Procesamiento del Lenguaje Natural, Revista nº 45, septiembre 2010, pp 291-294 recibido 21-06-10 revisado 05-07-10 aceptado 06-07-10

Determination of Features for a Machine Learning Approach to Pronominal Anaphora Resolution in Basque*

Determinación de características en una aproximación basada en el aprendizaje automático para la resolución de anáforas pronominales en euskara

O.Arregi, K.Ceberio, A.Díaz de Illarraza, I.Goenaga, B.Sierra, A.Zelaia
University of the Basque Country
Manuel Lardizabal pasealekua 1, 20018 Donostia-San Sebastián
olatz.arregi@ehu.es, jipdisaa@si.ehu.es, b.sierra@ehu.es, ana.zelaia@ehu.es

Resumen: En este trabajo presentamos una primera aproximación basada en el aprendizaje automático para resolver la anáfora pronominal en euskara. Asímismo, determinamos las características más relevantes para esta tarea.
Palabras clave: Resolución de anáfora, aprendizaje automático

Abstract: In this paper we present the preliminaries for a machine learning approach to resolve the pronominal anaphora in Basque language. In this work we determine the appropriate features to be used in this task.
Keywords: Anaphora resolution, machine learning

1. Introduction

Pronominal anaphora resolution is related to the task of identifying noun phrases that refer to the same entity mentioned in a document.

Anaphora resolution is crucial in real-world natural language processing applications e.g. machine translation or information extraction. Although it has been a wide-open research field in the area since 1970, the work presented in this article is the first dealing with the subject for Basque, especially in the task of determining anaphoric relationship using a machine learning approach.

The first problem to carry out is the lack of a big annotated corpus in Basque. Mitkov in [5] highlights the importance of an annotated corpus for research purposes: *The annotation of corpora is an indispensable, albeit time-consuming, preliminary to anaphora resolution (and to most NLP tasks or applications), since the data they provide are critical to the development, optimization and evaluation of new approaches.*

Recently, an annotated corpus has been published in Basque with pronominal anaphora tags [2] and thanks to that, this work could be managed.

Although the literature about anaphora resolution with machine learning approaches is very large, we will concentrate on those references directly linked to the work done here. In [10] they apply a noun phrase (NP) coreference system based on decision trees to MUC6 and MUC7 data sets. It is usually used as a baseline in the coreference resolution literature.

The state of the art of other languages varies considerably. In [8] they propose a rule-based system for anaphora resolution in Czech. In [11] the author uses a system based on a loglinear statistical model to resolve noun phrase coreference in German texts. On the other hand, [6] and [7] present an approach to Persian pronoun resolution based on machine learning techniques. They developed a corpus with 2,006 labeled pronouns.

2. Selection of Features

Basque is not an Indo-European language and differs considerably in grammar from languages spoken in other regions around. It is an agglutinative language, in which grammatical relations between components within a clause are represented by suffixes. This is a distinguishing characteristic since morphological information of words is richer than in the surrounding languages. Given that Basque is a head final language at the syntactic level, the morphological information of the phrase, which is considered to be the head,

* This work was supported by KNOW2 (TIN2009-14715-C04-01) and Berbatek (IE09-262) projects.

ISSN 1135-5948

is in the attached suffix. That is why morphosyntactic analysis is essential.

In this work we specifically focus on the pronominal anaphora; concretely, the demonstrative determiners when they behave as pronouns. In Basque there are not different forms for third person pronouns and demonstrative determiners are used as third person pronominals. There are three degrees of demonstratives that are closely related to the distance of the referent: hau (this/he/she/it), hori (that/he/she/it), hura (that/he/she/it). As we will see in the example of Section 2.2 demostratives in Basque do not allow to infer whether the referent is a person (he, she) or it is an impersonal one (it).

Moreover, there is no gender distinction in the Basque morphological system; the gender is not a valid feature to detect the antecedent of a pronominal anaphora.

2.1. Determination of Feature Vectors

In order to use a machine learning method, a suitable annotated corpus is needed. We use part of the Eus3LB Corpus[1] which contains approximately 50.000 words from journalistic texts previously parsed. It contains 349 annotated pronominal anaphora.

In this work, we first focus on features obtainable with our linguistic processing system proposed in [1]. We can not use some of the common features used by most systems due to linguistic differences (i.e. gender). Nevertheless, we use some specific features that linguistic researchers consider important for this task.

The features are grouped in three categories: features of the anaphoric pronoun, features of the antecedent candidate, and features that describe the relationship between both.

- Features of the anaphoric pronoun

f_1 - *dec_ana*: Declension case of anaphor.
f_2 - *sf_ana*: Syntactic function of anaphor.
f_3 - *phrase_ana*: Whether the anaphor has the phrase tag or not.
f_4 - *num_ana*: Number of anaphor.

