Tagging Spanish Texts: the Problem of *'se'*

G. Aguado de Cea, J. Puche Alosete, J.A. Ramos

Ontology Engineering Group, Universidad Politécnica de Madrid

Campus de Montegancedo, Avda. Montepríncipe s/n, Boadilla del Monte, Madrid, Spain

E-mail: lupe@fi.upm.es, javier.puche@educa.madrid.org, jarg@fi.upm.es

Abstract

Automatic tagging in Spanish has historically faced many problems because of some specific grammatical constructions. One of these traditional pitfalls is the *'se'* particle. This particle is a multifunctional and polysemous word used in many different contexts. Many taggers do not distinguish the possible uses of *'se'* and thus provide poor results at this point. In tune with the philosophy of free software, we have taken a free annotation tool as a basis, we have improved and enhanced its behaviour by adding new rules at different levels and by modifying certain parts in the code to allow for its possible implementation in other EAGLES-compliant tools. In this paper, we present the analysis carried out with different annotators for selecting the tool, the results obtained in all cases as well as the improvements added and the advantages of the modified tagger.

1. Introduction

Automatic tagging in Spanish has historically faced many problems because of the difficulty of some specific grammatical constructions. One of these traditional pitfalls is the *'se'* particle. This particle is a multifunctional and polysemous word used in many different contexts. When extracting taxonomical relations from annotated texts dealing with the classification language, the *'se'* particle appears in many grammatical constructions, such as *se clasifica(n), se divide(n)...* The type of taxonomical relations shown is SUBCLASS_OF. This relation is fundamental in several fields, such as knowledge extraction and mapping discovery between ontologies. Thus, the need to rely on a tagger that annotates this usage became a key issue. However, many taggers do not distinguish the possible uses of the Spanish *'se'* and thus provide poor results. Our aim was not to develop a new complete tagger for Spanish, as we did not want to "reinvent the wheel", but rather to reuse and improve a current tool. Reusing and enhancing a free tool was our priority idea, in order to make the most of free resources. According to this criterion, several tools were analyzed to check their behaviour in this point. In this paper we present the analysis carried out and the results obtained as well as the improvements added and the advantages of this new resource.

2. Brief overview of linguistic taggers in Spanish

Nowadays, automatic taggers seem to have left aside the problem of differentiating the grammatical, semantic and pragmatic values that 'se' can take in Spanish. From the linguistic viewpoint, this particle has been widely studied in conventional Spanish grammars (Seco, 1954; Gómez Torrego, 1997; RAE, 1999), as well as in other works specifically devoted to this particle (Sánchez-López, 2002; González- Vergara, 2006). But, although most scholars seem to agree on the 'pronoun' nature of the particle, there are other occurrences of 'se' that do not fit exactly under this label, as Sánchez-López, and González-Vergara point out. From the computational view, this general undefinability is also reflected in the tagging tools for Spanish, that usually simplify the possible values of 'se'. However, some authors, (Fernández et al., 2004; Aguado et al., 2003) have made an interesting attempt to add other values. In the latter work, the authors propose 4 values based on EAGLES (1996): reflexive pronoun (Juan se lava la cara), reciprocal pronoun (No se saludan), personal pronoun (Se lo daba), and passive or impersonal marker and others (Se aplaudió a los artistas).

Based on these four values we carried out a comparative analysis of 6 tagging tools (SVMTool (Giménez and Márquez, 2004), FDG Connexor1, DataLexica2, Wraetlic tools 2.03, TreeTagger4, and FreeLing 1.5 (2007)) with the same 60 Spanish sentences with 'se' (see Annex II). The rate of scores obtained for the 'se' particle was very low, for our purposes, ranging from 0% up to 13,33%, as most tools did not consider the problem of the *'se'* particle at the fine-grained level that we needed. FreeLing 1.5 achieved 65% of scores, tagging only two values. This evaluation can be consulted at http://webode.dia.fi.upm.es/Puche/ Puche.html. Thus, it was clear the need of a tool that could improve this score.

