Procesamiento del Lenguaje Natural, Revista nº 45, septiembre 2010, pp 281-285 recibido 21-06-10 revisado 05-07-10 aceptado 06-07-10

A machine learning method for identifying impersonal constructions and zero pronouns in Spanish*

Un método de aprendizaje automático para la identificación de construcciones impersonales y pronombres cero en español

Luz Rello
Research Group in Computational Linguistics
University of Wolverhampton
luzrello@gmail.com

Pablo Suárez
Dpto. de Física Teórica II
Universidad Complutense de Madrid
psuarez@fis.ucm.es

Ruslan Mitkov
Research Group in Computational Linguistics
University of Wolverhampton
R.Mitkov@wlv.ac.uk

Resumen: En este trabajo se presenta un método basado en aprendizaje automático para la clasificación de la elipsis del sujeto como referencial o no referencial en español. Se trata, tal como se desprende de la revisión bibliográfica realizada, del primer intento de identificar construcciones impersonales no referenciales en esta lengua. Una evaluación del sistema con un corpus de entrenamiento formado por 6.827 verbos anotados ha mostrado que alcanza una exactitud del 87%.
Palabras clave: elipsis de sujeto, construcción impersonal, pronombre zero

Abstract: In this paper, we present a machine learning system for classifying subject ellipsis in Spanish as either referential or non-referential. To the best of our knowledge, this is the first attempt to automatically identify non-referential ellipsis in Spanish. An evaluation of our system against 6,827 finite verbs shows an accuracy of 87%.
Keywords: subject ellipsis, impersonal construction, zero pronoun.

1 Introduction

In zero anaphora resolution, referential subject ellipses (zero pronouns) are identified and subsequently non-referential cases (impersonal constructions) need to be filtered out (Mitkov, 2010). Coreference and anaphora resolution, and in particular zero anaphora resolution, has been found to be crucial in a number of NLP applications including, but not limited to, information extraction (Chinchor and Hirschman, 1997), machine translation (Peral and Ferrández, 2000), text categorization (Yeh and Chen, 2003), and automatic summarization (Steinberger et al., 2007).

Zero anaphora is fairly frequent in Spanish because it is a pro-drop language (Chomsky, 1981). By way of example, out of the 6,827 annotated verbs in our corpus 26% have an omitted subject and 3% present an elliptic but non-referential subject. Automatic identification of referential subject ellipsis in Spanish has been discussed in the computational linguistics literature (Ferrández and Peral, 2000). However, the filtering task of identification of non-referential subject ellipsis in Spanish has not yet been addressed.

A linguistically motivated classification system was developed for referential and non-referential subjects (both explicit and elliptic), and it is detailed in Section 2. Related work on zero pronouns and non-referential expressions identification is outlined in Section 3, while Section 4 discusses the description of the training data, the preprocessing tools needed for its preparation, and the features used in the current machine learning approach. The results are presented in Section 5 and an evaluation of our system is presented in Section 6.

2 Classification

This project aims to deliver a three-fold classification of subjects as (1) explicit and referential (2) elliptic and referential (zero pronouns) and (3) elliptic and non-referential (impersonal constructions).

* We thank Richard Evans for his guidance and support. We are also thankful to Prof. José María Brucart for his wise comments. Many thanks are also due to Molino de Ideas s.a. for providing verb lists.

ISSN 1135-5948

2.1 Explicit subject

This class is composed of verbs whose subject is both explicit (phonetically realized) and belonging to the same clause as the verb occurs. The explicit subject (in italics) can precede or follow the verb. It can be formed not only by a noun phrase but also by an infinitive, an infinitival phrase, an adjectival group or a prepositional group, among others (Real Academia Española, 2009).

(a) *Las fuentes del ordenamiento jurídico español* son la ley, la costumbre y los principios generales del derecho.
The sources of the Spanish legal system are the law, the judicial costume and the general principles of law.

2.2 Zero pronoun

An elliptic subject or zero pronoun is the resultant "gap" (zero anaphor) where zero subject anaphora or ellipsis occurs (Mitkov, 2002). Since zero pronoun are referential, they can be lexically retrieved. The elision of the subject can affect not only the entire noun phrase (b), but also the noun phrase head alone (c). The elision of the head of a noun phrase is only possible when a definite article occurs (Brucart, 1999). Both cases of subject ellipsis are included in the class.

(b) Las leyes no tendrán efecto retroactivo si Ø no dispusieren lo contrario.
The law will not have a retroactive effect unless otherwise *(they)* specify it.

