METHODS AND TOOLS OF COMPUTATIONAL LINGUISTICS FOR THE CLASSIFICATION OF NATURAL NON-REFERENTIAL ELLIPSIS IN SPANISH (REVIEW)

Vera Danilova

Abstract: *This article represents a brief survey of the few works, dedicated to the modern approaches of natural language processing (NLP) to the analysis of impersonal sentences in Spanish. Such an analysis consists in classification of non-referential ellipsis that can be used in machine translation systems. The NLP approaches related with Spanish are mainly based on the work of Rello published in 2010. These approaches do not make use of a proper classification of impersonal models, but of a relative descriptive distribution without strict criteria. The structured classification presented in this article, based on historical and semantic data of interlingual nature, can be also applied for creation of linguistically-motivated classes for machine learning methods. The automatic classification method, employed in the work of Rello, is based on the use of the well-known WEKA package instance-based learner.*

Keywords: impersonal construction, non-referential ellipsis, machine translation

ACM Classification Keywords*: I.2.7 Natural Language Processing - Language models*

Introduction

The analysis of impersonal constructions in modern linguistics implies creation of appropriate classification of models which can be applied to the further detection and extraction of this kind of models in NLP. This can be an important and necessary task for improving the accuracy of machine translation.

The identification of impersonal sentences is closely related to the extraction of different types of subject ellipsis in case of zero anaphora resolution. There are few recently published works which describe the tools created for this kind of extraction in Spanish.

Certain aspects of the machine learning method for detection of non-referential subject in Spanish, described below, are based on the approaches, dedicated to the classification and extraction of pleonastic *it* in English [Evans, 2001; Mitkov et al., 2002; Boyd et al., 2005].

Elliphant and other NLP methods for classification of subject ellipsis in Spanish

The machine learning method Elliphant which is considered to be the first attempt of non-referential ellipsis classification in Spanish was created and described in 2010 by Luz Rello [Rello 2010].

With respect to previous approaches, this one has a number of serious advantages. In the first place, it uses a ternary classification of subjects: explicit, zero pronouns and impersonal constructions. In the previous work by the same author [Rello and Illisei, 2009] a binary classification was used, which included elliptic and non-elliptic variants of subjects, so that it couldn´t permit the further identification of impersonal constructions. Also, the rule-based method was used, which can be applied only to zero pronouns.

Another previous work by Ferrández and Peral [Ferrández and Peral, 2000] represents an algorithm for zero pronoun resolution. By means of their system, the Slot Unification Parser for Anaphora Resolution, anaphoric (the referent of zero pronoun appear in the previous explanation), exophoric (the referent can be found outside the text) and cataphoric (the explanation lies after the verb) zero pronouns´ references can be classified. These authors also used a rule based approach and a binary classification, which implied no identification of non-referential zero pronouns.

According to the Elliphant method, subjects located within the clause, before and after the verb or irrespective ones are included in the explicit class. One of the main advantages of the present system is also the use of an instance-based learner for classification (K* instance-based learner of the WEKA package). All the previous methods, as it has been already mentioned, are rule-based and exclude the possibility of non-referential zero anaphora classification.

Thus, the Elliphant system identifies non-elliptic (explicit) and elliptic (omitted) subjects. The elliptic ones are divided into referential/argumental (zero pronouns) and non-referential/non-argumental (impersonal constructions). Explicit subjects represent the first class in the present study. Zero pronouns (the second class) are defined in this context as subjects that are omitted, but nevertheless understood (e.g. those, which are not phonetically realised in a pro-drop language like Spanish). This referential type is in turn subdivided into specific (*No vendrán* “They won’t come”), unspecific (*Dicen que vendrá* “They say he will come”). In some cases both interpretations are possible, therefore, they were included in the same class. Impersonals sentences (the third class) are classified according to the distribution given in the Grammar of Real Academy of the Spanish language [RAE, 2009]. The omitted subject of an impersonal construction can’t be lexically retrieved. The construction itself can be formed with so-called “impersonal verbs” or it can be a reflexive impersonal clause with the element “*se*”. According to the used classification, there are non-reflexive impersonal clauses (IC), which may contain verbs of meteorological phenomena, *haber* (to be), *hacer* (to do), *ser* (to be), *estar* (to be), *ir* (to go)and *dar* (to give), other verbs, such as *sobrar con* (to be too much), *bastar con* (to be enough), *faltar con* (to have lack of) and pronominal unipersonal verb with subject zero *tratarse de* (to be about). In regard to the reflexive IC, they are considered to have non-specific referents which cannot be retrieved.

Thus, in order to perform the classification, a linguistically motivated classification system was developed for all instances of subject ellipsis. The training data (**ESZIC** (explicit subjects, zero-pronouns and impersonal constructions corpus)) was composed of seventeen legal and health texts written in Peninsular Spanish. It was compiled and an annotation tool was created.