- Features of the antecedent candidate

f_5 - *word*: Word of antecedent.
f_6 - *lemma*: Lemma of antecedent.
f_7 - *cat_np*: Syntactic category of NP.
f_8 - *dec_np*: Declension case of NP.
f_9 - *num_np*: Number of NP.
f_{10} - *degree*: Degree of the NP that contains a comparative.
f_{11} - *np*: Whether the noun phrase is a simple NP or a composed NP.
f_{12} - *sf_np*: Syntactic function of NP.
f_{13} - *enti_np*: Type of entity.

- Relational features

f_{14} - *dist*: The distance between the anaphor and the antecedent candidate in terms of number of Noun Phrases.
f_{15} - *same_sent*: If the anaphor and the antecedent candidate are in the same sentence.
f_{16} - *same_num*: Besides to singular and plural numbers, there is another one in Basque: the indefinite. Thus, this feature has more than two possible values.

In summary we would like to remark that we include morphosyntactic information in our pronoun features such as the syntactic function it accomplishes, the kind of phrase it is, and its number. We also include the pronoun declension case. We use the same features for the antecedent candidate and we add the syntactic category and the degree of the noun phrase that contains a comparative. We also include information about name entities indicating the type (person, location and organization). The word and lemma of the noun phrase are also taken into account. The set of relational features includes three features: the distance between the anaphor and the antecedent candidate, a Boolean feature that shows whether they are in the same sentence or not, and the number agreement between them.

2.2. Generation of Training Instances

The method we use to create training instances is similar to the one explained in [10]. Positive instances are created for each annotated anaphor and its antecedent. Negative instances are created by pairing each annotated anaphor with each of its preceding noun phrases that are between the anaphor and the antecedent. When the antecedent candidate is composed, we use the information of the last word of the noun phrase to create the features due to the fact that in Basque this word is the one that contains the morphosyntactic information.

[1] Eus3LB is part of the 3LB project [9]

In order to clarify the results of our system, we introduce the following example:

Ben Amor *ere ez da Mundiala amaitu arte etorriko Irunera*, **honek** *ere Tunisiarekin parte hartuko baitu Mundialean.*

(**Ben Amor** *is not coming to Irun before the world championship is finished, since* **he** *will play with Tunisia in the World Championship*).

The word *honek* (he) in bold is the anaphor and *Ben Amor* its antecedent. The noun phrases between them are *Mundiala* and *Irunera*. The next table shows the generation of training instances from the example.

Antecedent	Anaphor	Positive
Ben Amor	honek (he/it)	1
Mundiala	honek (he/it)	0
Irunera	honek (he/it)	0

Generating the training instances in that way, we obtained a corpus with 968 instances; 349 of them are positive, and the rest, 619, negatives.

3. Evaluation

In order to evaluate the performance of our system, we use the above mentioned corpus. Due to the size of the corpus, a 10 fold cross-validation is performed. It is worth to say that we are trying to increase the size of the corpus.

We consider different machine learning paradigms from Weka toolkit [3] in order to find the best system for the task. The classifiers used are: SVM (polynomial kernel), Multilayer Perceptron, Naïve Bayes (NB), k-NN ($k = 1$), Random Forest (RF), NB-Tree and Voting Feature Intervals (VFI). We tried some other traditional methods like rules or simple decision trees, but they do not report good results for our corpus.

Table 1. shows the results obtained with these classifiers.

	Precision	Recall	F-measure
VFI	0,653	0,673	0,663
Perceptron	0,692	0,682	**0,687**
RF	0,666	0,702	0,683
SVM	0,803	0,539	0,645
NB-tree	0,771	0,559	0,648
NB	0,737	0,587	0,654
k-NN	0,652	0,616	0,633

Cuadro 1: Results of different algorithms

The best result is obtained by using the Multilayer Perceptron algorithm, F-measure 68.7 %. In general, precision obtained is higher than recall. The best precision is obtained with SVM (80.3 %), followed by NB-tree (77.1 %). In both cases, the recall is similar, 53.9 % and 55.9 %.

These results are not directly comparable with those obtained for other languages such as English, but we think they are a good baseline for Basque language. We must emphasize that only the pronominal anaphora is treated here, so actual comparisons are difficult.

4. Contribution of Features Used

To better understand which of the features used are more efficient, we evaluate the weight of attributes by different measurements: Information Gain, Relief algorithm, Symmetrical Uncertainty, Chi Squared statistic, and Gain Ratio. The order of features derive from each of the measurements is quite similar in all cases except for the Relief algorithm [4]. Although the first four features are the same in all cases (with slight order variations), the Relief algorithm shows a different order beyond the fifth feature, giving more weight to *word* or *lemma* features than to others relating to anaphor.

