Pattern	Suggested	Annotator1	Annotator 2	Annotator 3	Average
HEARST1	2	40.00%	40.00%	60.00%	46.66%
DEFINITE1	19	21.05%	36.84%	36.84%	31.56%
DEFINITE2	74	91.36%	93.83%	96.30%	93.83%
APPOSITION	28	56.00%	62.00%	62.00%	60.00%
COPULA	22	66.67%	66.67%	63.64%	65.66%
ALL	188	69.15%	73.40%	74.47%	72.34%

Table 2: Accuracy of each of the patterns

Pattern	Relative Weight
HEARST1-4	5
DEFINITE1	3
DEFINITE2	9
APPOSITION	6
COPULA	7

Table 3: Relative weights of the patterns

Table 2 gives the accuracy for all the patterns based on the answers of the human subjects to the suggestions of the system. Unfortunately, no HEARST2, HEARST3 or HEARST4 instances were found in the texts, which shows that they are actually the ones which occur most rarely.

The above task can be seen as a classification task of the suggested instance-concept relationships into the four categories: *correct*, *add concept*, *wrong* and *doubt*. Thus, we can measure the categorial agreement between the three annotators as given by the Kappa statistic ([5]). In fact, when computing the average of the pairwise agreement between the annotators, we yield a Kappa value of K=66.19%. Thus the agreement seems quite reasonable and according to [5] is almost in a range from which 'tentative conclusions' can be drawn. This in turn means that our task is well defined.

The results itself show that in general the accuracy of the patterns is relatively good, i.e., almost 3/4 of the suggested instance-concept relations are correct. It also shows that the Hearst patterns are extremely rare. It is also interesting to notice that the *DEFINITE1* and *DEFINITE2* patterns, though they share the same rationale, have a completely different performance in terms of accuracy. Finally - and most importantly - the results show that the performance of each of the patterns in terms of accuracy is very different such that there is an actual need of weighting the contribution of each pattern. As a first approximation of setting the weights of the patterns to maximize the overall accuracy of the approach, we decided to weight the patterns relatively to each other proportionally to their accuracy. In particular we used the relative weights in Table 3. However, we found out that weighting the patterns in this linear fashion makes the results actually worse. In fact, the best F-Measure was $F_{1,avg} = 24.54\%$ ($t = 290$) and the best accuracy was $Acc_{avg} = 21.48\%$ ($t = 0$). As a further experiment we also tried to find optimal weights by training a neural network as well as other classifiers. However, due to the lack of a representative number of (positive) training examples, the model learned by the classifiers was worse than our baseline.

4.3.3 Interactive Selection

When using the interactive selection variant, i.e., $R^5_{b,0}$, if one of the top 5 answers of our system coincides with the one given by the annotator, we count it as a correct answer. Thus, we obviously get higher F-Measure, Precision and Accuracy (Recall) values. They are depicted in comparison to the baseline in Figure 4. The best accuracy here is Acc_{avg}=49.56%. This means in practice that for almost half of the instances in a web page, we provide the user already with the correct answer, thus notably reducing the annotation time and cost.

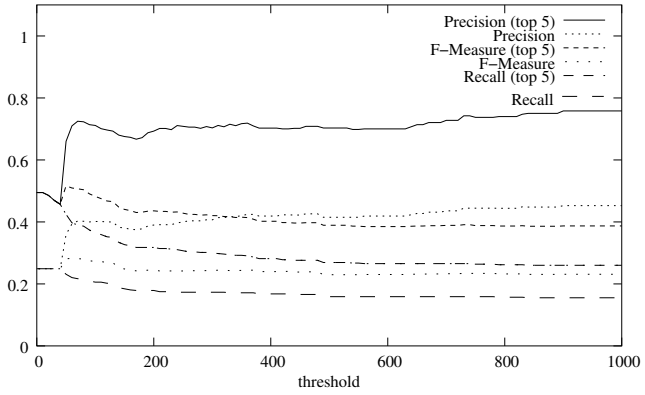

Figure 4: Precision, F-Measure and Accuracy/Recall for $R^5_{b,0}$ compared to $R_{b,\theta}$ zoomed into interval [0..1000]

4.4 Discussion

The results of the experiment described above are certainly very encouraging. As Table 1 shows, the overall results of our automatic classification seem quite reasonable. Some instance-concept relationships are certainly spurious, such as, for example, that *South Africa* is a *town*. In fact, the second best ranked category of our approach for *South Africa* is the correct one, i.e., *country*. Thus, a semi-automatic use of our approach in which the users are asked to select one of the highest ranked categories increases considerably the performance of our approach (compare section 4.3.3).

From a quantitative point of view, the best Accuracy of 24.9% is comparable to state-of-the-art systems performing a similar classification task, especially given the fact that our approach is unsupervised and does not require text preprocessing methods (see Section 6). The performance of our system is still far away from the human performance on the task ($F_1 = 62.09\%$), but it is also quite away from a random decision procedure. Thus, the results of our approach seem very promising. In future experiments we will verify if these results are scalable to a larger set of concepts such as the ca. 1200 considered by Alfonseca and Manandhar ([2]).

5. INTEGRATION INTO CREAM

We have integrated PANKOW into the CREAM framework [16] extending the CREAM implementation OntoMat by a plugin. The plugin has access to the ontology structure and to the document