A Part-of-Speech Tagger for Esperanto oriented to MT

Carlo Minnaja
Università di Padova & AIS San Marino
minnaja@math.unipd.it; ais@inthenet.sm

Laura G. Paccagnella
Università di Padova
laurap@math.unipd.it

Dipartimento di Matematica Pura ed Applicata,
Via Belzoni 7 - IT 35131 Padova, Italy.

Abstract

In this paper we propose a PoS tagger for Esperanto; this language has been already used as an intermediate language for MT, but the very promising work stopped nearly ten years ago at the level of a prototype English-French. Esperanto benefits of the fact that in absolutely most cases the ending of a single word determines its part of speech, and also the accusative case for the direct object or for phrases of motion. A short list of words where this does not happen was identified and stored apart. Then a tagset was built up, the first one for Esperanto, which is quite different from the various tagsets for English. The tag identifies much more than merely the part of speech, as also accusative cases and a large category of intransitive verbs can be determined, and this is a big help in the construction of trees. In the paper we present the tagset, the tagging algorithm and the results of an automatic tagging.

Introduction

The first attempt to involve the international language Esperanto in MT was in the early eighties. In the period 1984-1990 BSO, a Dutch software company, carried out the Distributed Language Translation project. As the name suggests, it was a system planned for multilingual online services in computer networks, based on an intermediate language (IL). The system was planned to be fully automatic, although a disambiguating dialogue technique was helpful in the step from the source language to the IL. The project was stopped when a prototype for an English-French MT was completed (Sadler 1991). Actually, the IL utilised was very similar to Esperanto, and only a very little number of modifications was needed in order to make it a fully satisfactory IL. Some personal pronouns were added, and some other slight changes were made; they concerned, for instance, ambiguities of preposition *de* (of, by) and a too large use of the accusative. The use of such an IL had two advantages. The first was that it was very flexible, clear, and widely unambiguous; the second one was that a man knowing Esperanto could check the two steps of the translation, from English into the IL and from the IL into French.

Besides the mentioned prototype, the main by-product of the DLT project was the Bilingual Knowledge Bank (BKB). It comes out from two separate knowledge banks: one English/Esperanto, the other Esperanto/French. The BKB English/French is built up automatically merging the combinations of the two. At that time the syntactic word categories, the Parts of Speech (PoS), were stored in the BKB manually, as well as nearly 76,000 example sentences or phrases, organised in trees and subtrees.

A prototype of an MT system having directly Esperanto as a source language was ROMANO, treated in de Kat (1985). But it remained at the level of some trials.

1. Some basic principles

Redundancy is present in every language: the "principle of redundancy" states that every communication system must have a certain amount of redundancy, so that the receiver of the message can reconstruct it even if some elements have been lost. On the other hand, the "principle of language opportunity" (Wüster 1966) gives the user the possibility to express himself by the least effort[1]. Esperanto is a good compromise between both. When it comes to word endings, plural is used both in nouns and in adjectives; participles have different endings when they are nouns, or adjectives (singular and plural), and they have the adverbial ending when their verbal function is prevailing. Accusative, too, is inflected both in noun and in adjectives. On the other hand, in Esperanto a participle functioning as a noun can avoid a full relative sentence.

The DLT project utilised a hidden tagset for Esperanto, and no critical reasons have been presented about a specific choice rather than another. The problem of tagging words was not the main one in the project. So it was nearly impossible to take advantage from the work already done in the DLT project.

Our work starts nearly at the beginning in order to build a tagger for Esperanto, which automatically gives a unique tag to each word, when processing a text. As the Esperanto grammar has also features which can be exploited when it comes to build up trees for MT, we kept in mind this while building the tagset.

2. The corpus

Esperanto uses the Latin alphabet with diacritics. Nevertheless, a machine-readable form for a text is not straightforward: although a lot of texts and reviews are electronically processed, there is no standard way in processing the diacritics. The most largely read review, organ of the Universal Esperanto-Association, uses an ASCII transcription with a backslash followed by numbers, and this messes up the endings as well as the statistical data about word length. Other systems use "Latin extended 3", a set of characters having the Esperanto letters: they are successfully processed by some printers, but not by all programs.

