A Combination of Classifiers for the Pronominal Anaphora Resolution in Basque

Ana Zelaia Jauregi, Basilio Sierra, Olatz Arregi Uriarte, Klara Ceberio, Arantza Díaz de Illarraza, and Iakes Goenaga

University of the Basque Country
ana.zelaia@ehu.es

Abstract. In this paper we present a machine learning approach to resolve the pronominal anaphora in Basque language. We consider different classifiers in order to find the system that fits best to the characteristics of the language under examination. We apply the combination of classifiers which improves results obtained with single classifiers. The main contribution of the paper is the use of bagging having as base classifier a non-soft one for the anaphora resolution in Basque.

1 Introduction

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

According to [5], *anaphora, in discourse, is a device for making an abbreviated reference (containing fewer bits of disambiguating information, rather than being lexically or phonetically shorter) to some entity (or entities).*

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.

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. Combination methods have been recently applied to coreference resolution problems. In [11] the authors use bagging and boosting techniques in order to improve single classifiers results.

The state of the art of other languages varies considerably. In [8] they propose a rule-based system for anaphora resolution in Czech. They use the Treebank data, which contains more than 45,000 coreference links in almost 50,000 manually annotated Czech sentences. In [12] the author uses a system based on a loglinear statistical model to resolve noun phrase coreference in German texts.

I. Bloch and R.M. Cesar, Jr. (Eds.): CIARP 2010, LNCS 6419, pp. 253–260, 2010.

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.

The paper we present describes a baseline framework for Basque pronominal anaphora resolution using a machine learning approach. In Section 2 some general characteristics of Basque pronominal anaphora are explained. Section 3 shows the results obtained for different machine learning methods. The combination of classifiers is presented in Section 4, and finally, in Section 5, we present some conclusions and point out future work lines.

2 Pronominal Anaphora Resolution in Basque

2.1 Main Characteristics of Pronominal Anaphora in Basque

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 (number, case, etc.), which is considered to be the head, 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 2.3 demostratives in Basque do not allow to infer whether the referent is a person (he, she) or it is an impersonal one (it).

Moreover, demostrative determiners do not have any gender in Basque. Hence, the gender is not a valid feature to detect the antecedent of a pronominal anaphora because there is no gender distinction in the Basque morphological system.

2.2 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 the linguistic processing system proposed in [1]. We can not use some of the common features used by

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

most systems due to linguistic differences. For example the gender, as we previously said. Nevertheless, we use some specific features that linguistic researchers consider important for this task.

The features used 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*: The declension case of the anaphor.
  - f_2 - *sf_ana*: The syntactic function of the anaphor.
  - f_3 - *phrase_ana*: Whether the anaphor has the phrase tag or not.
  - f_4 - *num_ana*: The number of the anaphor.
- Features of the antecedent candidate
  - f_5 - *word*: The word of the antecedent candidate.
  - f_6 - *lemma*: The lemma of the antecedent candidate.
  - f_7 - *cat_np*: The syntactic category of the NP.
  - f_8 - *dec_np*: The declension case of the NP.
  - f_9 - *num_np*: The number of the NP.
  - f_{10} - *degree*: The 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*: The syntactic function of the NP.
  - f_{13} - *enti_np*: The type of entity (PER, LOC, ORG).
- 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.

2.3 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.

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 sentence of the example.

Antecedent Candidate	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 Experimental Setup

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.

3.1 Learning Algorithms

We consider different machine learning paradigms from Weka toolkit [4] in order to find the best system for the task. On one hand, we use some typical classifiers like SVM, Multilayer Perceptron, Naïve Bayes, k-NN, and simple decision trees like C4.5 and REPTree. On the other hand, we use classifiers not so frequently used such as Random Forest (RF), NB-Tree and Voting Feature Intervals (VFI).

The SVM learner was evaluated by a polynomial kernel of degree 1. The k-NN classifier, $k = 1$, uses the Euclidean distance as distance function in order to find neigbours. Multilayer Perceptron is a neural network that uses backpropagation to learn the weights among the connections, whereas NB is a simple probabilistic classifier based on applying Bayes' theorem, and NB-Tree generates a decision tree with Naïve Bayes classifiers at the leaves. C4.5 and REPTree are well known decision tree classifiers. Random Forest and VFI are traditionally less used algorithms; however, they produce good results for our corpus. Random forest is a combination of tree predictors, such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. VFI constructs feature intervals for each feature. An interval represents a set of values for a given feature, where the same subset of class values is observed. Two neighbouring intervals contain different sets of classes.

3.2 Results for Single Classifiers

The results obtained with these classifiers are shown in Table 1. The best result is obtained by using the Multilayer Perceptron algorithm, an F-measure of 68.7%.

Table 1. Results for different algorithms

	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
C4.5	0.736	0.438	0.549
REPTree	0.715	0.524	0.605

In general, precision obtained is higher than recall. The best precision is obtained with SVM (80.3%), followed by NB-tree (77.1%). Althought C4.5 and REPTree are traditionally used for this task, they do not report good results for our corpus, as it can be observed in the table.

These results are not directly comparable with those obtained for other languages such as English, but we think that 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 Experimental Results

In this section the experimental results obtained are shown. It is worth to mention that one of the main contributions of this paper is concerned with the selection of single classifiers in order to perform the combination.

4.1 Combination of Classifiers

Classifier combination is very used in the Machine Learning community. The main idea is to combine some paradigms from the supervised classification trying to improve the individual accuracies of the component classifiers.

