IS-G: The Comparison of Different Learning Techniques for the Selection of the Main Subject References

Bernd Bohnet
University of Stuttgart
Visualization and Interactive Systems Group
Universitätstr. 58, 70569 Stuttgart, Germany
bohnet@informatik.uni-stuttgart.de

Abstract

The GREC task of the Referring Expression Generation Challenge 2008 is to select appropriate references to the main subject in given texts. This means to select the correct type of the referring expressions such as *name*, *pronoun*, *common*, or *elision (empty)*. We employ for the selection different learning techniques with the aim to find the most appropriate one for the task and the used attributes. As training data, we use the syntactic category of the searched referring expressions and additionally gathered data from the text itself.

1 Introduction

The training data of the GREC task consists of Wikipedia articles from five domains. The articles are about People, Cities, Countries, Rivers, and Mountains. An XML annotation replaces the original referring expressions in the articles with a set of alternative experssions which could be inserted in the empty space in order to complete the text. The training data additionally consists of the original referring expressions. From the annotations, we use for the training the *syntactic category* (SYNCAT) for the searched referring expression.

The annotation allows us easily to access additional data. One of the values that we calculate is the *distance* (DIST) to the last referring expression. The idea behind this value is that the references becomes with increasing *distance* unclear because of other content in the text and therefore, to provide a short form such as a pronoun is not enough or would be even misleading. Of cause there might be other reasons to use the name too. The next information, we use is the *count* (POS) of the referring expression in the text. Because we expect that the first referring expression is in most cases the name of the main subject and at the end of a text, it might be used less frequent. Then we use the type of the *last* (LAST) used referring expression in the text. This might be a good candidate because people want to avoid consecutive repetitions of the same word. The disadvantage of this attribute is that it is based itself on the classification of its predecessor and therefore, on insecure information. Finally, we use a second distance measure which provides the information if the last referring expression was in the same sentence (SENT).

2 Comparison of Learning Techniques

We tried several machine learning techniques and selected among them the three bests that are Bayesian Networks with the attribute selection method K2 (Cooper and Herskovits, 1992), decision trees C4.5 (Quinlan, 1993), and Multi Layer Perceptrons with a sigmoid function (Minsky and Papert, 1969). For the comparison with the three machine learning techniques, we provide in Table 1 a base line where we chose the type with the most occurrences in the training data of the domain.

Table 2 shows the results for the Bayesian Network. The results are significant better then the base line results. Table 3 shows the results of C4.5. The results are close to the results of the Bayesian Network. An advantage of decision trees is that they provide some explanation. A part of the decision tree is shown in Figure 1. The part selects a refer-

Set	Most frequent type	Accuracy
Cities	name	0.47
Countries	name	0.45
Mountains	name	0.39
People	pronoun	0.61
Rivers	name	0.43
Total	–	0.47

Table 1: Base Line

Set	Cities	Cou.	Mount.	Peo.	Ri.	Total
Acc	0.48	0.62	0.63	0.673	0.7	0.62

Table 2: Results for Bayesian Networks (K2)

ring expression in the case that the last expression was already within the same sentences.

Set	Cities	Cou.	Mount.	Peo.	Ri.	Total
Acc	0.545	0.63	0.641	0.673	0.7	0.638

Table 3: Results for C4.5

The uppercase words are the attributes followed by the value for the branch. If the value of a distinct instance is in the range of the value then the algorithms chooses the branch until it reaches a leaf. The leafs are labelled with the result of the decision and with information of an evaluation that provides the information how many training instances (cases) are classified correct / wrong. Interesting for the case are the following observations:

- The text writers chose nearly always (>99%) an other referring expression than the name.
- They select more frequent pronouns and an elisions (empty) compared to common names.
- The writers select common names in case of a high distance to the last referring expression.

Table 4 shows the results for the Multi Layer Perceptron, which performed best compared to the other learning techniques.

3 Conclusion

We calculated the information gain of each attribute to get an overview of the relevance of the attributes. The most releveant attribute is DIST (0.32) followed by POS (0.24), LAST (0.239), SENT (0.227), and SYNCAT (0.19).

```
SENT = true
— SYNCAT = subj-det
— — DIST <= 158: pronoun (103.0/8.0)
— — DIST > 158: common (4.0/1.0)
— SYNCAT = np-subj
— — DIST <= 33: pronoun (32.0/15.0)
— — DIST > 33
— — — LAST = common
— — — — DIST <= 123: empty (25.0/5.0)
— — — — DIST > 123: common (2.0/1.0)
— — — LAST = pronoun: empty (69.0/22.0)
— — — LAST = name
— — — — POS <= 2: empty (17.0/2.0)
— — — — POS > 2
— — — — — POS <= 15
— — — — — — DIST <= 79
— — — — — — — DIST <= 39: pronoun (3.0)
— — — — — — — DIST > 39: empty (46.0/15.0)
— — — — — — DIST > 79
— — — — — — — POS <= 5: pronoun (3.0)
— — — — — — — POS > 5
— — — — — — — — POS <= 11
— — — — — — — — — DIST <= 113: common (4.0/1.0)
— — — — — — — — — DIST > 113: name (2.0/1.0)
— — — — — — — — POS > 11: pronoun (2.0)
— — — — — POS > 15: common (2.0)
— — — LAST = empty: pronoun (5.0/1.0)
— SYNCAT = np-obj
— — DIST <= 109
— — — POS <= 9: pronoun (23.0/12.0)
— — — POS > 9: common (10.0/4.0)
— — DIST > 109: common (17.0/3.0)
SENT = false
...
```

Figure 1: Part of the decision tree

The results of all three learning techniques are significant better than the base line which has in average an accuracy of 0.47. The multi layer perceptron provides the best results with an average accuracy of 0.66.

References

G. F. Cooper and E. Herskovits. 1992. A Bayesian Method for the Induction of Probabilistic Networks from Data. In *Machine Learning 9*.

M. Minsky and S. Papert. 1969. *Perceptrons*. MIT Press.

J. R. Quinlan. 1993. *C4.5 Programs for Machine Learning*. Morgan Kaufmann, California.

Set	Cities	Cou.	Mount.	Peo.	Ri.	Total
Acc	0.545	0.64	0.65	0.668	0.8	0.66

Table 4: IS-G: Multi Layer Perceptron