A First Machine Learning Approach to Pronominal Anaphora Resolution in Basque

O. Arregi, K. Ceberio, A. Díaz de Illarraza,
I. Goenaga, B. Sierra, and A. Zelaia

University of the Basque Country
olatz.arregi@ehu.es

Abstract. In this paper we present the first machine learning approach to resolve the pronominal anaphora in Basque language. In this work we consider different classifiers in order to find the system that fits best to the characteristics of the language under examination. We do not restrict our study to the classifiers typically used for this task, we have considered others, such as Random Forest or VFI, in order to make a general comparison. We determine the feature vector obtained with our linguistic processing system and we analyze the contribution of different subsets of features, as well as the weight of each feature used in the task.

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 [7]: *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.

The first problem to carry out is the lack of a big annotated corpus in Basque. Mitkov in [12] 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.

2 Related Work

Although the literature about anaphora resolution with machine learning approaches is very large, we will concentrate on those references directly linked to

A. Kuri-Morales and G. Simari (Eds.): IBERAMIA 2010, LNAI 6433, pp. 234–243, 2010.

the work done here. In [20] they apply a noun phrase (NP) coreference system based on decision trees to MUC6 and MUC7 data sets ([15], [16]). It is usually used as a baseline in the coreference resolution literature.

Kernel functions to learn the resolution classifier are applied in [23]. They use structured syntactic knowledge to tackle pronoun resolution, and the results obtained for the ACE dataset show an improvement for all the different domains.

In [22] the authors propose kernel-based methods to resolve three coreference resolution subtasks (binding constraint detection, expletive identification and aliasing). They conclude that using kernel methods is a promising research direction to achieve state of the art coreference resolution results.

A rich syntactic and semantic processing is poposed in [5]. It outperforms all unsupervised systems and most supervised ones.

The state of the art of other languages varies considerably. In [18] 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 [21] the author uses a system based on a loglinear statistical model to resolve noun phrase coreference in German texts. On the other hand, [13] and [14] present an approach to Persian pronoun resolution based on machine learning techniques. They developed a corpus with 2,006 labeled pronouns.

A similar work was carried out for Turkish [24]. They apply a decision tree and a rule-based algorithm to an annotated Turkish text.

3 Selection of Features

3.1 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 [11]. 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 3.3 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.

3.2 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 our linguistic processing system proposed in [1]. We can not use some of the common features used by most systems ([20], [17], [23]) 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. Its possible values are from 1 to 15, the maximum distance shown in the corpus from an anaphor to its antecedent. The distance is measured in terms of number of Noun Phrases.
  - f_{15} - *same_sent*: If the anaphor and the antecedent candidate are in the same sentence the value is 0, otherwise the value is 1.
  - f_{16} - *same_num*: Its possible values are 0, 1, 2, and 3. If the anaphor and the antecedent candidate agree in number the value is 3, otherwise the value is 0. When the number of the noun phrase is unknown the value is 1. If the noun phrase is an entity, its number is indefinite and the anaphor is singular, then the value is 2. This last case is needed in Basque because person entities do not have singular or plural tags, but indefinite tag.

[1] Eus3LB is part of the 3LB project [19].

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.

3.3 Generation of Training Instances

The method we use to create training instances is similar to the one explained in [20]. 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.

4 Evaluation

In order to evaluate the performance of our system, we use the above mentioned corpus, with 349 positive and 619 negatives instances. 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.

4.1 Learning Algorithms

We consider different machine learning paradigms from Weka toolkit [6] in order to find the best system for the task. The classifiers used are: SVM, Multilayer Perceptron, NB, k-NN, 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.

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 that NB is a simple probabilistic classifier based on applying Bayes' theorem, and NB-Tree generates a decision tree with naive Bayes classifiers at the leaves. Random forest and VFI are traditionally less used algorithms; however, they produce the best 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 [3]. 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 [4].

4.2 Overall Results

Table 1. shows the results obtained with these classifiers.

Table 1. Results of 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

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 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.

