A MORPHOLOGICAL ANALYSER FOR MACHINE TRANSLATION BASED ON FINITE-STATE TRANSDUCERS

Alberto Sanchis[1], **David Picó**[1], **Joan Miquel del Val**[2], **Ferran Fabregat**[2], **Jesús Tomás**[3], **Moisés Pastor**[1], **Francisco Casacuberta**[1], **Enrique Vidal**[1]

[1]Institut Tecnològic d'Informàtica
Departament de Sistemes Informàtics i Computació
Universitat Politècnica de València
46071 València, SPAIN
{asanchis,dpico,moises,fcn,evidal}@iti.upv.es

[2] Servei de Normalització Lingüística
Universitat de València - Estudi General
46010 València, SPAIN
{Joan.M.Val,fabregat}@uv.es

[3] Departament de Comunicacions
Escola Politècnica Superior de Gandia
46730 Gandia, SPAIN
jtomas@dcom.upv.es

Abstract

A finite-state, rule-based morphological analyser is presented here, within the framework of machine translation system TAVAL. This morphological analyser introduces specific features which are particularly useful for translation, such as the detection and morphological tagging of word groups that act as a single lexical unit for translation purposes. The case where words in one such group are not strictly contiguous is also covered. A brief description of the Spanish-to-Catalan and Catalan-to-Spanish translation system TAVAL is given in the paper.

Keywords: Morphological Analysis, Finite-State Transducers, Word Graphs

Introduction

Rule-based techniques are the usual approach for building general machine translation systems. However, example-based approaches have experienced an increasing interest in different problems of machine translation during the last decade. These approaches have shown competitive performance in dealing with translation tasks in a restricted-domain language and have also been useful in particular subproblems that arise in rule-based translation, such as POS tagging (Abney, 1997; Dagan et al., 1997) or finding non-strictly linguistic relations among words or phrases in specific tasks (Brown et al., 1990; Brown et al., 1993; Al-Onaizan et al., 1999).

Finite-state machines have been successfully used for the implementation of both rule-based and example-based machine translation systems and tools for natural language processing (Rochar & Schabes, 1995; Oflazer, 1996; Mohri, 1997; Mohri et al., 2000). Finite-state techniques are very appreciated for their simplicity and high time performance.

TAVAL is a Spanish-to-Catalan and Catalan-to-Spanish translation system that aims to combine adequately both rule-based and example-based techniques on a finite-state framework. TAVAL takes advantage of the high degree of sequentiality existing between Spanish and Catalan that makes possible to avoid the analysis of certain complex word alignments and changes in syntactic structures. Up to now, there are a few machine translation systems that deal with these two languages. The newspaper *El Periódico* produces a Catalan version of news that are originally written in Spanish by using a memory-based translator. The SALT2 system is a rule-based computer-assisted translation system. INCYTA is a rule-based system that needs post-process corrections. InterNostrum is a rule-based system developed by a team of the Alacant University that uses finite-state technology (Canals et al., 2000). The TAVAL system will be tested for the translation of certain types of bilingual official publications by the Valencian government.

The TAVAL translator performs a partial analysis of the segment of text to be translated using a knowledge-based dictionary. An efficient finite-state representation of the dictionary was chosen following some guidelines from the grammatical inference framework (Oncina et al., 1993). The analysis is performed by means of a Viterbi-like algorithm that is used for translation with finite-state transducers (Amengual et al., 2001). The output of the analysis is a word-graph that represents different possible analysis (and

consequently different possible translations) of the source sentence.

Morphological tagging of the input language in a translation system presents some specific problems that do not usually come up in morphological tagging of a single language. The translation between two languages very often implies the detection of word groups in the source language that act as a single lexical unit for the purpose of translation. For instance, the Spanish expression "por favor" most of times should be tagged as a word group working as an adverb, rather than as two different words working as a preposition and a noun, respectively. Other more complex examples include idiomatic expressions, periphrasis, etc., that are not translated to the target language word by word but as a whole, and therefore need to be detected and morphologically tagged as a whole in the source language.

The morphological analyser in TAVAL differs from other ones in the treatment of such word groups. TAVAL can detect these idiomatic expressions and is able to produce an ambiguous lexical tagging in which words may have been interpreted individually or belonging to lexically significant groups. This kind of analysis makes possible to improve the quality of translations.

The purpose of this paper is to report on the structure and implementation details of the morphological analyser. This is explained in section "TAVAL morphological analyser". In order to put the analyser into context an overview of the complete architecture of the TAVAL translation system is given in the next section. Finally, some conclusions are given together with the main guidelines for future work.