(c) *El Ø* que está obsesionado con ser observado.
The (one) who is obsessed about being observed.

2.3 Impersonal construction

Impersonal constructions with no subjects are non-referential and elliptic. This class is composed of impersonal constructions (d) and impersonal clauses with *se* (e) (Brucart, 1999). The subject cannot be lexically retrieved in either type of clause.

(d) Cuando hay un diagnóstico.
When *(there)* is a diagnosis.

(e) Se podrá hablar de trastorno de la personalidad cuando [...].
(it) will be possible to speak about personality disorder when [...].

The set of impersonal constructions with no subjects (non-referential and elliptic) includes verbs denoting natural phenomena *llover* (to rain); temporal expressions with verbs such as the copulative verb *ser* (to be) and the auxiliary verb *haber* (to have); existential use of verb *haber* (there is, there are); impersonal expressions with locative verbs such as *sobrar* (to be too much of), *bastar* (to be enough) or *faltar* (to have a lack of); pronominal unipersonal verbs such as *tratarse de* (to be about); fixed constructions such as *es decir* or *es que* (that is); *ir para* plus a temporal expression, and the verb *poder* (to be able) as an auxiliary verb; among others (Real Academia Española, 2009).

3 Related Work

Related work on this topic can be classified as (1) literature related to zero pronouns which is mainly related to its identification (Han, 2004), resolution (Okumura and Tamura, 1996) and generation (Peral and Ferrández, 2000) and (2) literature related to the identification non-referential constructions (Evans, 2001).

While in previous work zero pronouns have been addressed in Spanish, no specific studies on the identification of non-referential constructions were found in Spanish, although it has been indicated to be a necessary task (Ferrández and Peral, 2000; Recasens and Hovy, 2009) in anaphora resolution.

Both ruled-based and machine learning approaches for zero pronoun identification and resolution have been carried out in pro-drop languages such as Japanese (Okumura and Tamura, 1996), Chinese (Zhao and Ng, 2007), Korean (Han, 2004) and Spanish (Ferrández and Peral, 2000).

Our work is different to previous research in Spanish in the approach taken. While our method presents a three-fold classification using machine learning techniques, other work in Spanish, both (Ferrández and Peral, 2000) and (Rello and Illisei, 2009) take a rule-based approach for identifying zero pronouns. They present, as a result, a binary classification: verbs with explicit subjects and verbs with elliptic subjects or zero pronouns. In (Ferrández and Peral, 2000) the implementation of a zero pronoun identification and resolution module is part of an anaphora res-

olution system [1] and outperforms the (Rello and Illisei, 2009) approach.

Literature on identifying non-referential expressions is focused on English pleonastic *it* (Evans, 2001) and French expletive pronouns (Danlos, 2005). Machine learning approaches to the identification of pleonastic *it* present better results than rule based as was shown in Boyd *et al.* (2005).

4 Methodology

4.1 Training data

The training data is composed of seventeen texts, originally written in Peninsular Spanish, belonging to two genre: legal[2] and health[3]. The corpus contains 6,827 finite verbs and out of these verbs, 71% have an explicit subject, 26% have a zero pronoun and 3% are non-referential constructions. The texts were first parsed by Connexor's FDG-Parser (Tapanainen and Järvinen, 1997), whose output returns the part of speech and morphological lemma of words in the text, as well as the dependence relations between words.

Next, all finite verbs were extracted and assigned a feature from those listed below. The human annotator was then presented with the sentence in which the verb appears and was asked to classify it. The vectors, along with their manual classification, were written to the training file.

4.2 Features

Fourteen features were proposed in order to classify instances according to the types presented in Section 2. The values for the features were derived from information provided by both the parser and a set of lists.

An additional program was implemented in order to extract the feature values for every instance in the corpus. For example, Connexor's FDG-Parser does not provide any information about the boundaries of clauses within sentences and to this end we developed a purpose-built clause splitter which identifies not only clauses but also their grammatical types (copulative, relative, *etc.*).

Corpus	Tokens	Sentences	Clauses
Legal text 1	9,972	941	600
Legal text 2	1,147	47	56
Legal text 3	17,960	1,035	1,181
Legal text 4	3,578	189	191
Legal text 5	12,456	746	891
Legal text 6	3,962	130	219
Legal text 7	2,159	131	136
Legal text 8	5,219	291	282
Health text 1	2,753	110	270
Health text 2	11,339	658	1,028
Health text 3	1,854	47	140
Health text 4	1,937	84	124
Health text 5	2,183	93	148
Health text 6	1,568	63	210
Health text 7	1,296	69	89
Health text 8	1,687	53	127
Health text 9	12,441	525	1,394
Total	93,511	5,212	7,086

Table 1: Corpus characteristics.

