Feature	Description
MarginalPopularity	The meaning candidate's marginal popularity score
ConditionalPopularity	The meaning candidate's conditional popularity score
WikiPopularity	The meaning candidate's Wiki popularity score
ContextSimilarity	TFIDF cosine similarity between meaning context and acronym context
LeftNeighborScore	Probability of acronym and meaning sharing the same immediate left word
RightNeighborScore	Probability of acronym and meaning sharing the same immediate right word
FullNamePercentage	Percentage of meaning candidate's component words appearing in acronym context

Table 1: Candidate Ranking Features

Feature	Description
Top1Score	Top 1 ranked meaning candidate's ranking score
Top1&2ScoreDiff	Difference between 1st and 2nd ranked meaning candidates' ranking score
Top1&2CtxSimDiff	Difference between 1st and 2nd ranked meaning candidates' context similarity score
Top1WikiPopularity	Top 1 ranked meaning candidate's Wiki popularity score
MaxWikiPopularity	Max Wiki popularity score among all the meaning candidates
MaxWP&MPGap	Max gap between Wiki and marginal popularity among all the meaning candidates
MaxWP&CPGap	Max gap between Wiki and conditional popularity among all the meaning candidates

Table 2: Confidence Estimation Features

The goal of the final selection model is similar to that of the confidence estimation model. In confidence estimation, we judge whether the top-ranked answer is correct; while in final selection, we check whether the most popular external meaning is correct. Thanks to this similarity, we can reuse the data, features and training algorithm in confidence estimation model. We take the same training data in Section 5.2.1 and update the labels correspondingly: if the genuine answer is the most popular external meaning, we generate a positive example; otherwise we make a negative one.

6 Experiments

6.1 Data

6.1.1 Mining and Training Corpus

We use both the Microsoft Answer Corpus (MAC) and the Microsoft Yammer Corpus (MYC) as the mining corpus. These corpus are kindly shared to us by Microsoft for research purpose. MAC contains 0.3 million web pages from a Microsoft internal question answering forum. MYC is consisted of 6.8 million posts from Microsoft's Yammer social network. In total, our mining module harvested 5287 acronyms and 17258 meaning candidates from this joint corpus.

For model training, the confidence estimation model and final selection model need to be trained on a different corpus than the candidate ranking model on MAC, with 12500 training examples being automatically generated; and train the confidence estimation and final selection model on MYC, with 40000 training instances being generated.

6.1.2 Evaluation Datasets

We prepared four datasets[3] for evaluation purposes. The first one ***Manual*** is obtained from the recent pages of Microsoft answer forum. Note these pages are disjoint from those used as mining/training corpus. We randomly sampled 300 pages and filtered out pages which do not contain ambiguous acronyms. After filtering, 240 test cases were left and we manually labeled them.

The second one ***Distant*** is generated via distant labeling on Microsoft Office365 documents. We sampled 2000 documents which contain at least one occurrence of a meaning candidate. Then we replaced the meanings with the corresponding acronyms and treat the meanings as ground truths. We manually checked through this dataset to remove some bad cases (e.g., "AS" for "App Store"). This resulted in a test set of 1949 test cases.

Comparing the ***Manual*** dataset with the ***Distant*** dataset, the ***Manual*** one, though in smaller size, can more accurately evaluate the system performance, since the target acronyms in it are sampled from the real distribution, while in the ***Distant*** dataset acronyms are artificially generated from

[3]Due to data confidentiality issue, we were unable to directly release these datasets.