Procesamiento del Lenguaje Natural, Revista nº 47 septiembre de 2011, pp 223-230 recibido 03-05-2011 aceptado 23-05-2011

Error Analysis for the Improvement of Subject Ellipsis Detection*

Análisis de Errores para Mejorar la Detección de Elipsis de Sujeto

Luz Rello
NLP and Web Research Groups
Pompeu Fabra University
luz.rello@upf.edu

Gabriela Ferraro, Alicia Burga
NLP Research Group
Pompeu Fabra University
{gabriela.ferraro,alicia.burga}@upf.edu

Resumen: En este trabajo se presenta el análisis de los errores de un método de detección de elipsis de sujeto en español, con el fin de mejorar el sistema en el futuro. El sistema que se evalúa utiliza aprendizaje automático y alcanza una exactitud del 85,3%. El análisis se ha realizado extrayendo de los datos de aprendizaje las instancias que el sistema clasifica erróneamente (1.001), con objeto de establecer una tipología de errores. Cada tipo de error se ha considerado teniendo en cuenta tanto los valores de las características de las instancias como los patrones lingüísticos involucrados. Finalmente, se proponen nuevas características y un conjunto de reglas que puedan aportar una mayor precisión al método.
Palabras clave: elipsis de sujeto, construcción impersonal, pronombre zero, análisis de errores, análisis lingüístico, aprendizaje automático.

Abstract: This paper presents an analysis of the errors of a machine learning method that allow us to propose changes to improve it in future developments. The evaluated system detects Spanish subject ellipsis and yields an accuracy of 85.3%. We extract the erroneously classified instances of our training data (1,001) and classify the errors. We perform an analysis of these instances taking into account the features and the linguistic patterns involved, which motivate the inclusion of new features and rules in the system.
Keywords: subject ellipsis, impersonal construction, zero pronoun, error analysis, linguistic analysis, machine learning.

1 Introduction

A detailed error analysis is a crucial step in the development of natural language processing (NLP) systems. The training of statistical classifiers for NLP tasks requires a careful selection of parameters, as well as a thorough error analysis for their verification and adjustment (Chiarcos and Ritz, 2010). Here, we present the error analysis of a machine learning (ML) system which detects Spanish subject ellipsis[1] by performing a classification of the finite verbs present in a text. We consider as subject ellipsis not only zero pronouns – a missing referential subject– but also non-referential impersonal constructions, where there is no subject. Since the features of the system are linguistically motivated, we find pertinent to perform a linguistic analysis of the erroneously classified instances, to find out which patterns are more difficult to classify and, consequently, to propose new linguistically motivated features or rules to improve the identification of ellipsis.

The high performance of various NLP tasks depends on the identification of ellipsis. Its detection becomes decisive when processing pro-drop languages such as Spanish, since subject ellipsis is a highly recurring phenomenon in these languages (Chomsky, 1981). Subject ellipsis identification is necessary for zero anaphora resolution (Mitkov, 2002), for co-reference resolution

* We thank Ricardo Baeza-Yates and Julio Gonzalo for their wise comments.

[1] Subject ellipsis is the omission of the subject in a sentence.

(Ng and Cardie, 2002) and it has been found to be helpful in a number of NLP applications. These include, but are not limited to, machine translation (Peral and Ferrández, 2000), text categorization (Yeh and Chen, 2003), salience identification (Iida, Kentaro, and Matsumoto, 2009) and parser performance evaluation (Foster, 2010). The difficulty in detecting missing subjects and non-referential pronouns has been acknowledged since the first studies on computational treatment of anaphora (Bergsma, Lin, and Goebel, 2008; Mitkov, 2010). We focus on Spanish where the necessity of this task has been specifically highlighted in (Ferrández and Peral, 2000; Recasens and Hovy, 2009).

Next section explains the classes that the ML system outputs and Section 3 shows the method and its results. Section 4 is devoted to the error analysis. First, the errors are classified in four different types for the analysis of the ML features (Section 4.1) and their linguistic properties are analysed (Section 4.2). Finally, we draw the conclusions in Section 5.

2 Classification

The evaluated system outputs a ternary classification which covers all the elements of the subject position in the clause of a sentence.

Literature related to ellipsis in NLP (Ferrández and Peral, 2000; Evans, 2001; Mitkov, 2010) and linguistic theory (Bosque, 1989; Brucart, 1999; Real Academia Española, 2009) has served as a basis for establishing the three linguistically motivated classes and the annotation criteria of this work.