The corpus included 6,825 finite verbs, 71% with an explicit subject, 26% with a zero pronoun and 3% of those verbs forming the impersonal constructions. The parser was based on the Functional Dependency Grammar [Järvinen and Tapanainen, 1998] and it was able to return the POS, the lemma of words in a text and their dependence relations. Every finite verb received a feature and thereupon it was classified by a human annotator. The training file was then provided with vectors, along with the manual classification.

The appropriate fourteen features were selected from the corpus by means of the corresponding tool, developed *ad hoc*, as the parsers used do not provide the information concerning the limits of the clauses within sentences. The extracted features were subsequently used for classification by machine learning method (WEKA package tool). The features belonging to nine classes, defined broadly by the author, are described in detail below. The feature description is taken from [Rello, Suarez, Mitkov, 2010: 283-284]

Table 1: The description of features.

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 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	
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	

Connexor´s Machinese syntax parser was exploited to perform the comparative evaluation of the system. Its results provided an accuracy of 74. 9%. The corresponding table is presented below.

Table 2: The results of Connexor's parser.

Class	Precision	Recall	F-feature
Explicit subject (non-elliptic and referential)	0.991	0.716	0.802
Zero Pronoun and Impersonal Construction (elliptic, referential or non-referential)	0.543	0.829	0.656

The results of the evaluation, performed with K* instance-based learner algorithm [Cleary and Trigg, 1995], are presented below. The estimated accuracy of the method is 86. 9%.

Table 3: The results of K instance-based learner algorithm.*

Class	Precision	Recall	F-feature
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

Hereby, the K* instance-based learner algorithm outperforms Connexor's Machinese syntax parser in the total accuracy and therewith the latest algorithm doesn't distinguish between referential and non-referential constructions.

It should be mentioned, that the author [Rello, 2010: 284] acknowledges the difficulty of automatic non-referential identification and the necessity of the system improvement by, *inter alia*, parameter optimisation related to feature selection.

Linguistic approaches to the classification of impersonal constructions

Linguistic theory is the basis for deriving linguistically motivated classes and the criteria for annotation. In modern linguistics there is no unified approach developed for the classification of impersonal constructions, as the semantic aspect of this issue has many interpretations. The scale, presented in the work of J.C.Moreno [Moreno 1987: 250-280], seems to demonstrate the most adequate criteria of impersonal semantics.

Table 4: The scale of impersonality by J.C.Moreno.

Scale of Impersonality		
External (-animate): *Llueve; Nieva*	Internal (+animate) (-controller): *Me pesa de mis culpas; Me da miedo de confesártelo*	Generic/ non-specific (+animate) (+controller) (-specific): *Se vende pisos;* *Aquí se trabaja bien*
Agentless		Agentive

The main problem of the classification of impersonal models used in the previously examined Elliphant method is that the term “impersonal” aims to be applied for the constructions containing non-referential omitted subjects only, meanwhile the approaches of descriptive and Real Academy Spanish grammars (RAE´s distribution is the basis of the linguistically motivated classes in [Rello 2010]), as those of most European linguists, include in the notion of this term the constructions with animate active unspecific participants (represented by omitted 2Sg, 3Pl and the most part of constructions with “se”-element, depending on the semantics of the verb), which are considered to be referential [Soriano, 1999; Mendikoetxea, 1999; RAE, 2009]. The use of the verb “dar” (to give) and other verbs should be more specified, as there are cases with agreement, which are parallel to those with prepositions, however, they cannot be classified as impersonal due to the presence of the explicit cataphoric subject.

Hereby, the impersonal constructions can be split into agentive (non-specific interpretation, the referent subject is animate) and agentless (no reference, the null element is inanimate and the construction has no controller). Agentless models are considered to be authentic impersonals and can be broadly divided into proper IC (external class in Moreno’s classification), dative IC (internal class) and modal IC (both external and internal classes). Internal class includes dative animate participant, which is considered to represent the experiencer of a situation.

The detailed classification of IC is presented in the table 5.

Table 5: The classification of impersonal structures.*

<table>
<tr><th>IC without “se”-element</th><th>Type of verb/model</th><th>IC containing “se”-element</th><th>Type of verb/model</th></tr>
<tr><td rowspan="3">Proper IC</td><td>• verbs and constructions with estar, ser, hacer, denoting natural phenomena</td><td rowspan="7">Modal IC</td><td>• models, related to necessity or possibility with deber, poder</td></tr>
<tr><td>• hacer (temporal IC)</td><td rowspan="6">• models with verbs denoting osmesis (se sabe, se huele)</td></tr>
<tr><td>• haber (existencial IC)</td></tr>
<tr><td rowspan="3">Dative IC</td><td>• Loc/Dat models with valuation meaning (Loc/Dat+sobrar/bastar/valer+con+Obj, Loc/Dat+sobrar/faltar+de+nada, todo, Loc/Dat+ser/estar+bastante/suficiente+con+ Obj, Loc/Dat+estar bien+con+Obj);
• models related to the sensations of experiencer (Dat+doler/picar/escocer+Loc)
• models related to an occurrence (Loc/Dat + pasar/suceder/ocurrir+de+todo)</td></tr>
<tr><td>• verbs, related to the mental state of the experiencer (Dat+dar/entrar+vergüenza /lástima/miedo/alegría+de+Inf)</td></tr>
<tr><td>• occasional models with dative experiencer or beneficiant (Dat+ir muy bien (mal, estupendamente, fatal, regular)+con+Obj)</td></tr>
<tr><td>Modal IC</td><td>• models of necessity (Hay que+Inf, Hace falta+Inf);
• models with verbs denoting osmesis ((se) huele, (se) sabe, apesta)</td></tr>
</table>