For the purpose of description, it is convenient to describe each of the features as broadly belonging to one of nine classes, detailed below[4].

F1 Presence or absence of subject, as identified by the parser.

F2 Clause type.

F3-5 Morphological information features of the verb (number and person) and lexical information extracted from the parser (the lemma of the finite verb).

F6 Features which take into account the tense of the clause verb and its agreement in person, number and tense with the previous main clause verb and the previous clause verb.

F7-9 Candidates for the subject of the clause: number of noun phrases in the clause before the verb, total number of noun phrases in the clause, and the number of infinitival forms.

F10 The appearance of the particle *se* close to the verb (when *se* occurs immediately before or after the verb or with a maximum of one token lying between the verb and itself).

F11 The appearance of a prepositional phrase with an *a* preposition.

[1] known as the Slot Unification Parser for Anaphora resolution (SUPAR) (Ferrández, Palomar, and Moreno, 1999).

[2] The legal texts are composed of laws from the Penal Code, Civil Code, Spanish Constitution, Law for Universities, Law for Associations, Law for Advertisement, Law for Administration, and Law for Unfair Competition.

[3] The psychiatric scientific papers were compiled from the Spanish on-line Psychiatry Journal. Available at: http://www.psiquiatria.com/

[4] For further explanations see (Rello, 2010).

Class	P	R	F
Subject (non-elliptic and referential)	0.901	0.924	0.913
Zero pronoun (elliptic and referential)	0.774	0.743	0.758
Impersonal construction (elliptic and non-referential)	0.889	0.626	0.734

Accuracy (correctly classified instances): 86.9%

Table 2: Results.

F12-13 The parts of speech of eight tokens: four words prior to and four words after the verb.

F14 Type of verb: a copulative verb, a verb with an impersonal use, a pronominal verb, and its transitivity[5].

5 Evaluation

This memory-based learning method was tested using a leave-one-out cross validation on the training file. The algorithm employed was the K* instance-based learner (Cleary and Trigg, 1995) with a blending parameter of 40%. This algorithm is available from the *WEKA* package (Witten and Frank, 2005; Hall et al., 2009). The results obtained are explained below. The total accuracy (correctly classified instances) was 87%.

Due to the lack of previous work on this topic, a comparison with other methods is not possible. Ferrández and Peral (2000) method identifies zero pronouns performing a higher accuracy (0.98) but lower accuracy when detecting non-omitted subjects (0.80), and it does not recognize impersonal constructions. Also, the corpora used in both approaches are different. As their corpora have been tagged and manually reviewed, their method relies on a recall rate of 100%, and therefore no errors are expected on verb detection. Contrastingly, our method is fully automatic and errors in verb detection can occur because of the performance of the parser.

As a guideline, we give the results offered by Connexor's parser regarding the existence (or not) of a subject inside the clause. Since this parser does not distinguish between referential and non-referential elliptic subjects, both categories have been merged into one. Needless to say, the comparison between these results should be done with caution and they are only presented here as a point of reference. It is clear from the figures that the machine learning method not only improves the f-measure of both elliptic subject classes, but in addition to that also improves f-measure of the non-omitted subject class.

[5]Following the criteria of the Royal Spanish Academy Dictionary (Real Academia Española, 2001). A set of five verb lists were used for this purpose: impersonal verbs, pronominal verbs, potential pronominal verbs, transitive verbs, and intransitive verbs.

Class	P	R	F
Explicit subject (non-elliptic and referential)	0.911	0.716	0.802
Zero pronoun & Impersonal construction (elliptic, referential (or non-referential)	0.543	0.829	0.656

Accuracy (correctly classified instances): 74.9%

Table 3: Connexor's Results.

6 Conclusions

In this paper we have presented a method for the classification of all instances of elliptic and non-elliptic as well as referential and non-referential subjects. The difficulty in detecting non-referential pronouns has been acknowledged since computational resolution of anaphora was first attempted (Bergsma, Lin, and Goebel, 2008). The corpus and the method developed will soon be freely available in internet[6]. Future research goals are improvement of the system by identifying the most significant features and their combinations, as well as parameter optimisation in relation to feature selection.

References

Bergsma, S., D. Lin, and R. Goebel. 2008. Distributional identification of non-referential pronouns. In *Proceedings of the 46th Annual Meeting of the ACL/HLT-08*, pages 10–18.