Each of the verbs in our training set is classified into one of this classes: (a) explicit subjects, (b) referential elliptic subjects (zero pronouns) and (c) non-referential elliptic subjects (impersonal constructions). These three classes cover all instances of the training data in subject position.

(a) Explicit subjects: non-elliptic and referential. This class is composed of verbs whose subject is both explicit and belonging to the same clause as the verb occurs. 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) *Los Jueces y Tribunales* establecerán sus resoluciones.
> *The judges and the courts* will establish their resolutions.

(b) Zero pronouns: elliptic and referential. An elliptic subject or zero pronoun is the resultant "gap" where zero subject anaphora or ellipsis occurs (Mitkov, 2002). Since zero pronoun are referential, they can be lexically retrieved.

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

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

3 Machine Learning Algorithm

The training data used in the learning process of the tool was obtained from the Explicit Subjects, Zero-Pronouns and Impersonal Constructions (ESZIC) corpus created specifically for this purpose[2]. The corpus is composed of texts from legal and health genre originally written in Spanish. It was parsed by Connexor's Machinese Syntax (Connexor Oy, 2006), which returns information on the part-of-speech (POS), morphological lemma of words in a text and the dependency relations between words. The instances of the corpus were manually annotated by three people presenting an overall Fleiss kappa (Fleiss, 1971) inter-annotator agreement of 0.90 (Rello, Baeza-Yates, and Mitkov, 2011). The training data is composed of 6,827 instances. Each instance corresponds to one finite verb extracted from the corpus and they are composed by 14 linguistically motivated features (see Section 4.1) derived from the

[2] Available at: http://www.luzrello.com/Projects_files/elliphant_eszic_es_corpus.zip.

corpus. There is a training set but no explicit test set, since we use cross-validation instead.

We use TiMBL, the Tilburg memory-based learning classifier (Daelemans and Bosch, 2005), which is a descendant of the k-nearest neighbor approach. To optimize TiMBL we chose the Inverse Linear Distance (Dudani, 1976) as the class voting weights for extrapolation of the 10 nearest neighbors (that is, k=10). Using leave-one-out evaluation, the overall accuracy is 85.3%: 5,825 out of the 6,827 instances are correctly classified (see Table 1).

Using the K* algorithm (Cleary and Trigg, 1995) from Weka package (Witten and Frank, 2005) the accuracy reaches 86.7%. However, given that Weka does not provide the error instances, we use TiMBL to perform the error analysis.

Class	P	R	F
Zero Pronoun	73.5%	73.8%	73.7%
Explicit Subject	89.7%	90.8%	90.3%
Impersonal	82.4%	52.5%	64.2%

Table 1: Precision, Recall and F-Measure.

4 *Error Analysis*

For each of the selected errors, we extract the clauses where they appear in the corpus from which the training data was generated. In Table 2 we show the distribution of the true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN) per class. Out of all the instances, 71% of the verbs (4,855 occurrences) have an explicit subject, 26% (1,793) have a zero pronoun and 3% of the verbs (179) take part in impersonal constructions.

Class	TP	FP	TN	FN
Zero Pronoun	1323	476	4558	470
Explicit Subject	4409	505	1467	446
Impersonal	94	20	6628	85

Table 2: Classification's TP, FP, TN and FN.

The classification of impersonal constructions is less balanced than the ones for explicit subjects and zero pronouns. If the number of true positives of impersonal constructions were smaller, the overall recall would considerably increase. In the confusion matrix shown in Table 3 we can observe the distribution of the errors per class.

Class	Zero Pronoun	Explicit Subject	Impersonal
Zero Pronoun	1323	459 **(c)**	11
Explicit Subject	437 **(d)**	4409	9
Impersonal	39 **(a)**	46	94 **(b)**

Table 3: Confusion Matrix.

We do not consider the cases of zero pronouns classified as impersonal constructions and the cases of explicit subjects classified as impersonal constructions because there are too few of them. The different kind of errors taken into account are classified into the following four classes:

(a) Impersonal constructions classified as zero pronouns (39 errors).

(b) Impersonal constructions classified as explicit subjects (46 errors).

(c) Explicit Subjects classified as zero pronouns (437 errors).

(d) Zero pronouns classified as explicit subjects (459 errors).

We focus on the analysis of all the errors of classes (a) and (b) because they belong to the most unbalanced class and a refinement in their classification would mean a significant improvement in the performance of the system. We also observe that classes of errors (c) and (d) are symmetric because the zero pronouns that are classified as explicit subjects are similar in number to the explicit subjects that are classified as zero pronouns. For the analysis of the classes of errors (c) and (d) we took a sample of both groups composed of instances which share similar feature values.