3. Analysis criteria applied

The criteria used to select the tool were the following:
- Implementation of different annotation levels (POS, syntactic, chunking).
- Approaches applied in these levels: algorithms in stochastic models, linguistic rule-based systems, etc.
- Real coverage of the implemented leves to Spanish texts.
- Flexibility: ability to change the tool's behaviour through configuration, extension or even modification of the tool itself in the case of open source models.
- Availability: whether the tool is free, at least, for research purposes (our case).
- Ease of integration: in order to make it work together with the OntoTagger (Aguado et al., 2003) tool that only accepts external tools with a java API running on

---

[1] http://www.connexor.com/

[2] http://www.bitext.com

[3] http://alfonseca.org/eng/research/wraetlic.html

[4] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

- Windows platforms.
- Support level: type of support granted for reporting tool malfunctioning, bug corrections, suggestions and average response time.
- Documentation: quality and clarity of the documentation and the possibility of modifying its internal structure.

Once these criteria were analyzed in all tools, FreeLing 1.5 (2007) was selected mainly for the following reasons (for a more detailed account see (Puche, 2007)).

- It was the best according to the coverage criterion. It includes a EuroWordNet subset for free (full with license) with sense annotation, Named Entity Recognition and Classification, suffixation rules, stochastic and rule-based models for POS tagging and a complete Context-Free Grammar (CFG) with dependencies.
- It complied with the flexibility criterion, since it permits modification of the behaviour of all the NLP stages through well documented configuration changes and, being OpenSource, it allows modification of the tool itself, if needed.
- It also complied satisfactorily with the support level criterion. Specific forums are available to discuss any problem or suggestion about the tool and the response time from the development team and/or volunteers is really short.
- It is free of charge, and use, distribution and modification through the LGPL licence.
- It is distributed with high quality documentation together with the tool's internals for possible modifications.

4. Enhancing FreeLing

The improvements and modifications in FreeLing 1.5 were carried out at six levels. We first started at the POS level because this level is the most mature level in taggers. Secondly, these POS annotation corrections are necessary to obtain better results in the other levels.

4.1. Lexicon

As the lexicon is an important part, we have updated it to cover the four different uses of '*se*' suggested in the OntoTag Tagset (Aguado et al., 2003):
- P0000000: passive and impersonal.
- P0300000: reflexive pronoun.
- P03NP000: reciprocal pronoun.
- PP3CN000: personal pronoun substituting *le/les*.
- P0400000: undefined at this processing stage.

FreeLing has a configuration file that contains the lecial probability of assigning a label to a given word. This file present the following value for 'se':

```
se P0-PP P0 1166 PP 172
```

This file has not been modified because it only states the probability of '*se*' as a personal pronoun. However, this value can be changed by training Freeling with a corpus, although in our experience this is not necessary for the moment being.

4.2. Re-tokenization rules

We also changed the tokenization rules in order to include the new values of '*se*', adding its use as a verbal suffix: *dáselo, irse,* etc. In some cases, such as '*Van a venderse los coches*', '*No paran de saludarse*'. The rules for '*se*' are as follows:

```
se *  ^VMM03 *  0  1  0  1  $$+se:$$+PP
se *  ^V  *  1  1  0  1  $$+se:$$+PP
```

In this case, we have modified the second one as:

```
se *  ^V  *  1  1  0  1  $$+se:$$+P04
```

4.3. Morphosyntactic rules

At the POS level we improved the program in order to include the following rules:
- To take into account phrases with auxiliary verbs and periphrases, such as '*Se ha dado cuenta*', '*Se va a dar cuenta*'.
- To verify that the reflexive use only allows for third person verbal structures, as in the folloeing sentence: '*Se lo desabroché*'.
- To detect all reflexive verbs whether they are followed by a noun or not, a in '*Se lavan coches*' , '*Juan se lava la cara*'.
- To cover the reflexive use of some verbs indicating change when accompanied by adjectival predicates. Some of these examples correspond to support verbs, such as '*hacer*', for instance '*Se hacen fuertes*'.
- To include those cases where 'se' accompanying certain verbs is considered a personal pronoun if followed by another pronoun. For instance, '*Se lo voy a dar*', '*Se me ha caído*'.
- To mark the reflexive or reciprocal use of 'se' depending on the near words accompanying it: 'sí mismo', 'mutuamente', 'entre ellos/as', 'el uno al otro', etc. Among these cases, we can mention '*Se ayudan mutuamente*', '*Se besan entre ellas*', '*Se apoyan el uno al otro*'.
- To cater for those cases where verbs are reflexive if they take an animate subject. As we did not have ontological information at this stage, we included personal pronouns and proper names. Among the latter we also considered Named Entities (towns, etc.) so as to encompass metaphorical uses. The integration of the information with the top ontology could not be carried out because of time constraints whether pronominal verbs or periphrases, some examples corresponding to this group are: '*Se descarta él mismo*', '*Ella misma se prepara*', '*En Valencia se venden coches*', '*Él y Juan se odian*'.
- A complementary rule to the previous one gives priority to the passive use of '*se*' when the verb is followed by a noun, or a subject is preceded by a preposition.