5 Contribution of Features Used

Our next step is to determine the attributes to be used in the learning process. When there is a large number of attributes, even some relevant attributes may be redundant in the presence of others. Relevant attributes may contain useful information directly applicable to the given task by itself, or the information may be (partially) hidden among a subset of attributes [10].

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 [8]. 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 weight of these features taking into account all the measurements used.

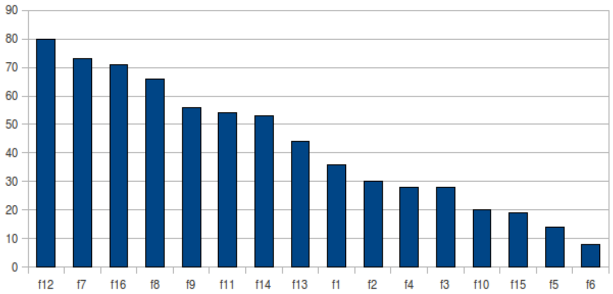

Fig. 1. The average weight of features

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.

If we test the algorithms presented in Section 4.1, taking into account the new order of features, and considering smaller subsets of features, the results are similar to the originals. In general, decreasing the number of features gives lower results. The best result (70%) is obtained with 14 features: the original set without the features *word* and *lemma*.

Table 2. shows the best F-measure results obtained with the classifiers mentioned above, taking into account different feature subsets. Only five methods are shown here, due to the fact that results obtained with SVM and NB-tree are not meaningful. SVM method does not improve the first result (64.5%) and NB-tree provides similar results to the ones obtained by simple NB.

Table 2. Results of five algorithms with different number of features

Number of features	VFI	Perceptron	RF	NB	k-nn
16	0.663	0.678	0.683	0.654	0.633
15	0.669	0.669	0.678	0.656	0.648
14	0.671	**0.692** all - $\{f_1, f_2\}$	**0.7** all - $\{f_5, f_6\}$	0.665	0.655
13	0.670	0.678	0.679	0.663	0.666
12	0.669	0.671	0.677	0.665	0.662
11	0.672	0.670	0.690	0.666	**0.674**
10	0.675	0.679	0.674	**0.669**	0.656
9	0.674	0.687	0.679	0.666	0.665
8	0.674	0.672	0.682	0.661	0.661
7	0.677	0.668	0.661	0.655	0.644
6	**0.684**	0.652	0.664	0.650	0.640
5	0.673	0.645	0.652	0.640	0.625
4	0.655	0.619	0.628	0.632	0.600
3	0.646	0.639	0.661	0.619	0.616
2	0.629	0.635	0.626	0.607	0.617

Although the two best results were obtained with 14 features, 69.2% (perceptron) and 70% (RF), the set of attributes selected in both cases is different, since in the first case the best selection of features is produced by the relief algorithm (all features except *sf_ana* and *dec_ana*), and in the second case features were chosen following the order established by the Gain Ratio measurement (all features except *word* and *lemma*). For the rest of the algorithms, the best results are obtained by using a smaller set of attributes (from 6 to 11); nevertheless these results are lower than those mentioned above. For all the algorithms we obtained a higher value than the original F-measure. Table 3. shows these values.

Table 3. Results obtained with different subsets of features

	original	best	Number of
	F-measure	F-measure	features
VFI	0.663	0.684	6
Perceptron	0.687	0.692	14
RF	0.683	**0.700**	14
NB	0.654	0.669	10
k-nn	0.633	0.670	11

For the k-NN method the measurement which offers the best results is, in most cases, the Relief algorithm. This result was expected as this algorithm evaluates the weight of a feature by repeatedly sampling an instance and considering the value of the given feature for the nearest instance of the same and different class. So, given an instance, Relief algorithm searches for its two nearest neighbours, and the k-NN algorithm is based on the same idea. The selection of the nearest neighbours is crucial in Relief. The purpose is to find the nearest neighbours with respect to important attributes [9].

5.1 The Contribution of Single Attributes

If we use a single attribute each time for the classification process, we can determine that the best attribute is *sf_np*, that is, the syntactic function of the noun phrase, with an F-measure equal to 0.480 but a precision of 0.905.