In order to have a corpus as representative as possible, we collected texts from various sources, and we made them uniform by transcription of every text to the same system. The sources we collected from are newspapers[2], literary magazines, private messages in the net, documents downloaded from Internet. Such different sources and subjects offered a very wide set of linguistic and technical situations: abbreviations, acronyms, in-word dashes, periods not followed by a capital, special symbols as @, quotation marks, different types of brackets, titles, and so on. Words having endings out of the Esperanto grammar have been classed as foreign words. A side consideration: errors, misprints, loss of parts when a file is sent to another server are *much more frequent* than one usually can guess.

3. The algorithm

We built up a tagset, according to the features of Esperanto, e.g. its inflection; the tagset is in Table 1, and is similar to the one proposed in Minnaja (1998).

	Tag	Description	Examples
1.	``	straight single (or double) quote	`` "
2.	$	currencies	$, £
3.	(	opening parenthesis	([{
4.	)	closing parenthesis	)] }
5.	,	comma	,
6.	.	sentence terminator	. ! ?
7.	:	colon, ellipsis	: ; ...
8.	CC	coordinative conjunction	kaj, ke, tamen
9.	NUM	numeral, cardinal	unu, dek, 100
10.	DT	determiner	la
11.	FW	foreign word	Gesellschaft
12.	PREP	preposition	de, kun
13.	ADJ	adjective	bona
14.	ADJPL	adjective, plural	bonaj
15.	ADJOBJ	adjective, object	bonan
16.	ADJPLOBJ	adjective, plural, object	bonajn
17.	N	noun	universitato
18.	NPL	noun, plural	universitatoj
19.	NOBJ	noun, object	universitaton
20.	NPLOBJ	noun, plural, object	universitatojn
21.	NP	noun, proper	Karlo, Jack
22.	PP	pronoun, personal	mi, ili
23.	PPOBJ	pronoun, personal, object	min, ilin
24.	PPS	pronoun, possessive	(la) mia
25.	PPSOBJ	pronoun, possessive, object	(la) mian
26.	PPSPL	pronoun, possessive, plural	(la) miaj
27.	PPSPLOBJ	pronoun, possessive, plural, object	(la) miajn
28.	ADV	adverb	certe
29.	ADVOBJ	adverb, direction	tien, aliloken
30.	PRT	particle	ĉu
31.	SYM	symbol	%, &, #
32.	UH	interjection	ho
33.	VBP	verb, present tense	faras
34.	VBD	verb, past tense	faris
35.	VBF	verb, future tense	faros
36.	VBIMP	verb, imperative	faru
37.	VBS	verb, subjunctive	farus
38.	VBI	verb, infinitive	fari
39.	VBNTP	verb, intrans., present	iĝas
40.	VBNTD	verb, intrans., past	iĝis
41.	VBNTF	verb, intrans., future	iĝos
42.	VBNTIMP	verb, intrans., imperative	iĝu
43.	VBNTS	verb, intrans., subjunctive	iĝus
44.	VBNTI	verb, intrans., infinitive	iĝi
45	NEG	negation	ne

Tab. 1

As it can be easily seen, some of the tags have the features of a *supertag*, following the concept of Bangalore and Joshi (1999), as they identify not only the PoS, but also the syntactical function. A comparison with the Penn Treebank tagset shows the lack of some tags, like the ones about comparative and superlative, or about the 3rd singular person of the verbal present tense. On the contrary other tags are added, like the one about the certainly intransitive verbs, or the plural of the adjectives, or the accusative case.

Then the algorithm moves according to different steps. It identifies nearly 280 words, mostly short, whose PoS tag is independent of the ending; these words have been stored apart and their tags have been attributed manually; then it assigns one tag, according to the word ending (one letter, or a group of letters). For a tagger oriented to MT, we had to distinguish possessive adjectives from possessive pronouns: they have the same ending in Esperanto, but they are different in some other languages (*my/mine*; *mon/le mien*). The algorithm looks at the previous word: if it is a determiner, the word processed is taken as a pronoun.

As far as we need, in order to run a program which gives the tags automatically, we did not distinguish the sex in the personal pronouns, but in the third singular person[3]. In addition, we did not distinguish the participles from other elements having the same PoS. Words such as *amato* (loved) or *amantino* (lover, feminine) are merely classed as nouns; *amante* (while loving) is classed as an adverb.