All these rules are added to the default rules and they are not incompatible with the existent rules because they only add corrections to the '*se*' sentences.

By way of example of some of these rules, here we present following rule that corresponds to the point mentioned in the morphosyntactic rules, as it is the case of the support verb '*dar*':

```
5.0 P03*
      (0 (se))
      (1 <dar>)
      (2 <cuenta>);
```

4.4. Grammar and dependency rules

We updated the CFG grammar rules and the dependency files distributed with FreeLing 1.5 so that the new tags added to the lexicon could be used in the grammatical and dependency analysis.

4.5. Integration of external grammar rules from Volem

In order to enlarge verbal knowledge we codified external rules using the Java API, that combine the data obtained by the FreeLing chunker with other verbal patterns from the Volem project (Fernández et al., 2002).

4.6. Combination of verbal information rules derived from SenSem

We also proposed some rules using the Java API that combine the data obtained in the syntactical and dependency analysis of FreeLing with the verbal subcategorization patterns of the SenSem project (Vázquez et al., 2004). This has allowed us to differentiate some uses of '*se*' depending on the implicit and explicit complements of the verb.

5.   Results

The results obtained with the modified tagger show the great improvements achieved. Out of the same 60 sentences we obtained 56 scores (93,33%). Moreover, we identified also why the remaining 4 were incorrect and these problems will be solved in due course. For instance, some verbal affixes are not tagged according to FreeLing rules as in "*toca lavarse*". In another case, the verb is classified in the list of verbs with reflexive preference as in "*Desde mi casa se ve la torre de la iglesia*", "*No se vende esta mesa*".

On the reciprocal basis of free software these results as well as the new rules developed and the modified ones are available at http://webode.dia.fi.upm.es/Puche/Puche.html.

Instead of developing a new tool from scratch we tried to reuse, enhance and improve a free tool providing Windows portability and a Java API. By doing so, we achieved our first aim: reusing free resources. Besides, the rate of scores of this new tagger improved tremendously (93,3%) with all these additions, compared to the other tools analyzed (65%, 12%, 3%, 3%, 2% and 0%).

In summary, we modified successfully an annotation platform such as FreeLing and we merged some results from other projects such as SenSem and Volem.

Our future research will deal with extending the number of '*se*' values and refining the lists of verbs that are considered reflexive preferably. Moreover, a new version of FreeLing, that will solve the suffix rules problems, would have to be modified.

7.   References

Aguado, G., Álvarez-de-Mon, I., Gómez-Pérez, A., Pareja-Lora, A. (2003). OntoTag: XML/RDF(S)/OWL Semantic web Page Annotation in Content Web. EACL03. 10thProceedings of 3rd Workshop on NLP and XML (NLPXML-2003) Language Technology and the Semantic Web. Budapest, Hungary.

Eagles. (1996). Expert Advisory Group on Language Engineering   Standards. http://www.ilc.cnr.it/EAGLES96/home.html.

Fernández, A., Saint-Dizier, P., Vázquez, G., Benamara, F., Kamel, M. (2002). The VOLEM Project: a Framework for the Construction of Advanced Multilingual Lexicons. Proceedings of the Language Engineering Conference. Hyderabad, India.

Fernández, A., Vázquez, G., Castellón, I. (2004). La desambigüación automática de oraciones pronominales. J. Valera, J.M. Oró, J. Anderson (editores), Lengua y Sociedad: Lingüística aplicada en la era global y multicultural. Universidad de Santiago de Compostela:, p. 127-144. ISBN: 84-9750-398-9

Freeling. (2007). An Open Source Suite of Language Analyzers. http://garraf.epsevg.upc.es/freeling/

Giménez, J.; Márquez, L. (2004) SVMTool: A general POS tagger generator based on Support Vector Machines. Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04). Lisbon, Portugal. 2004

Gómez Torrego, L. (1997). Gramática didáctica del español. Madrid. Ediciones SM.