System architecture

The TAVAL system follows a modular architecture in which text is processed through a cascade of *black boxes* that handle different phases of the translation process. In an overview, the main modules are the following:

1. A *fragmenting* module that breaks up the input text into "fragmentation units" (FU), so that the subsequent translation problems to be treated by the next modules are simpler.

2. An *identifying* module able to identify and mark out *identifiable translation units* (ITU), i.e., translation units that may be identified with the available (local) context.

3. A *tagging* module that assigns the corresponding possible grammatical categories to each translation unit (TU) through a process of morphological analysis.

4. A statistical module for *category disambiguation* that decides what is the correct grammatical category to which each translation unit belongs, in the context of a given fragmentation unit.

5. A rule-based *transference* module that finds the equivalent sentence in Catalan for the Spanish input, given the segmentation and linguistic information yielded by the previous modules.

Our work until now has been centred on the first four modules. This paper is devoted to the the morphological analyser, module 3, which is described in section "TAVAL morphological analyser". In the following paragraphs we will give a glimpse of the other pieces of the system, so as to put the analyser into context.

In system TAVAL the linguistic information about translation units is compelled in one basic and several specific translation dictionaries containing morphological, syntactic, semantic, contextual and translation information for each translation unit (formed by individual words or by word groups). The basic dictionary refers the most statistically significant words in a general language register while the specific dictionaries deal with specialized subjects and are intended to be used for particular types of texts (technical and scientific texts, legal documents, commercial letters, etc.)

The aim of the fragmentation module is to break up the full-text input into linguistic fragments that can be viewed as a unit for translation purposes. In our project we are interested in detecting text sections, paragraphs, sentences and tokens, and also in marking out some identifiable translation units such as abbreviations, acronyms or proper nouns. The detection of articles and paragraphs is easy, since there usually are explicit marks in the source text that indicate the limits of this kind of units. However, the sentence-level fragmentation is a non-obvious problem. The major difficulty is that the symbol used most often to indicate the end of the sentence, the dot, is also used with others purposes –to indicate abbreviations or as a part of a numerical expression. Token detection encounters similar difficulties. The character more commonly used to separate tokens is the white space, but also other symbols such as hyphens or punctuation marks. Our fragmenting module has been implemented using finite-state techniques that compile knowledge-based rules and lists of known abbreviations, acronyms, proper nouns, etc., in a parsing mechanism that is both space- and time-efficient. We have performed experiments with an 18-million-words bilingual corpus extracted from the bilingual edition of the newspaper *El Periódico*. We have tested our system on 100 randomly chosen paragraphs. Sentences, numbers, proper nouns and abbreviations were detected with less than 1% of errors, and acronyms were detected with a 16% of errors.

Once the input text has been processed by the labelling module (described on the next section) we obtain an ambiguous word graph in which individual words and/or word groups may have received one or more labels indicating an entry in the dictionary and specific morphological information. This graph must be disambiguated so that labelling is reduced to one label per word or group of words. The technique employed in TAVAL is an adaptation of that described by Pla (2000) for POS tagging. This technique combines grammatical inference and statistical models in a machine learning paradigm. This same methodology has been generalized for modelling linguistic units such as noun phrases that will be necessary in the transference module for solving translation problems such as gender or number agreement.

For the transferring module it is important to remark that, given the important similarities existing between syntactic structures in Spanish and Catalan, *it is not necessary to specify a huge set of transference rules*, as it is when building ruled-based translation systems for more dissimilar languages. In essence, two types of transference rules need to be applied when translating from Spanish to Catalan: changes in gender and number of words and (contextual) semantic disambiguation. A set of specific transfer rules for particular problems such as the appearance of weak pronouns in Catalan, the change or fall of prepositions, or the change in verb tense, is also necessary.

TAVAL morphological analyser

A morphological analyser provides lexical, morphological and syntactical information for each lexical unit in the analysed sentence. Most morphological analysers can only analyse one-word lexical units. This fact implies an important limitation since there are groups of words that have their own lexical entity (idiomatic expressions, periphrasis, etc.) and therefore should be analysed as a unique, independent entity. A morphological analyser that is not able to detect these composed lexical units will force incorrect translations, since some special expressions that should be detected, morphologically tagged and translated as a single unit will receive instead a word-by-word analysis. The TAVAL morphological analyser can detect significant word groups and give specific morphological information for the word group.

The TAVAL morphological analyser is based on a morphological dictionary where morphological characteristics (lemma and inflection paradigm) for each lexical unit are stored, together with the corresponding syntactic category. This morphological dictionary is internally represented by a set of finite-state transducers (FST) that are automatically generated from a more general dictionary containing morphological, syntactic, semantic, contextual and translation information built by expert linguists. Details on this are given in the next subsection.

