A Non-Learning Approach to Spelling Correction in Web Queries

Jason Soo
Department of Computer Science
Georgetown University
jjs246@georgetown.edu

ABSTRACT

We describe an adverse environment spelling correction algorithm, known as Segments. Segments is language and domain independent and does not require any training data. We evaluate Segments' correction rate of transcription errors in web query logs with the state-of-the-art learning approach. We show that in environments where learning approaches are not applicable, such as multilingual documents, Segments has an F1-score within 0.005 of the learning approach.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Algorithms, Languages

Keywords

Learning vs non-learning; web spelling correction

1. INTRODUCTION

Segments is a spelling correction algorithm for use without available training data. Examples include multilingual document web searches. In this paper, we show that Segments achieves roughly similar results to the state-of-the-art learning algorithm using web query logs.

We enhance the Segments approach to aid in correcting transcription errors from online queries. Transcription errors are those where characters are inserted, deleted, swapped, or replaced. We compare Segments to a recent state-of-the-art learning-based spelling correction algorithm. We demonstrate that Segments – which has no dependence on language, training data or domain – achieves similar results.

2. RELATED WORK

There are many proposed algorithms for spelling correction. Approaches such as Soundex and D-M Soundex provide phonetic solutions, but have many problems [3]. Recent research has shown that an advanced n-grams approach can perform quite well [1], but n-grams does not consider non-sequential windows of characters less than n. Recent learning work has shown great promise [2]. As the work of Li et. al. provides the most recent and advanced work in this field, we evaluate the ability of Segments to correct transcription errors in comparison to Li et. al.'s work.

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

3. SEGMENTS

Segments is a hybrid approach that uses n-grams and a series of sub-string generation rules. n-grams is a powerful approach that requires no language dependencies, but suffers from problems associated with sequential windows. The inclusion of the sub-string generation rules prevents the shortcomings of sequential windows.

3.1 Algorithm

The Segments process works as follows:

For each word in the query, the following process is repeated: Search the lexicon for an exact match, and if found, return the result immediately. Otherwise, six sub-string generation rules are applied to the query term. After each iteration the newly generated query term is used to search the lexicon. The rules are as follows:

1. Replace the first and last character with a wildcard. Search and repeat.
2. Replace the middle character with a wildcard. Search and repeat.
3. Replace the first half of the query with a wildcard.
4. Replace the second half of the query with a wildcard.
5. Replace all but the first and last letter with a wildcard.
6. Replace all but the first two and last two letters with a wildcard.

These rules are demonstrated in Table 1. When a rule finds a candidate term, it casts a vote for said candidate. Votes can be weighted according to which rules find the candidate, and in which iteration.

When the rules have exhausted the possible sub-strings, or a threshold of searches is reached, we union the candidate sets. If the rules maintain a sufficient confidence – where confidence is computed using

$$\max_{i \in C} \left\{ \frac{\sum \text{votes for candidate } i}{\text{total votes}} \right\}$$

where C is the set of candidates being suggested – n-grams is not run. Otherwise, vanilla n-grams is used. In our testing, we set the sufficient confidence threshold to be ≥ 0.3. The

Rule #	Search Candidates		
1	%ississipp%	%ssissip%	%sissi%
2	Missi%sippi	Miss%ippi	Mis%ppi
3	%ssippi	-	-
4	Missis%	-	-
5	M%i	-	-
6	Mi%pi	-	-

Table 1: Sub-string generation rules applied to *Mississippi*. Each column is an rule iteration.

Algorithm	All	CF	CD
Segments	0.526	0.564	0.333
Li	0.531	N/A	N/A

Table 2: F1-score query results. CF is context-free queries. CD is context-dependent queries. All is the total F1-score.

confidence of the n-grams candidates is then compared to the confidence of the rules, and the candidates set with the higher confidence is then returned.

We then consider the permutation of candidate query strings to return to the user. The permutation of candidate terms with the highest confidence is returned to the user.