González-Vergara, C. (2006). Las construcciones no reflexivas con «se». Una propuesta desde la Gramática del Papel y la Referencia. Tesis doctoral. Universidad Complutense   de   Madrid http://www.gonzalezvergara.host.sk/escritos.htm

Puche, J. (2007). Etiquetado morfo-sintáctico automático de textos en castellano: el problema del 'se'. Pre-doctoral Dissertation. Universidad Politécnica de Madrid.

Real Academia Española. (1999). Gramática Descriptiva de la Lengua Española. Madrid. Espasa-Calpe.

Sánchez-López, C. (2002). Las construcciones con 'se'. Madrid. Visor.

Seco, R. (1954). Manual de Gramática Española. Madrid. Aguilar.

Vázquez, G., Fernández, A., Castellón I. (2004). El corpus Sensem: caracterización sintáctico-semántica de los verbos del español. XXXIV Simposio de la Sociedad Española de Lingüística. Madrid.

8. Annex I: List of Tools Analized

The following list presents the taggers evaluated and their url reference (see comparison in Table 1).

- FreeLing 1.5:
  http://garraf.epsevg.upc.es/freeling/
- VSIL CG-3:
  http://beta.visl.sdu.dk/cg3.html
- GATE:
  http://www.gate.ac.uk/
- LingPipe:
  http://www.alias-i.com/lingpipe/
- TreeTagger:
  http://www.ims.uni-stuttgart.de/projekte/corplex/
  TreeTagger/
- Wraetlic 2.0:
  http://alfonseca.org/eng/research/wraetlic.html
- @notate:
  http://www.coli.uni-sb.de/sfb378/negra-corpus/
  annotate.html
- AGFL:
  http://www.cs.kun.nl/agfl/
- Aries:
  http://www.mat.upm.es/~aries/
- Brill Tagger
  http http://www.cs.jhu.edu/~brill/RBT1_14.tar.Z
- Charniak:
  http://www.cs.brown.edu/people/ec/#software
- HMM Toolkit:
  http://htk.eng.cam.ac.uk/download.shtml
- Lovin:
  http://www.cs.waikato.ac.nz/~eibe/stemmers/
- Dan Bikel's:
  http://www.cis.upenn.edu/~dbikel/software.html
- Stanford NLP SW:
  http://nlp.stanford.edu/software/index.shtml
- TNT:
  http://www.coli.uni-sb.de/~thorsten/tnt/
- SnowBall:
  http://snowball.tartarus.org/
- Xerox Tagger:
  ftp://ftp.parc.xerox.com/
- YamCha:
  http://chasen.org/~taku/software/yamcha/
- SVM Tool
  http://www.lsi.upc.es/~nlp/SVMTool/

9. Annex II: Test sentences

The following sentences have been extracted from the grammars studied (see bibliography) enlarged with special conflictive sentences of all cases.

Valencia se lava la cara
En Valencia se venden coches baratos
En Valencia se lo montan muy bien
Se le cae la baba
Se alcanza a ver la torre desde aquí
Desde aquí a veces se ha alcanzado a ver Marruecos
Juan se moja
Juan se moja a sí mismo
Juan se ha mojado
Los niños se lavan
Se lavan los niños

Se lavan niños
Se lavan coches
Él se ve bien
Juan se ve bien
Pepe y el gato se miran mucho
Juan y Mara se van
Juan y Mara se besan
Mutuamente se ayudan
Se han vestido para la ocasión
Se ha bañado
Está claro que se miente a sí mismo constantemente
Se miente continuamente a sí mismo
Hoy se han cerrado Las Cortes
José se lava la cara delante del espejo
La carta se recibió oportunamente
La casa se hunde
La proposición se rechazó por todo el mundo
No se admiten propinas
No se atreve a irse (**first 'se'**)
No se vive muy bien que digamos en Madrid
Por la Dirección se han tramitado ya las órdenes oportunas
Se alquilan locales
Se cuentan verdaderos horrores de su crueldad
Se es cristiano
Se espera. (**ambiguous**)
Se espera al delegado
Se espera a los delegados
Se espera el premio
Se esperan los premios
Se está cayendo
Se habla de un nuevo gobierno
Se habla ya de un nuevo gobierno
Se ha hecho un traje
Se han escrito
Se les espera
Se suspenden las representaciones
Se vende
Se venden telas
Se venden telas baratas en el mercado
Se vive bien en Madrid
Si no quiere usted no se le obliga
Las montañas se han blanqueado
Se bailó hasta las tres
Valencia se achicharra
Toca lavarse
Desde mi casa se ve la torre de la iglesia
No se vende esta mesa
Se llevó las llaves.

