query is still a new challenge. Although there are some similarities in correction candidate generation and selection, these two settings are quite different in one fundamental problem: How to determine the validity of a search term. Traditionally, the measure is mostly based on a pre-defined spelling lexicon – all character strings that cannot be found in the lexicon are judged to be invalid. However, in the web search context, there is little hope that we can construct such a lexicon with ideal coverage of web search terms. For example, even manually collecting a full list of car names and company names will be a formidable task.

To obtain more accurate understanding of this problem, we performed a detailed investigation over one week's MSN daily query logs, among which found that 16.5% of search terms are out of the scope of our spelling lexicon containing around 200,000 entries. In order to get more specific numbers, we also manually labeled a query data set that contains 2,323 randomly sampled queries and 6,318 terms. In this data set, the ratio of out-of-vocabulary (OOV) terms is 17.4%, which is very similar to the overall distribution. However, only 25.3% of these OOV terms are identified to be misspelled, which occupy 85% of the overall spelling errors. All these statistics indicate that accurate OOV term classification is of crucial importance to good query spelling correction performance.

Cucerzan and Brill (2004) first investigated this issue and proposed to use query logs to infer correct spellings of misspelled terms. Their principle can be summarized as follows: given an input query string *q*, finding a more probable query *c* than *q* within a confusion set of *q*, in which the edit distance between each element and *q* is less than a given threshold. They reported good recall for misspelled terms, but without detailed discussions on accurate classification of valid out-of-vocabulary terms and misspellings. In Li's work, distributional similarity metrics estimated from query logs were proposed to be used to discriminate high-frequent spelling errors such as *massenger* from valid out-of-vocabulary terms such as *biocycle*. But this method suffers from the data sparseness problem: sufficient amounts of occurrences of every possible misspelling and valid terms are required to make good estimation of distributional similarity metrics; thus this method does not work well for rarely-used out-of-vocabulary search terms and uncommon misspellings.

In this paper we propose to use web search results to further improve the performance of query spelling correction models. The key contribution of our work is to identify that the dynamic online search results can serve as additional evidence to determine users' intended spelling of a given term. The information in web search results we used includes the number of pages matched for the query, the term distribution in the web page snippets and URLs. We studied two schemes to make use of the returning results of a web search engine. The first one only exploits indicators of the input query's returning results, while the other also looks at other potential correction candidate's search results. We performed extensive evaluations on a query set randomly sampled from search engines' daily query logs, and experimental results show that we can achieve 35.4% overall error rate reduction and 18.2% relative *F*-measure improvement on OOV misspelled terms.

The rest of the paper is structured as follows. Section 2 details other related work of spelling correction research. In section 3, we show the intuitive motivations to use web search results for the query spelling correction. After presenting the formal statement of the query spelling correction problem in Section 4, we describe our approaches that use machine learning methods to integrate statistical features from web search results in Section 5. We present our evaluation methods for the proposed methods and analyze their performance in Section 6. Section 7 concludes the paper.

2 Related Work

Spelling correction models in most previous work were constructed based on conventional task settings. Based on the focus of these task settings, two lines of research have been applied to deal with non-word errors and real-word errors respectively.

Non-word error spelling correction is focused on the task of generating and ranking a list of possible spelling corrections for each word not existing in a spelling lexicon. Traditionally candidate ranking is based on manually tuned scores such as assigning alternative weights to different edit operations or leveraging candidate frequencies (Damerau, 1964; Levenshtein, 1966). In recent years, statistical models have been widely used for the tasks of nat-