English Language Spelling Correction as an Information Retrieval Task Using Wikipedia Search Statistics

Dr Kyle Goslin, Dr. Markus Hofmann

TU Dublin – Blanchardstown Campus, Dublin, Ireland.

Abstract

Spelling correction utilities have become commonplace during the writing process, however, many spelling correction utilities suffer due to the size and quality of dictionaries available to aid correction. Many terms, acronyms, and morphological variations of terms are often missing, leaving potential spelling errors unidentified and potentially uncorrected. This research describes the implementation of WikiSpell, a dynamic spelling correction tool that relies on the Wikipedia dataset search API functionality as the sole source of knowledge to aid misspelled term identification and automatic replacement. Instead of a traditional matching process to select candidate replacement terms, the replacement process is treated as a natural language information retrieval process harnessing wildcard string matching and search result statistics.

The aims of this research include: 1) the implementation of a spelling correction algorithm that utilizes the wildcard operators in the Wikipedia dataset search API, 2) a review of the current spell correction tools and approaches being utilized, and 3) testing and validation of the developed algorithm against the benchmark spelling correction tool, Hunspell. The key contribution of this research is a robust, dynamic information retrieval-based spelling correction algorithm that does not require prior training. Results of this research show that the proposed spelling correction algorithm, WikiSpell, achieved comparable results to an industry-standard spelling correction algorithm, Hunspell.

Related Work

Spelling errors can exist in many forms. Examples of this include terms that have obvious misspellings (Gupta et al., 2019), terms that are correct although in the wrong context (Mays et al., 1999), and phonetic spelling errors that are typically performed by children (O'Neill et al., 2020). Spelling correction algorithms are highly dependent on having a quality source of language-specific corpora to utilize to function correctly, which for less widely used languages becomes an issue as they often do not exist in a complete form (Etoori et al., 2018). As a solution to this, spelling correction algorithms have been created that utilize dynamic corpora such as Wikipedia for many languages as the list of correct terms is never complete and always growing (Beeksma et al., 2018). In addition to this, grammatical error and correction corpora have also been generated for spelling correction tasks (Grundkiewicz and Junczys-Dowmunt, 2018).

There is no current standard approach to spelling identification and correction, leaving modern approaches to adopt a variety of implementations including language models, term co-occurrence statistics, and machine learning.

Implementation of WikiSpell

This section describes the design and implementation of the proposed algorithm, WikiSpell, that utilizes the Wikipedia Search API statistics and text results. The Wikipedia Search API results are used as a source of terms related to a single term in the original sequence of terms under analysis. For any given search result the article title and snippet were used as sources of terms for a given article. The summary of text is typically 25 terms. The number of results to return was set using *srlimit=100* parameter to limit to a max of 100 results. To avoid overfitting of terms, for any sequence of terms they are tokenized and submitted separately broadening the results collection.

In this algorithm, two core metrics for relevance are used. The first of these is the search result statistics. Given a string Q, the number of records returned relates to the number of Wikipedia articles where the string Q was found. The total hits were returned using *srinfo=totalhits* API parameter in the Search API. This statistic was used to:

- Identify that a term has been used before in the English Wikipedia corpus. This also serves to show if the spelling is correct as the number will be higher than those incorrectly spelled in the English language corpus.
- Identify if a sequence of terms has together been seen in context where the length of a sequence is > 1. The more frequent the use, the higher the chance the sequence is correct.

The hypothesis of this metric is that terms shown in context together provide an indicator of being semantically correct and a viable replacement for incorrect term sequences. Terms that are not frequently shown together would indicate that terms are semantically not a good fit. In addition to this, the Levenshtein distance was also used for small sets of terms to identify if a given term is a good replacement for a candidate term. For a collection of terms {T}, the lower the score for a single term Tn indicates that the term is a closer match.

The algorithm consists of two core processes, one to generate the candidate corpus from the Wikipedia search API using wildcards, and one to select replacement terms.

For the first process, the algorithm is responsible for generating four different permutations of the base term to search the Wikipedia search API. For each different character position in the term t between 0 and the length of t the following operations are performed:

- INSERT - A search wildcard is added which is represented by a single asterisk.
- INSERT - Two asterisk characters are inserted.
- REPLACE - A single character inside the string is replaced by a single asterisk.
- REMOVE - A single character is removed from the original string.

For the second process, to apply a relevance filter to the terms, a second loop is used to iterate over the terms stored. For each individual term, the Levenshtein distance is applied between the original query term t and the current term tn. Each of the terms and associated distance scores are appended to the *simvals* array. The distances are then stored in descending order. The top candidate replacements are then selected.

