An Information Retrieval Approach to Spelling Suggestion

Sai Krishna
Search and Information
Extraction Lab, LTRC,
IIIT Hyderabad, India
saikrishna@research.iiit.ac.in

Prasad Pingali
Search and Information
Extraction Lab, LTRC,
IIIT Hyderabad, India
pvvpr@iiit.ac.in

Vasudeva Varma
Search and Information
Extraction Lab, LTRC,
IIIT Hyderabad, India
vv@iiit.ac.in

ABSTRACT

In this paper, we present a two-step language-independent spelling suggestion system. In the first step, candidate suggestions are generated using an Information Retrieval(IR) approach. In step two, candidate suggestions are re-ranked using a new string similarity measure that uses the length of the longest common substrings occurring at the beginning and end of the words. We obtained very impressive results by reranking candidate suggestions using the new similarity measure. The accuracy of first suggestion is 92.3%, 90.0% and 83.5% for Dutch, Danish and Bulgarian language datasets respectively.

Categories and Subject Descriptors

I.7.1 [**Document and Text Editing**]: Languages, Spelling

General Terms

Algorithms, Design, Languages

Keywords

Spelling suggestion, Information retrieval, Language independent

1. INTRODUCTION

Brill[1] reports roughly 10-15% of search engine queries contain errors. Popular search engines like Google, Yahoo provide suggestions for misspelled queries. This conveys spelling suggestion is a vital feature for web search engines. However, many of the spelling suggestion techniques today depend on query logs. Query logs carry a lot of implicit and explicit information and the main motivation for using query logs is based on the notion that, "*People who can't spell a term correctly can be helped by people who spell it correctly*". Experiments have also demonstrated that query logs can be useful while suggesting misspelled words. However, not all search engines have the luxury of using query logs because their query log might not be large enough. Insufficient query logs cannot be an excuse for absence/poor behaviour of spelling suggestion.

In this paper, we present a language independent spelling suggestion component for a web search engine called Setooz[1], and is built using the web document collection alone and requires no training data or human intervention.

2. CANDIDATE LIST GENERATION

We use IR technique for retrieving candidate suggestion words when queried with a misspelt word. Here our documents are words and each word is indexed using the ngrams of the word whose length varies from two to five. For example, the word *October* is indexed using the following ngrams: *oc, ct, to, ob, be, er, oct, cto, tob, obe, ber, octo, ctob, tobe, ober, octob, ctobe* and *tober*. In addition, while searching, a misspelled word is decomposed into distinct ngrams of length varying from 2 to 5, that are used to search the index.

Tf-idf[3] is one the most popular weighting function used in IR. The basic motivation for usage of an *idf* factor is that terms that appear in many documents are not very useful for distinguishing a relevant document from a non-relevant document. Unlike in classical IR, where each query term(word) carries a semantic meaning, query terms in our case are just a sequence of characters(ngrams) and *df* helps in determining the validity of the sequence. Hence, in our approach *df* is used to compute the similarity instead of *idf*. In addition, we use the frequency of word in document collection in computing the similarity between candidate suggestion and misspelled word. To prevent the bias towards word frequency, the overall score is normalized using the Levenshtein distance(LD). Moreover, each query term is given a weight proportional to number of characters in it. Finally, the similarity between a misspelled word M and a candidate suggestion S_j, $sim(S_j, M)$ is computed using the formula:

$$sim(S_j, M) = \frac{b(S_j) * \sum_{t_i \in M'} tf(t_i, S_j') * df(t_i) * b(t_i)}{LD(S, M)} \qquad (1)$$

where $M^{'}$ and $S_j^{'}$ are ngram representations of M, S_j respectively, $b(S_j)$ is natural log of word S_j's frequency in document collection, $tf(t_i, S_j^{'})$ is frequency of ngram t_i in $S_j^{'}$, $df(t_i)$ is natural log of ngram t_i's frequency in vocabulary and $b(t_i)$ is length of ngram t_i.

3. RERANKING

Spelling mistakes are often encountered in the middle of the word and are rare at the tails(beginning or end) of a word. Following this intuition which was also demonstrated

WWW 2010, April 26–30, 2010, Raleigh, North Carolina, USA.
ACM 978-1-60558-799-8/10/04.

[1] http://www.setooz.com

in study [2], we developed a new string similarity measure called Tail Similarity(TSim). TSim is computed using the length of "Longest Common Substring"(LCS) at both tail ends of misspelled word and candidate suggestion word.

