where R_q is the web search results of query q; R_c is the web search results of c which is the top-ranked correction of q suggested by model M1.

The new feature set denoted with S2 is a set of document similarities between R_q and R_c, which includes different similarity estimations between the query and its correction at the document level using merely cosine measure based on term frequency vectors of R_q and R_c.

6 Experiments

6.1 Evaluation Metrics

In our work, we consider the following four types of evaluation metrics:

- **Accuracy**: The number of correct outputs proposed by the spelling correction model divided by the total number of queries in the test set
- **Recall**: The number of correct suggestions for misspelled queries by the spelling correction model divided by the total number of misspelled queries in the test set
- **Precision**: The number of correct suggestions for misspelled queries proposed by the spelling correction model divided by the total number of suggestions made by the system
- **F-measure**: Formula $F = 2PR/(P + R)$ used for calculating the f-measure, which is essentially the harmonic mean of recall and precision

Any individual metric above might not be sufficient to indicate the overall performance of a query spelling correction model. For example, as in most retrieval tasks, we can trade recall for precision or vice versa. Although intuitively F might be in accordance with accuracy, there is no strict theoretical relation between these two numbers – there are conditions under which accuracy improves while F-measure may drop or be unchanged.

6.2 Experimental Setup

We used a manually constructed data set as gold standard for evaluation. First we randomly sampled 7,000 queries from search engine's daily query logs of different time periods, and had them manually labeled by two annotators independently. Each query is attached to a truth, which is either the query itself for valid queries, or a spelling correction for misspelled ones. From the annotation results that both annotators agreed with each other, we extracted 2,323 query-truth pairs as training set and 991 as test set. Table 1 shows the statistics of the data sets, in which E_q denotes the error rate of query and E_t denotes the error rate of term.

	# queries	# terms	E_q	E_t
Training set	2,323	6,318	15.0%	5.6%
Test set	991	2,589	12.8%	5.2%

Table 1. Statistics of training set and test set

In the following experiments, at most 50 correction candidates were used in the maximum entropy model for each query if there is no special explanation. The web search results were fetched from MSN's search engine. By default, top 100 retrieved items from the web retrieval results were used to perform feature extraction. A set of query log data spanning 9 months are used for collecting statistics required by the baseline.

6.3 Overall Results

Following the method as described in previous sections, we first ran a group of experiments to evaluate the performance of each model we discussed with default settings. The detailed results are shown in Table 2.

Model	Accuracy	Recall	Precision	F
M0	91.8%	60.6%	62.6%	0.616
M1	93.9%	64.6%	77.4%	0.704
M2	94.7%	66.9%	78.0%	0.720

Table 2. Overall Results

From the table we can observe significant performance boosts on all evaluation metrics of M1 and M2 over M0.

We can achieve 25.6% error rate reduction and 23.6% improvement in precision, as well as 6.6% relative improvement in recall, when adding S1 to M1. Paired t-test gives p-value of 0.002, which is significant to 0.01 level.

M2 can bring additional 13.1% error rate reduction and moderate improvement in precision, as well as 3.6% improvement in recall over M1, with paired t-test showing that the improvement is significant to 0.01 level.