The endings taken into account are *o* (singular noun, nominative case), *a* (singular adjective, nominative), *j* (plural of nouns and adjectives, nominative), *e* (adverb), *n* (accusative case for nouns, adjectives, singular and plural and for moving to a direction), and the verbal ones, as *as* (presence), *is* (past), *os* (future), *us* (subjunctive), *i* (infinitive), *u* (imperative). The endings of the accusative cases (singular and plural) are useful in order to build up some trees. Other specific endings could be taken into account, if suffixes too would come into consideration: this would lead to a certain number of supertags. A specific work about this is now in progress. Nevertheless, in order to orient the tagger to MT we took into consideration one suffix, located right before the verbal ending, as it is linked to verb intransitivity. Suffix *-iĝ-* can be found only in intransitive verbs, and indicates that the subject is coming into the state expressed by the root. Of course, many verbs are intransitive without having these suffixes; so we classed the verbs into two categories: certainly intransitive, and others. With regard to the list of short words we mentioned before, they were set apart, and the respective tags were given manually. It deals with the definite article (*la*), personal pronouns (e.g. *mi, vi, ili, oni*), interjections (e.g. *ha!, hu!*), demonstratives (e.g. *tiu*) and others. This way ambiguity has been completely cut off.

The tagger works extremely well, with an accuracy of more than 99%: actually, it fails only at misprints. More details will appear in Minnaja and Paccagnella (2000).

4. An example

Let us present now an example of a sentence in Esperanto, and how it is processed by the tagger.

Dum vasta kaj multaspekta debato oni enfokusigis originalajn flankojn de la lingvo-problemo ne nur rilate la rolon de Esperanto en la Eŭropa Unio, sed ankaŭ koncerne aliajn problemojn. (During a large and various debate original aspects of the language problem were brought into focus, not only about the role of Esperanto in the European Union, but also about other problems.)

Dum#CC	la#DT	la#DT .
vasta#ADJ	lingvo-problemo#N	Eŭropa#ADJ
kaj#CC	ne#NEG	Unio#N
multaspekta#ADJ	nur#PREP	,#,
debato#N	rilate#ADV	sed#CC
oni#PP	la#DT	ankaŭ#ADV
enfokusigis#VBD	rolon#NOBJ	koncerne#ADV
originalajn#ADJPLOBJ	de#PREP	aliajn#ADJPLOBJ
flankojn#NPLOBJ	Esperanto#N	problemojn#NPLOBJ
de#PREP	en#PREP	.#.

Notes

(1) The principle of the least effort is used in various fields of linguistics; an example for anaphora is in Paccagnella (1998).

(2) Courtesy of the International Academy of Sciences San Marino, Universal Esperanto-Association and Italian Esperanto-Federation.

(3) The DLT system created some additional pronouns just to distinguish the sex also in the first and second person, and in the third plural too. We did not take care of this distinction.

References

Bangalore S. and Joshi A. (1999), Supertagging: An approach to almost parsing, *Computational Linguistics*, 25(2), pages 237-265.

de Kat J. O. (1985), 'Traduko el la internacia lingvo al naciaj' (Translation from the International Language to National Languages), in Koutny, I. (ed), Perkomputila tekstoprilaboro (Automatic Text Processing)}, Scienca eldona centro: Budapest, pages 259-266.

Minnaja C. (1998), Standartoj por aŭtomata etikedado de la parolelementoj – Standards for automatic part-of-speech tagging, in *Actes 15e Congr. Int. Cybernétique*, Association Internationale de Cybernétique: Namur, pages 745-750.

Minnaja C. and Paccagnella L. G. (2000), Anaphora with Relative Pronouns: an Algorithm for Italian, Esperanto and Latin, submitted to *Computational Linguistics*.

Paccagnella L. G. (1998), The minimization of effort in the use of anaphora, *Cybernetica*, XLI, 1, pages 57-65.

Sadler, V. (1991), Machine Translation Project Reaches Watershed, *Language Problems & Language Planning*, 15(1), pages 78-81.

Wüster, E. (1966), Internationale Sprachnormung in der Technik, Sprachforum, Beiheft 2: Bonn.