Assuming l_1 and l_2 to be lengths of LCS from begining and ending of the two words $\{W_1, W_2\}$, TSim between W_1 and W_2 can be computed using the formula:

$$\text{TSim}(W_1, W_2) = \frac{1}{4.0} * \begin{cases} 2.0 + 2.0 & \text{for } l_1 = 0 \text{ and } l_2 = 0 \\ 1/l_1 + 2.0 & \text{for } l_1 \neq 0 \text{ and } l_2 = 0 \\ 2.0 + 1/l_2 & \text{for } l_1 = 0 \text{ and } l_2 \neq 0 \\ 1/l_1 + 1/l_2 & \text{otherwise} \end{cases}$$

Whenever the common substring does not exist, the word is penalized with a score of 2.0 which is greater than worst possible score of 1.0(when the length of LCS is 1). The final score is normalized with 4.0, to ensure that TSim score lies between 0 and 1, which is later used for re-ranking the candidate suggestions. Similarity between two strings is inversely proportional to TSim score i.e., lower the TSim higher the similarity between two strings and vice versa.

$$\text{Example :}\quad \text{TSim}(\underline{\text{adel}}\underline{\text{ijk}}, \underline{\text{adel}}\text{l}\underline{\text{ijk}}) = (\tfrac{1}{4} + \tfrac{1}{3})/4.0 = 0.14$$

The candidate suggestions obtained using IR approach are reranked using the formula:

$$sim(S_j, M) * (1 - TSim(S_j, M)) \qquad (2)$$

4. EVALUATION

We have chosen three languages belonging to three different regions and two different scripts, to evaluate our system. Document collections of languages chosen are taken from documents collected by the Setooz crawler over a duration of six months. On analysing a random set of words from the document collection, we observed that frequency of most invalid words[2] is less than 50. Hence words with frequency greater than 50 are chosen to construct spelling suggestion index as described in section 2.

Table 1: Statistics of the languages chosen

Language	Documents	Vocabulary	Misspelled words
Bulgarian	21,264,379	1,807,536	321
Danish	5,802,076	596,797	231
Dutch	23,202,641	1,201,148	664

Misspelled words and their judgement(correct word) pairs to evaluate our approach are picked from Wikipedia, which has a list of commonly misspelled words for some languages. Wilbur et.al[4] report that 87% of misspelled words in their query logs are the result of a *single edit* mistake and also demonstrate that shorter words are very difficult to correct. Hence, we chose misspelled words that are of at least six characters and are due to single edit mistakes. Statistics about the three languages, i.e., number of documents in the web collection, size of vocabulary and number of misspelled words used for evaluation are summarized in Table 1.

The performance of our system is evaluated using accuracy as a metric and is defined as percentage of correct suggestions out of TopN candidate suggestions. Performance of various approaches and the baseline system GNU Aspell[3] are presentend in Table 2. It is evident from Table 2 that, *tf-df* outperforms *tf-idf* at any point, supporting our intuition of using $df(t_i)$ instead of $idf(t_i)$. Using frequency of word and length of query(ngram) term in modified IR approach(MIR) helped in achieving a better recall(more correct suggestions). Reranking the MIR candidate suggestions using TSim improved the precision of final system(FS) and has also outperformed Aspell for all three languages.

Table 2: Comparison of various approaches. TF-IDF, TF-DF refer to traditional *tf-idf* and *tf-df* ranking techniques respectively. MIR refers to the modified IR approach as described in section 2, FS is the final system.

Language	Approach	Top1	Top2	Top3	Top4	Top5
Bulgarian	TF-IDF	17.4	25.5	31.4	36.4	39.5
	TF-DF	20.2	29.2	36.4	41.7	45.5
	MIR	71.6	80.6	88.7	90.9	94.0
	FS	**83.5**	**88.7**	**94.3**	**95.6**	**97.1**
	Aspell	76.6	78.8	94.3	95.0	95.9
Danish	TF-IDF	20.7	32.0	40.2	46.7	50.2
	TF-DF	29.4	43.2	52.8	60.1	66.2
	MIR	84.4	93.0	97.4	98.2	98.2
	FS	**90.0**	**96.5**	**98.7**	**98.7**	**99.1**
	Aspell	69.1	77.8	80.4	81.4	82.1
Dutch	TF-IDF	12.5	22.2	30.7	36.2	40.8
	TF-DF	20.0	32.3	43.6	49.3	54.9
	MIR	85.8	94.2	96.9	98.6	99.2
	FS	**92.3**	**97.8**	**99.0**	**99.3**	**99.6**
	Aspell	65.3	68.9	84.6	87.0	89.3

5. CONCLUSION AND FUTURE WORK

In this paper we show that IR techniques can be used to address problems like spelling suggestions and demonstrate that *tf-df* performs much better than *tf-idf* in such context. We also show that using frequency of word in document collection helps in having better recall. Moreover, to improve the precision we used a novel string similarity measure called Tail Similarity. We show that our final system outperforms GNU Aspell for all three languages chosen for evaluation.

Having observed some spelling mistakes due to missing or misplaced vowels, we would like to explore if a little information about language, like vowels, helps in improving Tail Similarity measure.

6. REFERENCES

[1] S. Cucerzan and E. Brill. Spelling correction as an iterative process that exploits the collective knowledge of web users. In *Proceedings of the Conference on EMNLP*, 2004.

[2] S. J. R. Schiller N O, Greenhall J A and C. A. Serial order effects in spelling errors: evidence from two dysgraphic patients. *Neurocase*, 7:1–14, 2001.

[3] K. Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. pages 132–142, 1988.

[4] W. J. Wilbur, W. Kim, and N. Xie. Spelling correction in the pubmed search engine. *Inf. Retr.*, 9(5):543–564, 2006.

[2] words with no semantic meaning

[3] http://aspell.net/, a free spell checker currently available for 70 languages