http://www.cult.gva.es/DGOIEPL/SALT/salt_programes_salt2.htm

– Disambiguation in some verb pairs like: *ser* and *ir*, *creer* and *crear*, etc. since they have lexical forms in common.

Finally, a SisHiTra future perspective is the extension to other Romance languages (Portuguese, French, Italian, etc.).

References

1. Halteren, H.V., Zavrel, J., Daelemans, W.: Improving accuracy in word class tagging through the combination of machine learning systems. Computational Linguistics **27** (2001) 199–229
2. Pla, F., Molina, A.: Improving part-of-speech tagging using lexicalized hmms. Natural Language Engineering **10** (2004) 167–189
3. Roche, E., Schabes, Y.: Deterministic part-of-speech tagging with finite-state transducers. Computational Linguistics **21** (1995) 227–253
4. Aguirre, E., Rigau, G.: Word sense disambiguation using conceptual density. In: COLING. (1996)
5. Ide, N., Véronis, J.: Word sense disambiguation: The state of the art. Computational Linguistics **24** (1998) 1–40
6. Karttunen, L.: Finite-state lexicon compiler. Volume 1. (1994) 406–411
7. Roche, E.: Finite-state transducers: parsing fre and frozen sentences. In Kornai, A., ed.: Proceedings of the ECAI 96 Workshop on Extended Finite State Models of Language. (1996) 52–57
8. Roche, E., Schabes, Y.: Finite-state language processing. MIT Press, Cambridge, Mass. (1997) 1–65
9. Mohri, M.: Finite-state transducers in language and speech processing. Computational Linguistics **23** (1997)
10. Mohri, M., Pereira, F., Riley, M.: The design principles of a weighted finite-state transducer library. Theoretical Computer Science **231** (2000) 17–32
11. Oflazer, K.: Error-tolerant finite-state recognition with applications to morphological analysis and spelling correction. Computational Linguistics **22** (1996) 73–89
12. Canals, R., Esteve, A., Garrido, A., Guardiola, M., Iturraspe, A., Montserrat, S., Pastor, H., Pérez, P., Forcada, M.L.: The spanish-catalan machine translation system internostrum. In: MT Summit VIII: Machine Translation in the Information Age. (2001) 73–76
13. Sanchis, A., Picó, D., del Val, J., Fabregat, F., Tomás, J., Pastor, M., Casacuberta, F., Vidal, E.: A morphological analyser for machine translation based on finite-state transducers. In: MT Summit VIII: Machine Translation in the Information Age. (2001) 305–309
14. Viterbi, A.: Error bounds for convolutional codes and an asymtotically optimal decoding algorithm. IEEE Transactions on Information Theory (1967) 260–269
15. Molina, A., Pla, F., Moreno, L., Prieto, N.: Apoln: A partial parser of unrestricted text. In: Proceedings of 5th Conference on Computational Lexicography and Text Research COMPLEX-99, Pecs, Hungary (1999) 101–108
16. Carmona, J., Cervell, S., Màrquez, L., Martí, M., Padró, L., Placer, R., Rodríguez, H., Taulé, M., Turmo, J.: An environment for morphosyntactic processing of unrestricted spanish text. In: Proceedings of the 1st International Conference on Language Resources and Evaluation, LREC, Granada, Spain (1998) 915–922