*Further details in [Danilova, 2011]

Conclusion

This paper aims to demonstrate recent approaches to the classification of impersonal constructions. Linguistic theory is meant to be the basis for selecting appropriate classes and features in order to perform the further analysis in NLP, using corresponding tools. As a result of the present review, the further conclusions are made:

- Elliphant is the first and rather effective automatic system for the classification of anaphora resolution. It identifies both explicit and omitted subject and classifies elliptic ones as referential and argumental (zero-pronouns) or non-referential and non-argumental (impersonal constructions) by means of K* instance-based learner (WEKA package);

- The next step is the correct automatic classification of impersonal models. The Elliphant system makes use of the linguistic distribution given in the Real Academy Spanish Grammar [2009], which doesn't represent a proper classification, as it describes the whole variety of models without clearly stated criteria for their distribution. Therewith, this distribution includes in the notion of impersonal construction those models, which have non-specific animate referents, while proper impersonals must be non-referential. Thus, the distribution of the Real Academy Spanish Grammar is not rather appropriate for the formation of linguistically-motivated classes.
- In this article another classification with more strict criteria is presented. It is based on historical and semantic data. Hopefully, it can provide some additional useful information for further development of NLP methods. Hereby, the purpose of the further work in this field is to increase the accuracy of automatic non-referential ellipsis classification by selecting adequate basis for it.

Bibliography

[Boyd et al., 2005] A. Boyd, W. Gegg-Harrison, D. Byron. Identifying non-referential it: a machine learning approach incorporating linguistically motivated patterns. In: Proceedings of the ACL Workshop on Feature Engineering for Machine Learning in Natural Language Processing. 43rd Annual Meeting of the Association for Computational Linguistics (ACL-05), 40-47, 2005.

[Cleary and Trigg, 1995] J.G. Cleary, L.E. Trigg. K*: an instance-based learner using an entropic distance measure. In: Proceedings of the12th ICML-95, pages 108-114, 1995.

[Evans, 2001] R. Evans. Applying machine learning: toward an automatic classification of it. Literary and Linguistic Computing, 2001.

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

[Järvinen and Tapanainen, 1998] T. Järvinen, P. Tapanainen. Towards an implementable dependency grammar. In: A. Polguère & S. Kahane, eds., Proceedings of the Workshop on Processing of Dependency-Based Grammars. 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (ACL/COLING-98), 1-10, 1998.

[Mendikoextea, 1999] A. Mendikoetxea. Construcciones con se. In: Bosque I., Demonte V. Gramática descriptiva de la lengua española. Madrid, vol. 2, 1575-1630, 1999.

[Mitkov et al., 2002] R. Mitkov, R. Evans, C. Orasan. A new, fully automatic version of Mitkov's knowledge-poor pronoun resolution method. In: Proceedings of the 3rd International Conference on Computational Linguistics and Intelligent Text Processing (CICLing-02), 69-83, Springer, Berlin, Heidelberg, New York, Lecture Notes in Computer Science, Vol. 2276, 2002.

[Moreno, 1987] J.C. Moreno. Processes and actions: internal agentless impersonals in some European languages. Berlin, 1987.

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

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

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

[Rello, Suárez, Mitkov, 2010] L. Rello, P. Suárez, R. Mitkov. A machine learning method for identifying impersonal construction and zero pronouns in Spanish. In: Procesamiento del Lenguaje Natural, Revista nº 45, pp 281-285, 2010.

[Soriano, 1999] O. Soriano. Construcciones impersonales no reflejas. In: I. Bosque, V. Demonte. Gramática descriptiva de la lengua española. Madrid, vol. 2, 1723-1779, 1999.

[Danilova, 2011] V. Danilova. The paradigm of impersonal constructions in the Sephardic language in comparison with modern Spanish. Master's thesis, Saint-Petersburg State University, Russia, 2011. (in russian)

Authors' Information

Vera Danilova *– Saint-Petersburg State University, the Department of Romance Languages, master program student; Russia, Saint-Petersburg, Novosibirskaya 18/5-67, 197342; e-mail: maolve@gmail.com*

Major Fields of Scientific Research: Semantics of impersonal structures