Improving Anaphora Resolution performance in Italian and English using Search Engines

Fabio Celli
CLIC-CIMeC, University of Trento
fabio.celli@unitn.it
June 26, 2010

Abstract
In this paper I describe an heuristic method for improving anaphora resolution by classifying as co-referring the pairs in a high co-occurrence with taxonomic relations (Is-A) and as non-coreferring the pairs in a high co-occurrence with associative relations (Part-Of). The experiments reported here show that it is possible to classify correctly 65% of the errors of an anaphora resolution system by using simple pattern-based relation extraction from web.

1 Introduction and Background
Anaphora resolution, that is the task of recognizing coreferring expressions automatically in text, has proven to be very important in a number of other natural language processing tasks, like question answering and text summarization, to cite just a few. In order to improve the performance of anaphora resolution Ji et al. [6] used Semantic Relation Extraction obtaining good results in English and Chinese with ACE-style semantic relations like Person-Organization or Person-Location. Those relations have a very high degree of formalization and can be extracted only between entities (hence a Named Entity Recognizer is needed), getting harder the task.
In this paper I describe an heuristic method for improving anaphora resolution that is different (and simpler) from the one presented by Ji et al. for two aspects. First: it makes use of web search patterns and second: I applied it on Italian and English. In a previous work [1] I found that taxonomic relations, such as Is-A and Coordinate, are strongly associated to the presence of coreference while associative relations, such as Part-Of and Member-Of, are strongly associated to the absence of coreference. Thus the hypothesis is that: given a pair of expression, if I compare the frequency of that pair with a taxonomic and an associative relation pattern, I can retrieve a measure of the probability that the words in the pair are coreferring expressions. This could help in classifying correctly false positive and false negative anaphora examples. In another previous work [2] I found that this heuristic method works fine with anaphoric pairs with different heads. There is a long tradition in Is-A and Part-Of relation extraction via patterns, starting with Hearst 1992 [5], to Girju et al. 2006 [4].
and this can be done either in corpora or in the web, using search engines. In the next section I describe the experiments, and in section 3 I will discuss the results and draw some provisional conclusions.

2 Experiments

Experiment 1: Google search vs Yahoo search In the first experiment I compared the performance of Yahoo and Google web search. I randomly extracted 128 pairs of false positive and 128 pairs of false negative examples from the output of BART, a system for anaphora resolution developed by Versley et al [11], 500 examples in total for each dataset. I have 2 datasets one from ACE 2002 [8] (English) and the other one from ICAB [7] (Italian); for each pair I searched the web for the cooccurrences of each pair and the following patterns: “kind of”, “part of”, “same as”, “different from”, and the respective Italian translations “tipo di”, “parte di”, “uguale a”, “diverso da”. Results are reported in table 1 below. I tried the classification either with SVMs [3] and decision trees C4.5 [9] in Weka (see [10]), with a 10-fold cross validation. Results show that Google search counts yield better results using SVMs, while Yahoo yields better results using decision trees. No big difference is observed between the two results, so I chose Yahoo as search engine for the following experiments since Google has some limitation of query per day and collecting the data takes much more time.

<table>
<thead>
<tr>
<th>dataset</th>
<th>F1 (SVM)</th>
<th>F1 (C4.5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ace-google</td>
<td>0.663</td>
<td>0.647</td>
</tr>
<tr>
<td>ace-yahoo</td>
<td>0.647</td>
<td>0.652</td>
</tr>
<tr>
<td>icab-google</td>
<td>0.652</td>
<td>0.689</td>
</tr>
<tr>
<td>icab-yahoo</td>
<td>0.615</td>
<td>0.722</td>
</tr>
</tbody>
</table>

Table 1: Results of experiment 1.

Experiment 2: Learning curve I randomly sampled 500 false positive and 500 false negative examples either from ACE and ICAB processed with Bart, 1000 examples in total for English and 1000 for Italian. Then I tagged the false negative examples as coreferring expressions and the false positives as non-coreferring expressions, then I run a binary classification task using as features the frequency of the pair with the pattern described above: “kind of”, “part of”, “same as”, “different from”, and the respective Italian translations in the case of ICAB “tipo di”, “parte di”, “uguale a”, “diverso da”. The frequency is obtained from Yahoo APIs and normalized. I run the classification experiment three times using 256 examples (128 positive and 128 negative), 500 examples (250 positive and 250 negative) and the entire dataset. Like before I run the experiments in Weka, using C4.5 decision trees (dashed line in the figure) and SVMs (normal line in the figure). Evaluation is with a 10-fold cross validation. Results are reported in table 2 below and displayed as learning curve in figure 1:

The results show that the learning curve does not increase much adding more than 500 examples, and this is true both in English and in Italian. Decision
<table>
<thead>
<tr>
<th>dataset</th>
<th>F1 (SVM)</th>
<th>F1 (C4.5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ace-250</td>
<td>0.647</td>
<td>0.652</td>
</tr>
<tr>
<td>ace-500</td>
<td>0.61</td>
<td>0.634</td>
</tr>
<tr>
<td>ace-1000</td>
<td>0.658</td>
<td>0.658</td>
</tr>
<tr>
<td>icab-250</td>
<td>0.615</td>
<td>0.722</td>
</tr>
<tr>
<td>icab-500</td>
<td>0.504</td>
<td>0.662</td>
</tr>
<tr>
<td>icab-1000</td>
<td>0.532</td>
<td>0.666</td>
</tr>
</tbody>
</table>

Table 2: Results of experiment 2.

Figure 1: Learning curve of the classification task. Dashed line is the performance with C4.5, normal line is the performance with SVMs

trees work better than SVMs in this task, especially for Italian, while no big difference in using SVMs or decision trees is found with the ACE dataset. A discussion follow in the next section.

3 Discussion and Conclusion

By the one hand the results confirm that Google and Yahoo are both suitable choices for web search in the task of relation extraction via patterns. On the other hand results show that it is possible to improve the performance of anaphora resolution by using relation extraction from web as an euristic method for reducing anaphora classifier’s errors. Still is left open the question whether adding more patterns to the model yield better results or not, this could be a task for future work.

In conclusion I presented a model suitable as an euristic method that improves the performance of anaphora resolution exploiting relation extraction from web,
both in Italian and English, but this model could be extended easily to other languages supported by search engines, simply translating the patterns. This can be done also in a automated way.

References