1360

Table 1. Comparison of taggers

Tool	Functionality	Technologies	Spanish coverage	Language (.S.)	License type	Activity (Support)	Documentation (Papers)
FreeLing 1.5	Morpho, NERC, POS, Chunking, Shallow Grammar Parsing, Dependencies and Sense Annotation.	HMMtrigrams,constraint grammars(simplified), CFG, EWN lexicon, dependencies.	Included for all levels.	C++ (Unix)	LGPL	Very Active. Version history (forum)	Usage and internals (many)
VSIL CG-3	Constraint Grammar Parser..	Full Constraint Grammar Specification +Weights.	Proprietary.	C++ (Unix)	OpenSource	Finished (contact email)	User Manual. (many)
GATE	Integrator for other tools covering all the spectrum.	Diverse.	Chunker in beta-testing.	Java (independent)	OpenSource	Very active (distribution lists)	Usage and framework(many)
LingPipe	NER, k-best POS, Chunk.	confidence-scored.	Trainable.	Java (independent)	Non comertial sources	Active (forum)	User & tutorials (yes)
TreeTagger	POS.	Decision Trees.	Yes.	Executable for Linux and DOS	Non comertial No sources	Unknown (unknown)	Brief (2)
Wraetlic 2.0	Morpho, POS, NP Chunking, WSD.	Trigrams. Lesk.	Acceptable.	Java and C(Unix)	Free usage No sources	Incomplete, stopped since 2005 (none)	Scarse (1)
@notate	POS + NP and PP chunk.	TNT, Cascade Hidden Markov Model.	Unknown.	C + TclTk (Solaris,Linux)	Non comertial	Last release in 2006 (contact email)	Scarse (no papers)
AGFL	Shallow parsing.	Affix Grammars over a Finite Lattice CFGs.	Limited (CFG under construction)	C++ (Unix, Windows)	GPL for non comercial usage	Active (contact email)	Good (many)
Aries	Lexicón + Morpholical Análisis.	Chart, PATR-II.	Yes.	C++ (Unix, DOS)	Proprietary	Unknown (limited via email)	Unknown
Brill Tagger	POS.	Error-driven transformation-based lists.	Trainable..	C (many)	MIT OpenSource	Finished in 1994 (no)	Basic (influent)
Charniak	POS y parser estadístico.	Maximum-Entropy statistical + CFG.	Trainable.	C (many)	OpenSource	Finished in 2005 (contact email)	User (yes)
HMM Toolkit	POS.	HMM.	Trainable.	C (Unix,DOS)	OpenSource	Active (distrib. lists)	User (no)
Lovin	Stemmer.	Porter stemming.	Trainable.	C or Java	GPL	Finished (none)	User (yes)
Dan Bikel's	Shallow parser.	Stochastic.	Trainable.	Java(independent)	Free for research	Last release 2005 (no)	User (yes)
Stanford NLP SW	NER, POS, Shallow Parsing + dependencies.	Viterbi,Maximum Entropy, PCFG.	Trainable.	Java(independent)	GPL.	Last release in 2006 (distribution list)	User (yes)
TNT	POS.	Viterbi HMM Trigrams.	Trainable.	C (many)	Free for research	Finished (none)	User (influent)
SnowBall	Stemming.	Several.	Yes.	Several.	GPL	Active (distrib. lists)	User (yes)
Xerox Tagger	POS.	Baum-Welch HMM.	Trainable.	LISP(independent)	Sources	Finished in 1992 (no)	User (yes)
YamCha	POS, NEC, Chunk.	Support Vector Machines.	Trainable.	C++ (many)	OpenSource	Finished 2005 (email)	User (many)
SVM Tool	POS.	Support Vector Machines.	Yes, trained model.	C++ (Unix) or Perl (independent)	OpenSource	Active (forum)	Technical ([Giménez and Márquez, 2004])