This approach is customized for large lexicons and query logs, and operates in contrast to earlier evolutions of the Segments system [5, 4].

4. EVALUATION

Our evaluation goal is to – as best possible – compare the F1-score of Segments to Li et. al. We focus on transcription errors in part because Li et. al., in order to solve for other types of errors, leverages prior work by the Microsoft Web N-gram Services[1]. This approach does not work for multilingual document searching, or other adverse environments.

4.1 Data Set

We start with the same data set and lexicon as Li et. al. As they supplement their algorithm with knowledge of bigrams, so too do we, but with a different source[2] due to availability. The web query logs are released by Microsoft Research and are an annotated version of the TREC 2008 Million Query Track.[3] They contain suggested corrections for the 311 misspelled queries. We focus on a subset of the misspellings that are a) transcription error based and b) available within our lexicon. This leaves 93 usable queries.

Note this is not an exact match with the Li. et. al. data set, but rather a subset, due to our restrictions. We believe these restrictions are reasonable because 1) we are focused on transcription errors and 2) omission of a term from our lexicon guarantees Segments *cannot* correct it, whereas inclusion only guarantees Segments *could* correct it.

4.2 Metrics

We report the F1-score of the results from both research efforts. Since Segments only returns the most confident suggested query string to the user, precision and recall are ignored.

[1]http://research.microsoft.com/en-us/collaboration/focus/cs/web-ngram.aspx
[2]http://www.ngrams.info/download_coca.asp
[3]http://web-ngram.research.microsoft.com/spellerchallenge/Datasets.aspx

5. RESULTS

Using the queries from the Microsoft Research data set, we attempted to correct the 93 web queries which contain transcription errors. Table 2 shows these results. We measure our performance at three different levels. Each column in the table represents an F1 score. The *All* column is the score for all transcription error queries. Similarly, the *CF* and *CD* columns represents the scores for context-free and context-dependent queries, respectively.

The goal of this research was to compare a non-learning approach to a learning approach for correcting transcription errors in web query logs. Intuitively, the learning approach does outperform Segments. Interestingly though, it only has an F1-score of 0.005 higher. This clearly demonstrates that in the case of transcription errors, which account for nearly 66% of the spelling errors in the data set, Segments performs almost equally to the advanced learning approach.

Furthermore, we see Segments does considerably better in the cases where context is not required to correct the query. Li et. al. did not report specific results for context-based queries. For example, take the context-dependent query *capital hill*. While *capital* is a word and is spelled correctly, it is misspelled in the context of Washington D.C.'s *capitol hill*. While leveraging bigrams allows for Segments to account for some context-dependent cases, this is clearly an area where learning would provide improvement. However, when a learning approach is not an option, Segments provides nearly the same performance.

6. CONCLUSION

Spelling correction plays a crucial role in the outcome of online queries today. However, the ability to train a spelling correction approach is not always an option, like in multilingual documents. In this paper, we show our novel non-learning approach provides results that are nearly identical in F1-score to that of the state-of-the-art learning approach in tasks of correcting transcription errors.

7. REFERENCES

[1] Duan and Hsu. "Online spelling correction for query completion." Proceedings of the 20th international conference on World wide web. ACM, 2011.

[2] Li, Duan, and Zhai. "A generalized hidden Markov model with discriminative training for query spelling correction." Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval. ACM, 2012.

[3] Patman and Shaefer, "Is Soundex Good Enough for You? On the Hidden Risks of Soundex-Based Name Searching", Language Analysis Systems, Inc., Herndon, VA, 2003.

[4] Soo, Cathey, O. Frieder, Amir, G. Frieder. "Yizkor books: a voice for the silent past." Proceedings of the 17th ACM conference on Information and knowledge management. ACM, 2008.

[5] Soo and Frieder. "On foreign name search." 32nd European Conference on IR Research, ECIR 2010, Milton Keynes, UK, March 28-31, 2010.