Morphological analysis is performed by a Viterbi-like parsing of the input sentence through the finite-state network. Lexical ambiguity causes that for a given input string different possible analysis (and consequently different possible translations) of the input string can be performed. For this reason, the output of the algorithm is a word graph in which all the possible analysis of the input string (and consequently all the different possible translations) are compactly represented.

A morphological dictionary based on finite-state transducers

A FST is composed by a finite set of states and a set of transitions between pairs of states. Each transition is labelled by a symbol from the input vocabulary and by a string of symbols that belong to the output vocabulary. Some states can be final states, in which case may have an associated output string.

The TAVAL morphological dictionary is based on a set of FSTs. The main FST contains the morphological and syntactical information for all the (single and complex) lexical units in the dictionary. The input alphabet is composed of Spanish characters plus a set of special symbols that reference morphological patterns as it is explained below. The output alphabet is composed of special symbols that represent syntactical categories (nouns, adjectives, verbs, pronouns, etc.) and morphological characteristics (structure of each word or group of words).

Simple lexical units are represented in the main FST by a set of edges describing the spelling of the lemma (each edge is labelled with a single character) followed by some edges that describe the possible morphological patterns. Representation of complex lexical units is done by concatenating the morphological descriptions of the individual words. It is possible to insert special symbols that indicate that a complex lexical unit can be splitted by some particular kind of word (adverbs, adjectives and some other) –i.e., the Spanish expression *echar de menos*, meaning to miss someone or something, can appear modified by adverbs that are inserted immediately after the verb: *echar mucho de menos*, *echar a menudo de menos*, etc.– See an example of some entries of this dictionary in figure 1.

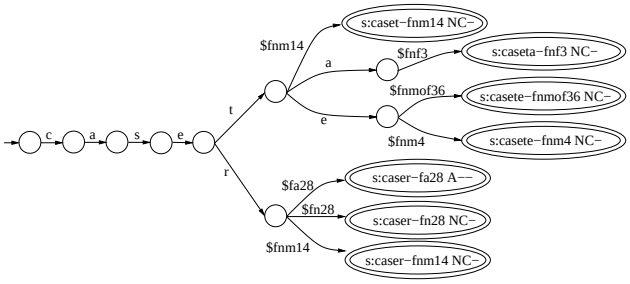

Figure 1: A simplified example of the morphological dictionary structure. The special symbols ($fnf3, $fnmof36, $fa28, etc.) serve as a reference to another subFST. An example of subFST (specifically $fa28) is showed in figure 2.

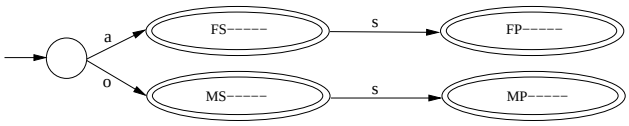

Figure 2: An example of subFST.

Lexical units share inflection paradigms, combination with pronouns, composition with suffixes, etc. Each one of these common morphological patterns is modelled by one *sub*FST. In this way, morphological information is efficiently represented in the main FST by using edges that are labelled with special symbols which serve as a reference to these subFSTs. Therefore, the main FST does not need to incorporate an explicit expansion of all the subFSTs (see figure 1). On the other hand, the main FST is structured as a prefix-acceptor tree so common prefixes of the lemmas are compactly represented.

Searching inside the dictionary

Given an input string w, the process of analysis searches for the sequence of states (path) in the network that parses

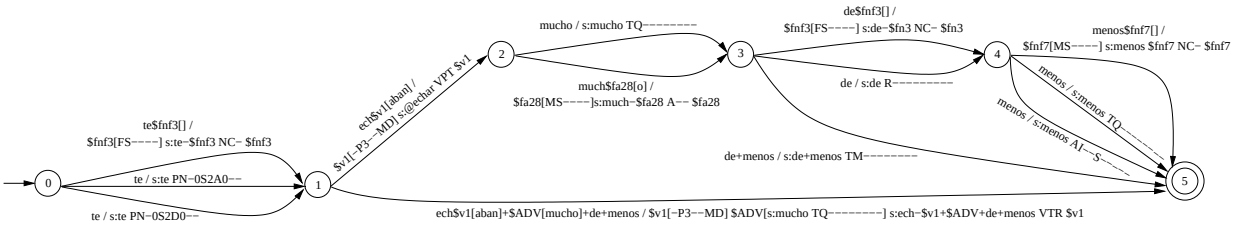

Figure 3: Word graph generated by the morphological analyser for the input Spanish sentence "te echaban mucho de menos". Two complex lexical units have been detected: "echaban mucho de menos" (transition:1-5) and "de menos" (transition:3-5).

w and goes from the initial state to a final state, and then outputs the sequence Ω of output symbols that is associated to the sequence of states. The output string Ω will contain the syntactical and morphological information for the input string w.

