Table 5: Resolution accuracy (%)

Dataset		ANC	AQT	MUC
1	Previous noun	36.7	34.5	30.8
2	No Prob. Features	58.1	60.9	49.7
3	No Prob. Gender	65.8	71.0	68.6
4	No MI	71.3	73.5	69.2
5	No $C(p)$	72.3	73.7	69.8
6	Full System	73.9	75.0	71.6
7	Upper Bound	93.2	92.3	91.1

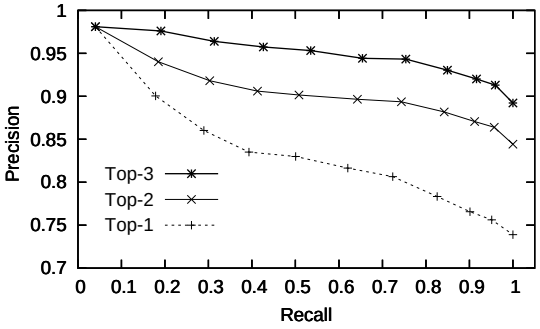

Figure 2: ANC pronoun resolution accuracy for varying SVM-thresholds.

were set using cross-validation on the training set; test sets were used only once to obtain the final performance values.

Evaluation Metric: We report results in terms of accuracy: Of all the anaphoric pronouns in the test set, the proportion we resolve correctly.

6 Results and Discussion

We compare the accuracy of various configurations of our system on the ANC, AQT and MUC datasets (Table 5). We include the score from picking the noun immediately preceding the pronoun (after our hard filters are applied). Due to the hard filters and limited search window, it is not possible for our system to resolve every noun to a correct antecedent. We thus provide the performance upper bound (i.e. the proportion of cases with a correct answer in the filtered candidate list). On ANC and AQT, each of the probabilistic features results in a statistically significant gain in performance over a model trained and tested with that feature absent.[5] On the smaller MUC set, none of the differences in 3-6 are statistically significant, however, the relative contribution of the various features remains reassuringly constant.

Aside from missing antecedents due to the hard filters, the main sources of error include inaccurate statistical data and a classifier bias toward preceding *pronouns* of the same gender/number. It would be interesting to see whether performance could be improved by adding WordNet and web-mined features. Path coreference itself could conceivably be determined with a search engine.

Gender is our most powerful probabilistic feature. In fact, inspecting our system's decisions, gender often rules out coreference regardless of path coreference. This is not surprising, since we based the acquisition of $C(p)$ on gender. That is, our bootstrapping assumption was that the majority of times these paths occur, gender indicates coreference or lack thereof. Thus when they occur in our test sets, gender should often sufficiently indicate coreference. Improving the orthogonality of our features remains a future challenge.

Nevertheless, note the decrease in performance on each of the datasets when $C(p)$ is excluded (#5). This is compelling evidence that path coreference is valuable in its own right, beyond its ability to bootstrap extensive and reliable gender data.

Finally, we can add ourselves to the camp of people claiming semantic compatibility is useful for pronoun resolution. Both the MI from the pronoun in the antecedent's context and vice-versa result in improvement. Building a model from enough text may be the key.

The primary goal of our evaluation was to assess the benefit of path coreference within a competitive pronoun resolution system. Our system does, however, outperform previously published results on these datasets. Direct comparison of our scoring system to other current top approaches is made difficult by differences in preprocessing. Ideally we would assess the benefit of our probabilistic features using the same state-of-the-art preprocessing modules employed by others such as (Yang et al., 2005) (who additionally use a search engine for compatibility scoring). Clearly, promoting competitive evaluation of pronoun resolution scoring systems by giving competitors equivalent real-world preprocessing output along the lines of (Barbu and Mitkov, 2001) remains the best way to isolate areas for system improvement.

Our pronoun resolution system is part of a larger information retrieval project where resolution ac-

[5] We calculate significance with McNemar's test, p=0.05.