Brucart, J. M. 1999. La elipsis. In I. Bosque and V. Demonte, editors, *Gramática descriptiva de la lengua española*, volume 2. Espasa-Calpe, Madrid, pages 2787–2863.

Chinchor, N. and L. Hirschman. 1997. MUC-7 Coreference task definition (version 3.0). In *Proceedings of the MUC-97.*

Chomsky, N. 1981. *Lectures on government and binding.* Mouton de Gruyter, Berlin, New York.

[6]http://clg.wlv.ac.uk/resources/index.php

Cleary, J.G. and L.E. Trigg. 1995. K*: an instance-based learner using an entropic distance measure. In *Proceedings of the 12th ICML-95*, pages 108–114.

Danlos, L. 2005. Automatic recognition of French expletive pronoun occurrences. In Robert Dale, Kam-Fai Wong, Jiang Su, and Oi Yee Kwong, editors, *Natural language processing. Proceedings of the 2nd IJCNLP-05*, pages 73–78, Berlin, Heidelberg, New York. Springer. Lecture Notes in Computer Science, Vol. 3651.

Evans, R. 2001. Applying machine learning: toward an automatic classification of *it*. *Literary and Linguistic Computing*, 16(1):45–57.

Ferrández, A., A. Palomar, and L. Moreno. 1999. An empirical approach to Spanish anaphora resolution. *Machine Translation*, 14(3/4):191–216.

Ferrández, A. and J. Peral. 2000. A computational approach to zero-pronouns in Spanish. In *Proceedings of the 38th Annual Meeting of the ACL-2000*, pages 166–172.

Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. 2009. The WEKA data mining software: an update. *SIGKDD Explorations*, 11(1):10–18.

Han, N. 2004. Korean null pronouns: classification and annotation. In *Proceedings of the Workshop on Discourse Annotation. 42nd Annual Meeting of the ACL-04*, pages 33–40.

Mitkov, R. 2002. *Anaphora resolution*. Longman, London.

Mitkov, R. 2010. Discourse processing. In Alexander Clark, Chris Fox, and Shalom Lappin, editors, *The handbook of computational linguistics and natural language processing*. Wiley Blackwell, Oxford, pages 599–629.

Okumura, M. and K. Tamura. 1996. Zero pronoun resolution in Japanese discourse based on centering theory. In *Proceedings of the 16th COLING-96*, pages 871–876.

Peral, J. and A. Ferrández. 2000. Generation of Spanish zero-pronouns into English. In D. N. Christodoulakis, editor, *Natural Language Processing. Proceedings of the 2nd International Conference on NLP-2000*. Springer, Berlin, Heidelberg, New York, pages 252–260. Lecture Notes in Computer Science, Vol. 1835.

Real Academia Española. 2001. *Diccionario de la lengua española*. Espasa-Calpe, Madrid, 22 edition.

Real Academia Española. 2009. *Nueva gramática de la lengua española*. Espasa-Calpe, Madrid.

Recasens, M. and E. Hovy. 2009. A deeper look into features for coreference resolution. In Lalitha Devi Sobha, António Branco, and Ruslan Mitkov, editors, *Anaphora Processing and Applications. Proceedings of the 7th DAARC-09*. Springer, Berlin, Heidelberg, New York, pages 29–42. Lecture Notes in Computer Science, Vol. 5847.

Rello, L. 2010. Elliphant: A machine learning method for identifying subject ellipsis and impersonal constructions in spanish. Master's thesis, University of Wolverhampton, UK.

Rello, L. and I. Illisei. 2009. A rule-based approach to the identification of Spanish zero pronouns. In *Student Research Workshop. RANLP-09*, pages 209–214.

Steinberger, J., M. Poesio, M. A. Kabadjov, and K. Jeek. 2007. Two uses of anaphora resolution in summarization. *Information Processing and Management*, 43(6):1663–1680.

Tapanainen, P. and T. Järvinen. 1997. A non-projective dependency parser. In *Proceedings of the 5th Conference on ANLP-97*, pages 64–71.

Witten, I. H. and E. Frank. 2005. *Data mining: practical machine learning tools and techniques*. Morgan Kaufmann, London, 2 edition.

Yeh, C. and Y. Chen. 2003. Zero anaphora resolution in Chinese with partial parsing based on centering theory. In *Proceedings of the International Conference on NLP-KE-03*, pages 683–688.

Zhao, S. and H.T. Ng. 2007. Identification and resolution of Chinese zero pronouns: a machine learning approach. In *Proceedings of the 2007 Joint Conference on EMNLP/CNLL-07*, pages 541–550.