In the analysis, we first study the relationship between errors and the patterns observed in their feature values (See Section 4.1). The exploration of the features allows us to generate smaller samples of the groups of errors (c) and (d) for a further linguistic analysis. Then, we explore the linguistic characteristics of the instances (See Section

Feature	Definition	Value
1 PARSER	Parsed subject	True, False
2 CLAUSE	Clause type	Main, Rel, Imp, Prop, Punct
3 LEMMA	Verb lemma	Parser's lemma tag
4 NUMBER	Verb morphological number	SG, PL
5 PERSON	Verb morphological person	P1, P2, P3
6 AGREE	Agreement in person, number, tense and mood	FTFF, TTTT, FFFF, TFTF, TTFF, FTFT, FFFT, TTTF, FFTF, TFFT, FFTT, FTTT, FTTF, TFFF, TFTT, TTFT
7 NHPREV	Previous noun phrases	Number of noun phrases previous to the verb
8 NHTOT	Total noun phrases	Number of noun phrases in the clause
9 INF	Infinitive	Number of infinitives in the clause
10 SE	Spanish particle *se*	True, False
11 A	Spanish preposition *a*	True, False
12 POS$_{pre}$	Four parts of the speech previous to the verb	292 different values combining the parser's POS tags,*i.e.:* @HN, @CC, @MAIN, etc.
13 POS$_{pos}$	Four parts of the speech following the verb	280 different values combining the parser's POS tags,*i.e.:* @HN, @CC, @MAIN, etc.
14 VERB$_{type}$	Type of verb: copulative, impersonal pronominal, transitive and intransitive	CIPX, XIXX, XXXT, XXPX, XXXI, CIXX, XIPT, XXXX, XIXI, CXPI, XXPI, XIPI XXPT, XIPX, XXEX

Table 4: Features: definitions and values.

4.2) by examining the clause in which the instance appears in our corpus from which our training data was generated.

4.1 Machine Learning Features

In this section we see the interaction of the features values and the incorrectly classified instances. Each instance is composed by 14 linguistically motivated features shown in Table 4 (Rello, Baeza-Yates, and Mitkov, 2011).

TiMBL assigns a weight to each feature determining their relevance in the classification applying a GainRatio measure. By using the three most relevant features (PERSON, NHPREV, NHTOT) the overall accuracy is 71.1% and the confusion matrix shows that all the instances are classified as subjects. Then, if we use the set of features considered as the second best set (PARSER, LEMMA, NHTOT, POS$_{pre}$, POS$_{pos}$) the accuracy yields a 74.9% and no instance is classified as an impersonal construction (See Table 5).

An ablation study shows that each of the features does not have a meaningful contribution by itself, while the interaction between them is the most determining factor for the classification.

Class	**Zero Pronoun**	**Explicit Subject**	**Impersonal**
Zero Pronoun	529	1264	0
Explicit Subject	265	4590	0
Impersonal	67	112	0

Table 5: Confusion Matrix: Second Best Feature Set.

4.1.1 Impersonals as Zero Pronouns and Explicit Subjects

Here, we explore the feature patterns found in the impersonal constructions which were incorrectly classified as zero pronouns and explicit subjects, groups (a) and (b). The 39 cases of impersonal constructions classified as zero pronouns form the most homogeneous group of the four error types. All of these instances are verbs conjugated in third person singular (NUMBER and PERSON) and have no noun phrases before the verb in the clause (NHPREV). The distribution of the instances in terms of the clause type (CLAUSE) is regular. On the other hand, the set of impersonal constructions classified as explicit subjects is very heterogenous and no striking trends were observed. However, noun phrases in the clause (NHPREV, NHTOT) were scarce.

Feature LEMMA	Group (a)	Group (b)	Training data
tratar	**13.04%**	**5.12%**	0.77%
haber	**15.21%**	**7.69%**	1.61%
poder	8.69%	**17.94%**	8.73%
ser	13.04%	17.94%	14.07%

Table 6: Percentages of LEMMA values.

It is worth noticing that in both types of errors there are some verbs lemmas which are wrongly classified and their presence in the errors is much higher than in the training data. The verb *haber ('to have, there is/are')* appears in the errors an average of 7 times more than in the training data, the verb *tratar ('to be about', 'to deal with')* appears 12 times more. Although the presence of verbs *ser ('to be')* and *poder ('to can')* is quite frequent in the errors, these verbs have a similar frequency in our training data (see Table 6).