Fig. 1. shows the weights of these features taking into account all the measurements used.

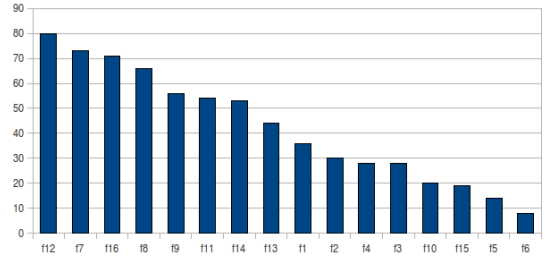

Figura 1: The average weight of features

As expected, the features *word* and *lemma* do not contribute much to the classification process, and we can say that, in general, features relating to the anaphor are not very important for this task, while relational features like *same_num* (agreement in number) or *dist* (distance) appeared to be important. Moreover, all measurements show that features corresponding to the noun phrase are meaningful for this task, as indicated by other authors ([8], [10]).

5. Conclusions and Future Work

This is the first study carried out on resolution of pronominal anaphora in Basque using a machine learning approach. It has been a useful start in defining criteria for anaphora resolution. The results obtained from this work will be helpful for the development of a better anaphora resolution tool for Basque.

We consider seven machine learning algorithms for our first approach in order to decide which kind of method can be the best for this task. The best results are obtained with two classifiers (Random Forest and VFI) which are not the most used for this task in other languages. This may be due to the chosen feature set, the noise of the corpus, and the Basque language characteristics. Traditional methods like SVM, give us a good precision but an F-measure four points below the best system. Anyway, the corpus used in this work is quite small, so we think that the results we obtain can be improved with a larger corpus.

The combination of classifiers has been intensively studied with the aim of improving the accuracy of individual components. We intend to apply a multiclassifier based approach to this task and combine the predictions generated applying a Bayesian voting scheme.

We plan to expand our approach to other types of anaphoric relations with the aim of generating a system to determine the coreference chains for a document.

Finally, the interest of a modular tool to develop coreference applications is unquestionable. Every day more people research in the area of the NLP for Basque and a tool of this kind can be very helpful.

Referencias

[1] Aduriz, I., Aranzabe, M. J., Arriola, J.M., Díaz de Ilarraza, A., Gojenola, K., Oronoz, M., Uria, L.: A Cascaded Syntactic Analyser for Basque. CICLing 2004. Seoul, Korea (2004)

[2] Aduriz, I., Aranzabe, M. J., Arriola, J.M., Atutxa, A., Díaz de Ilarraza, A., Ezeiza, N., Gojenola, K., Oronoz, M., Soroa, A., and Urizar, R.: Methodology and steps towards the construction of EPEC, a corpus of written Basque tagged at morphological and syntactic levels for the automatic processing. Language and Computers, Corpus Linguistics Around the World. Edited by Andrew Wilson, Dawn Archer, Paul Rayson, pp. 1 – 15(15). Rodopi, Netherlands (2006)

[3] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I. H.: The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1 (2009)

[4] Kira, K., Rendell, L. A.: A Practical Approach to Feature Selection. Ninth International Workshop on Machine Learning, pp. 249 – 256, (1992)

[5] Mitkov, R.: Anaphora resolution. London: Longman, (2002)

[6] Moosavi, N. S., and Ghassem-Sani, G.: Using Machine Learning Approaches for Persian Pronoun Resolution. Workshop on Corpus-Based Approaches to Coreference Resolution in Romance Languages. CBA-08, (2008)

[7] Moosavi, N. S., and Ghassem-Sani, G.: A Ranking Approach to Persian Pronoun Resolution. Advances in Computational Linguistics. Research in Computing Science 41, pp. 169 – 180, (2009)

[8] Nguy and Zabokrtský.: Rule-based Approach to Pronominal Anaphora Resolution Method Using the Prague Dependency Treebank 2.0 Data. . Proceedings of DAARC 2007 (6th Discourse Anaphora and Anaphor Resolution Colloquium), (2007)

[9] Palomar, M., Civit, M., Díaz, A., Moreno, L., Bisbal, E., Aranzabe, M. J., Ageno, A., Martí, M.A. and Navarro, B.: 3LB: Construcción de una base de datos de árboles sintáctico-semánticos para el catalán, euskera y español. XX. Congreso SEPLN, Barcelona, (2004)

[10] Soon, W. M., Ng, H. T., and Lim, D. C. Y.: A Machine Learning Approach to Coreference Resolution of Noun Phrases. Computational Linguistics, 27(4):521 – 544, (2001)

[11] Versley, Y.: A Constraint-based Approach to Noum Phrase Coreference Resolution in German Newspaper Text. In Konferenz zur Verarbeitung Natürlicher Sprache KONVENS, (2006)