Table 4. shows the results obtained for this test applying Random Forest algorithm. Unsurprisingly many of the attributes result in zero. It should be noted that as in other works [20], selected attributes provide high values for precision, although the recall is very low. The first four attributes of the table, which are the same as those selected by the measurements introduced at the beginning of this section, provide a precision above 65%, reaching to 90% in the case of the first attribute (*sf_np*). In contrast, the F-measure values are lower than 50%.

Table 4. Results obtained using just one attribute at a time

	Precision	Recall	F-measure
sf_np	0.905	0.327	0.480
cat_np	0.659	0.309	0.421
same_num	0.811	0.123	0.214
dec_np	0.837	0.249	0.384
lemma	0.421	0.381	0.400
word	0.378	0.347	0.362
dist	0.364	0.011	0.022
Rest of attributes	0.000	0.000	0.000

6 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.

We also analyzed the contribution of features used in order to decide which of them are important and which are not. With a good combination of features we obtain an F-measure of 70%, which is the best result obtained in this work.

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.

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.

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.: Random Forests. Machine Learning 45(1), 5–32 (2001)
4. Demiroz, G., Guvenir, A.: Classification by voting feature intervals. In: 9th European Conference on Machine Learning, pp. 85–92 (1997)
5. Haghighi, A., Klein, D.: Simple Coreference Resolution with Rich Syntactic and Semantic Features. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Singapore, pp. 1152–1161 (2009)
6. 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)
7. Hirst, G.: Anaphora in Natural Language Understanding. Springer, Berlin (1981)
8. Kira, K., Rendell, L.A.: A Practical Approach to Feature Selection. In: Ninth International Workshop on Machine Learning, pp. 249–256 (1992)
9. Kononenko, I.: Estimating Attributes: Analysis and Extensions of RELIEF. In: European Conference on Machine Learning, pp. 171–182 (1994)
10. Kononenko, I., Hong, S.J.: Attribute Selection for Modeling. Future Generation Computer Systems 13, 181–195 (1997)

11. Laka, I.: A Brief Grammar of Euskara, the Basque Language. Euskarako errektore-ordetza, EHU (2000), `http://www.ehu.es/grammar`
12. Mitkov, R.: Anaphora resolution. Longman, London (2002)
13. Moosavi, N.S., Ghassem-Sani, G.: Using Machine Learning Approaches for Persian Pronoun Resolution. In: Workshop on Corpus-Based Approaches to Coreference Resolution in Romance Languages. CBA 2008 (2008)
14. 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)
15. MUC-6.: Proceedings of the Sixth Message Understanding Conference (MUC-6). Morgan Kaufmann, San Francisco, CA (1995)
16. MUC-7.: Proceedings of the Seventh Message Understanding Conference (MUC-7). Morgan Kaufmann, San Francisco, CA (1998)
17. Ng, V., Cardie, C.: Improving Machine Learning Approach to Coreference Resolution. In: Proceedings of the ACL, pp. 104–111 (2002)
18. Nguy, Zabokrtský: 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)
19. 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. XX. Congreso SEPLN, Barcelona (2004)
20. 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)
21. Versley, Y.: A Constraint-based Approach to Noum Phrase Coreference Resolution in German Newspaper Text. In: Konferenz zur Verarbeitung Natrlicher Sprache KONVENS (2006)
22. Versley, Y., Moschitti, A., Poesio, M., Yang, X.: Coreference System based on Kernels Methods. In: Proceedings of the 22nd International Coreference on Computational Linguistics (Coling 2008), Manchester, pp. 961–968 (2008)
23. Yang, X., Su, J., Tan, C.L.: Kernel-Based Pronoun Resolution with Structured Syntactic Knowledge. In: Proc. COLING/ACL 2006, Sydney, pp. 41–48 (2006)
24. Yldrm, S., Klaslan, Y., Yldz, T.: Pronoun Resolution in Turkish Using Decision Tree and Rule-Based Learning Algorithms. In: Human Language Technology. Challenges of the Information Society. LNCS. Springer, Heidelberg (2009)