The lexical ambiguity causes that for a given input string w more than one path can be found between the initial state and final states. Therefore, different possible analysis of the input string can be performed (and consequently different possible translations can be obtained). For this reason, the output of the algorithm is a *word graph* in which all the possible analysis of the input string are compactly represented. Each word graph is a FST where every path from the (unique) initial state to the (unique) final state represents one possible way to analyse the original input string w. The different analysis are described by the input symbols in the word graph, while the corresponding morphological information for each analysis is recorded using output symbols. See an example of a word graph in figure 3.

The analysis is performed using a Viterbi-like algorithm. During the search process when a reference to a morphological pattern is found the associated FST is expanded dynamically. In this way, the use of FSTs during the process of analysis naturally indicates the morphology of the analysed sentence. As said above, the algorithm outputs a word graph that records all the paths within the finite-state network that have reached a final state and, therefore, are valid hypothesis. The word graph can be seen as a compact representation of all possible analysis of the input sequence by the morphological analyser.

Conclussions

An efficient finite-state morphological analyser has been described. The aim of this analyser is its use as the first step in a Spanish-to-Catalan translation system. The efficiency of the analyser is due to the use of finite-state models and associated search algorithms. An example of a word graph produced by the analyser is presented in figure 3 and illustrate the behaviour of the analyser for different types of sentences.

The next step is to prune the unnecessary paths in the output word graph in order to produce a unique analysis of the input string. This step will be carried out using statistical methods for category disambiguation.

Acknowledgements

This work has been partially founded by the Spanish CICYT-FEDER under grant TIC 1FD1997-1433.

References

Abney, S. (1997). "Part-of-Speech Tagging and Partial Parsing". Corpus-Based Methods in Language and Speech Processing. S. Young and G. Bloothooft (eds.). Kluwer Academic Publishers.

Al-Onaizan, Y. and Curin, J. and Jahr, M. and Knight, K. and Lafferty, J. and Melamed, D. and Och, F.J. and Purdy, D. and Smith, N.A. and Yarowsky, D. (1999). "Statistical machine translation". Tech. Rep. Final Report, JHU Workshop, John Hopkins University.

Amengual, J.C. and Benedí, J.M. and Casacuberta, F. and Castaño, A. and Castellanos, A. and Jiménez, V.M. and Llorens, D. and Marzal, A. and Pastor, M. and Prat, F. and Vidal, E. and Vilar, J.M. (2001). "The EUTRANS-I Speech Translation System". Machine Translation Journal (to be published).

Brown, P.F. and Cocke, J. and Della Pietra, S.A. and Della Pietra, V.J. and Jelinek, F. and Lafferty, J.D. and Mercer, R.L. and Roosin, P.S. (1990). "A statistical approach to machine translation". Computational Linguistics, vol. 16, no. 2, pp. 79–85.

Brown, P.F. and Della Pietra, S.A. and Della Pietra, V.J. and Mercer, R.L. (1993). "The mathematics of statistical machine translation: Parameter estimation". Computational Linguistics, vol. 19, no. 2, pp. 263–310.

Canals, R. and Garrido, A. and Guardiola, M. and Iturraspe, A. and Montserat, S. and Pastor, H. and Forcada, M.L. (2000). "Herramientas para la construcción de sistemas de traducción automática: aplicación al par castellano-catalán". Proceedings of the IV Congreso de Lingüística General.

Dagan, I. and Lee, L. and Pereira, F. (1997). "Similarity-Based Methods for Word Sense Disambiguation". Proceedings of the 35th Annual Meeting of the ACL and 8th Conf. of the EACL. Madrid (Spain), pp. 56-63.

Mohri, M. (1997). "Finite-state transducers in language and speech processing". Computational Linguistics, vol. 23, no. 2.

Mohri, M. and Pereira, F. and Riley, M. (2000). “The design principles of a weighted finite-state transducer library”. Theoretical Computer Science, vol, 231, pp. 17-32.

Oflazer, K. (1996). “Error-tolerant Finite-state Recognition with Applications to Morphological Analysis and Spelling Correction”. Computational Linguistics, vol, 22, no. 1, pp. 73-89.

Oncina, J. and Garcia, P. and Vidal, E. (1993). “Learning Subsequential Transducers for Pattern Recognition Interpretation Tasks”. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 15, no. 5, pp. 448-458.

Pla, F. (2000). “Etiquetado léxico y análisis sintáctico superficial basado en modelos estadísticos”. Ph.D.Thesis. Universidad Politécnica de Valencia.

Rochar, E. and Schabes, Y. (1995). “Deterministic Part-Of-Speech Tagging with Finite State Transducers”. Computational Linguistics, vol. 21, no. 2, pp. 227-253.