4.1.2 Explicit Subjects as Zero Pronouns and *vice versa*

We explore the feature patterns in all the instances of the groups of errors (c) and (d). There are 172 different values patterns in group (c) and 187 in group (d). More than a half of the errors present unique patterns of features while we distinguish 8 feature patterns which include 150 instances, as shown in Tables 7 and 8. These patterns enlighten the interaction of the features. For instance, the features which take into account the number of noun phrases per clause (NHPREV and NHTOT) are highly ranked by the system. When an instance has no noun phrases in the clause, the system tends to classify it as a zero pronoun and, when there are noun phrases, the preference is explicit subject. Also, we observe the relevance of the clause type where the instance is found, relative clauses for zero pronouns (CLAUSE = REL) and clause starting with an improper conjunction (CLAUSE = IMP).

4.2 Linguistic Analysis

The feature analysis serves as an starting point for a more refined linguistic analysis of the errors. Since we found a great variety of different patterns in groups (c) and (d), for the linguistic analysis of these groups we only take into account the instances belonging to the most frequent vector patterns, that is, patterns 1 and 2 in Table 7 and patterns 1, 2 and 3 in Table 8. We mention only the linguistic characteristics in the errors which are different from the general trends observed in the corpus.

4.2.1 Impersonal Constructions Classified as Zero Pronouns

This set is composed of sentences of diverse length, although it is noticeable the presence of very long sentences (more than 30 tokens). It is frequent the presence of a post-verbal complement introduced by a preposition which is not necessary "a" (the preposition taken into account in the features). In the health texts, preverbal clitics are included in those constructions *i. e. no se le dice, ('he is not told')*.

4.2.2 Impersonal Constructions Classified as Explicit Subjects

When the system points out as subject zero, there is much more flexibility in tenses than when the system detects explicit subject. There is no regularity considering the position within the sentence, or the presence and size of a post-verbal complement. Even though there are few cases where the verb is indicative present, most times the verb is either future or subjunctive. When the post-verbal complement is plural, negation precedes the verb. If the verb is in the main clause, the most common pattern previous to the verb is composed by adverbial clauses introduced by prepositions. In both groups (a) and (b) when the system detects a zero subject with the verb *ser ('to be')*, it always refers to the idiomatic form *es decir ('that is')*.

4.2.3 Explicit Subjects Classified as Zero Pronouns

The three analyzed patterns present homogeneous linguistics features (see Table 8). In pattern 1, instances incorrectly classified are those where verbs appear in the main clause, in which the subject is followed by a comma. Concessive, adverbial and relative clauses also present difficulties on the task of subject identification. In most of the cases, subjects tend to be post-verbal and the verb tends to be in subjunctive mode.

The second pattern presents also post-verbal subjects and most of the cases occur within embedded clauses. Most of the cases

Pattern	No. Instances	PARSER	CLAUSE	NUMBER	PERSON	NHPREV	NHTOT	INF	SE	A
1	25	True	REL	SG	P3	1	1	0	false	false
2	22	False	REL	SG	P3	1	1	0	false	false
3	19	True	REL	PL	P3	1	1	0	false	false
4	18	False	IMP	SG	P3	0	1	0	false	false

Table 7: Feature Values Patterns of Zero Pronouns Classified as Explicit Subjects.

Pattern	No. Instances	PARSER	CLAUSE	NUMBER	PERSON	NHPREV	NHTOT	INF	SE	A
1	20	False	IMP	SG	P3	0	0	0	false	false
2	18	False	IMP	PL	P3	0	0	0	false	false
3	16	True	IMP	PL	P3	0	0	0	false	false
4	12	False	Main	SG	P3	0	0	0	false	false

Table 8: Feature Values Patterns of Explicit Subjects Classified as Zero Pronouns.

use subjunctive tense. Complex subjects[3] and topicalized subjects (d) are more difficult to be detected :

(d) *Los antisociales*, más que valientes, son temerarios.
Antisocial people, instead of brave, are reckless.

The errors from pattern 3 mainly occur in embedded clauses, in long clauses where the head noun is an indefinite pronoun or the subject is far from its head (sometimes the head of the subject is ten locations far from the head verb). There are less cases of subjunctive. There are mistakes in detecting complex subjects and subjects containing precise words such as, *i.e. ambos ('both') or todo ('all')*.