According to the architecture used to combine different single classifiers, there are three possible configurations: cascaded, parallel and hierarchical. In this paper we use two parallel combinations of classifiers. One of the ways to combine the classifiers in parallel consists of using several base classifiers, applying them to the database, and then combining their predictions using a vote process. But even with a unique base classifier, it is still possible to build an ensemble, applying it to different training sets in order to generate several different models. A way to get several training sets from a given dataset is bootstrap sampling, which is used in bagging [3].

4.2 Results Obtained

We tried both vote and bagging combination approaches based on the results obtained in the previous section for the single classifiers. We selected five single classifiers, which belong to different paradigms, and which obtain good results for our corpus: Multilayer Perceptron, Random Forest, VFI, NB and k-NN. We performed the experiments in the following way:

- We make a votation with those five classifiers. Three different voting criteria were used: Majority, average of probabilities and product of probabilities.
- We apply the bagging multiclassifier with those five single classifiers, using different number of classifiers: 10, 15, 20, 30 and 40.

Results obtained by applying the vote combination schema are shown in Table 2. As it can be seen a slight increase in results is obtained with the majority voting achieving an F-measure of 69.2%.

Table 2. Results for different voting criteria

Classifier voting criteria	F-measure
Majority voting	0.692
Vote: average of probabilities	0.684
Vote: product of probabilities	0.636

The bagging multiclassifier is supposed to obtain better results when "soft" base classifiers are used. Classification trees are a typical example of soft classifier. That is why, for comparison reasons, we applied a bagging combination of C4.5 and REPTree trees. In Table 3 just the best results obtained from the bagging process for each classifier are shown. Although it is not recommended, we applied bagging to the selected classifiers, some of which are not considered to be "soft". As it can be seen, results obtained using classification trees are worse than those obtained with the selected classifiers. However, they are the single classifiers which obtain the highest benefit from the combination.

The best result is obtained by the multilayer perceptron classifier as the base one, obtaining an F-measure of 70.3%.

Table 3. Results for the bagging multiclassifier

	Single	Bagging
C4.5	0.549	0.654
REPTree	0.605	0.657
VFI	0.663	0.664
Perceptron	0.687	**0.703**
RF	0.683	0.702
NB	0.654	0.654
k-NN	0.633	0.634

5 Conclusions and Future Work

This paper presents a study carried out on resolution of pronominal anaphora in Basque using a machine learning multiclassifier. The results obtained from this work will be helpful for the development of a better anaphora resolution tool for Basque.

We considered nine machine learning algorithms as single classifiers in order to decide which of them select to combine in a parallel manner. Two different classifier combination approaches were used: vote and bagging. The main contribution of the paper is the use of bagging having as base classifier a non-soft one for the anaphora resolution in Basque.

There are several interesting directions for further research and development based on this work. The introduction of other knowledge sources to generate new features and the use of composite features can be a way to improve the system.

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.

Acknowledgments

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

References

1. Aduriz, I., Aranzabe, M.J., Arriola, J.M., Daz de Ilarraza, A., Gojenola, K., Oronoz, M., Uria, L.: A cascaded syntactic analyser for basque. In: Gelbukh, A. (ed.) CICLing 2004. LNCS, vol. 2945, pp. 124–134. Springer, Heidelberg (2004)
2. Aduriz, I., Aranzabe, M.J., Arriola, J.M., Atutxa, A., Daz de Ilarraza, A., Ezeiza, N., Gojenola, K., Oronoz, M., Soroa, A., 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. In: Wilson, A., Archer, D., Rayson, P. (eds.) Language and Computers, Corpus Linguistics Around the World, Rodopi, Netherlands, pp. 1–15 (2006)
3. Breiman, L.: Bagging predictors. Machine Learning 24(2), 123–140 (1996)
4. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA Data Mining Software: An Update. SIGKDD Explorations 11(1) (2009)
5. Hirst, G.: Anaphora in Natural Language Understanding. Springer, Berlin (1981)
6. Moosavi, N.S., Ghassem-Sani, G.: Using Machine Learning Approaches for Persian Pronoun Resolution. In: Workshop on Corpus-Based Approaches to Conference Resolution in Romance Languages, CBA 2008 (2008)
7. Moosavi, N.S., Ghassem-Sani, G.: A Ranking Approach to Persian Pronoun Resolution. Advances in Computational Linguistics. Research in Computing Science 41, 169–180 (2009)

8. Nguy, G.L., Zabokrtský, Z.: Rule-based Approach to Pronominal Anaphora Resolution Method Using the Prague Dependency Treebank 2.0 Data. In: 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., Mart, M.A., Navarro, B.: 3LB: Construcción de una base de datos de árboles sintáctico-semánticos para el catalán, euskera y español. In: XX. Congreso SEPLN, Barcelona (2004)
10. Soon, W.M., Ng, H.T., Lim, D.C.Y.: A Machine Learning Approach to Coreference Resolution of Noun Phrases. Computational Linguistics 27(4), 521–544 (2001)
11. Vemulapalli, S., Luo, X., Pitrelli, J.F., Zitouni, I.: Using Bagging and Boosting Techniques for Improving Coreference Resolution. Informatica 34, 111–118 (2010)
12. Versley, Y.: A Constraint-based Approach to Noum Phrase Coreference Resolution in German Newspaper Text. In: Konferenz zur Verarbeitung Natrlicher Sprache KONVENS (2006)