4.2.4 Zero Pronouns Classified as Explicit Subjects

In patterns 1 and 2, post-verbal objects are commonly introduced by a preposition, *i.e. a ('to'), de ('of')*, *etc.*, or a conjunction, *i.e. que ('that'), como ('as'), etc.* These objects and the verb are in third person singular. Pattern 1 in this group did not present defined tendencies. There were also found some errors from the previous annotation process. Most cases are presented in embedded cases and most of them are within relative clauses. All instances belonging to Pattern 2 occur in embedded clauses of different sizes. Subjunctive mode is more frequent in this class than in the ESZIC corpus. The antecedent of the zero pronoun tends to appear as a prepositional complement or as the subject of another verb. Maybe this explains why they were classified as explicit subjects by the system. The relative pronoun in relative clauses cases is not the subject but a complement of the verb. There are some idiomatic expressions involving verbs, *i.e. lo que sea ('whichever it is'), pase lo que pase ('whatever it happens'), etc.*

5 *Conclusions and Future Work*

Although we expected to find semantic and pragmatic characteristics shared in the errors, we only found grammatical information. However, this information is enough for improving the system.

From the analysis presented we propose a set of features for future work which takes into account the following linguistic characteristics:

– Post-verbal prepositions.

– Auxiliary verbs.

– Verbal tense (future tenses).

– Verbal mode (subjunctive).

– Clause length.

– Punctuation marks appearing before the verb and the preceding noun phrases.

– Concessive and adverbial subordinate clauses.

– Negation.

We also propose to enrich the ML system with rules and lists:

[3] Subjects which include a conjunction of singular elements, either directly through commas and/or a coordinative conjunction or through combinations, such as *i. e., tanto... come ('as... as')*.

- List of idioms which include verbs with impersonal uses, such as *es decir ('that is to say')* and *es que ('that is')*; *ir para ('go for')* plus a temporal expression; the pronominal unipersonal verbs such as *tratarse de ('to be about')*; impersonal expressions with locative verbs such as *sobrar ('to be too much of')*, *bastar ('to be enough')* or *faltar ('to have a lack of')*.

- List of words which can be subject on their own, *i.e. ambos ('both'), todo ('all'), etc.*

- Rules for specific verbs *i.e. poder ('can'), ser ('to be'), haber ('to have'), etc.*

- Rules for verbs which tend to have post-verbal subjects *i.e. gustar ('to like'), preocupar ('to be worried'), importar ('to matter'), doler ('to hurt'), quedar ('to be left'), etc.*

- Rules for detecting complex noun phrases.

For instance, we tried so far to add one new feature (SPECIFICVERB), which indicates the presence of verbs *haber* and *tratar* and the system accuracy rises from 85,3% to 85,4%.

For further work we also consider the possibility of applying the features in cascade with different ML algorithms. In that case, the features which deal with noun phrases would be used first and the features that take into account the lemma and morphology of the verb would be applied later.

References

Bergsma, S., D. Lin, and R. Goebel. 2008. Distributional identification of non-referential pronouns. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL/HLT-08)*, pages 10–18.

Bosque, I. 1989. Clases de sujetos tácitos. In Julio Borrego Nieto, editor, *Philologica: homenaje a Antonio Llorente*, volume 2. Servicio de Publicaciones, Universidad Pontificia de Salamanca, Salamanca, pages 91–112.

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.

Chiarcos, D. and J. Ritz. 2010. Qualitative and quantitative error analysis in context. In *Proceedings of the 10th Conference on Natural Language Processing (KONVENS-2010).*

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

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

Connexor Oy, 2006. *Machinese language model.*

Daelemans, W. and A. V. Bosch. 2005. *Memory-Based Language Processing.* Cambridge University Press, Cambridge.

Dudani, S.A. 1976. The distance-weighted k-nearest neighbor rule. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-6:325–327.

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

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

Fleiss, J. L. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Foster, Jennifer. 2010. "cba to check the spelling": Investigating parser performance on discussion forum posts. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 381–384.

Iida, R., I. Kentaro, and Y. Matsumoto. 2009. Capturing salience with a trainable cache model for zero-anaphora resolution. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Conference on Natural Language Processing of the Asian*

Federation of Natural Language Processing (ACL/AFNLP-09), pages 647–655.

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.

Ng, V. and C. Cardie. 2002. Identifying anaphoric and non-anaphoric noun phrases to improve coreference resolution. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING-02)*, pages 1–7.

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

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 Discourse Anaphora and Anaphor Resolution Colloquium (DAARC-09).* Springer, Berlin, Heidelberg, New York, pages 29–42. Lecture Notes in Computer Science, Vol. 5847.

Rello, L., R. Baeza-Yates, and R. Mitkov. 2011. Improved subject ellipsis detection in Spanish. *submitted.*

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. Using zero anaphora resolution to improve text categorization. In *Proceedings of the 17th Pacific Asia Conference on Language, Information and Computation (PACLIC-03)*, pages 423–430.