Methodology

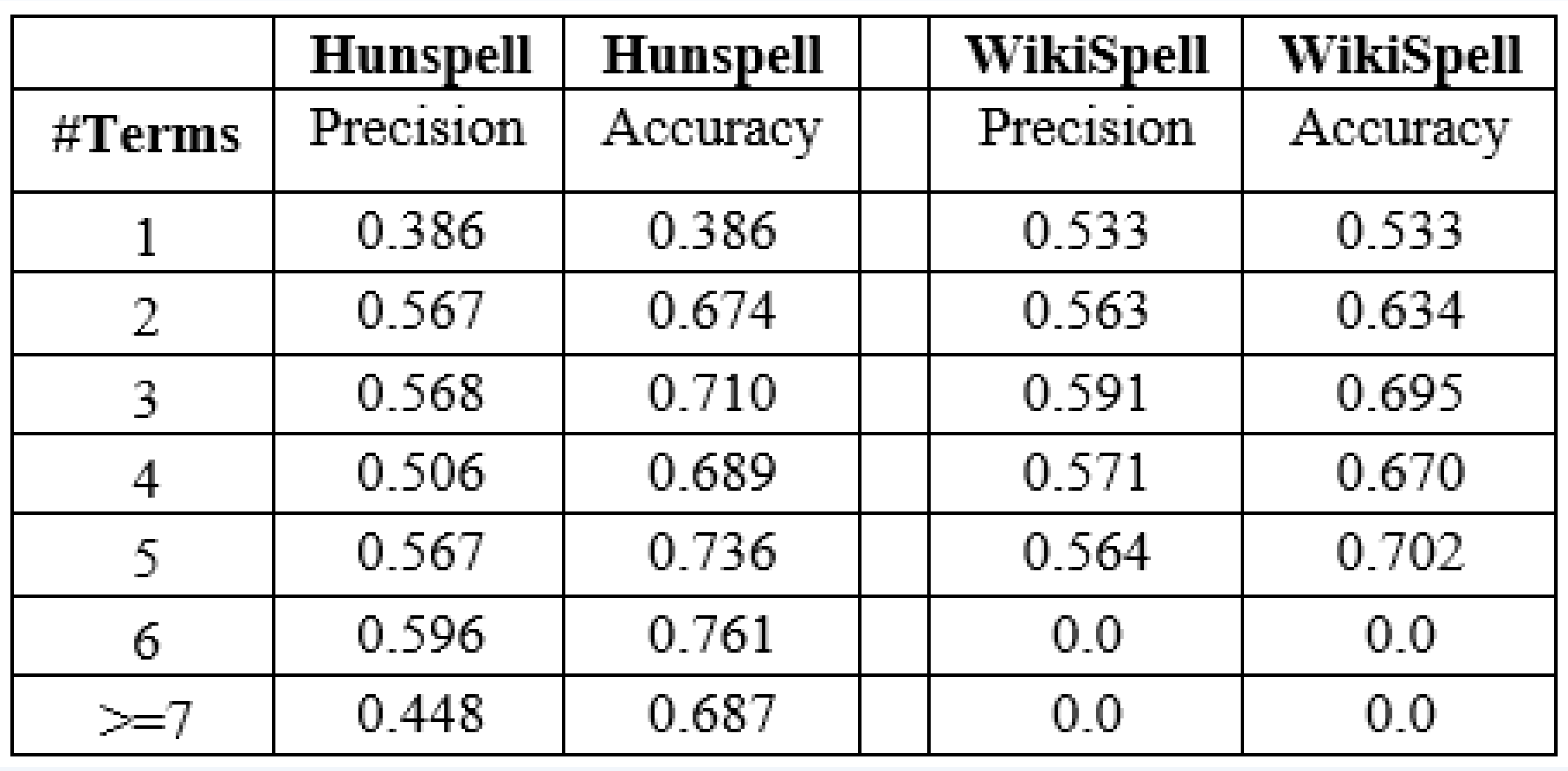
To test both Hunspell and the proposed algorithm, a synthetic test set of terms was designed to implement common spelling errors. These errors included the addition of one or two individual random characters for each term at any position in the string. These errors were introduced into the TREC 2009 Million Query Track 20001-60000 data set that contains 2,000 queries of varying lengths.

This approach provided both an original query as a reference point and a modified error version of the query for testing. From this set, queries 20001 – 22001 were utilized. These queries contain common user queries to a search engine that include abbreviations, acronyms, and common phrases on a wide variety of topics.

The default en_US dictionary for Hunspell was utilized for testing, as used in previous studies. The first replacement term from the suggested list of terms for an identified misspelling was utilized. For the proposed algorithm the number of records during consideration was set to 5, the number of search results records included (max) = 100 and the threshold for relevance = 500.

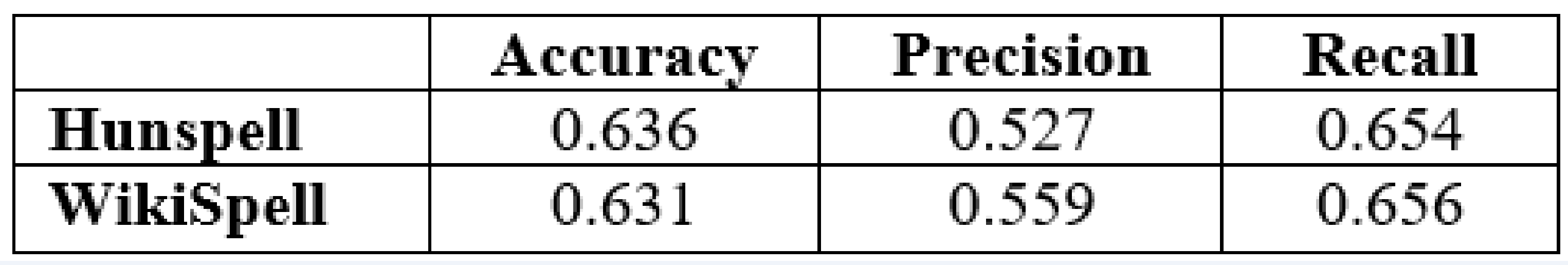
Results

Table 2 shows the precision and accuracy results from the 2000 queries with spelling errors that were corrected. They are shown as the term length of the original query. Table 3 shows the average results for all runs.



	Hunspell	Hunspell		WikiSpell	WikiSpell
#Terms	Precision	Accuracy		Precision	Accuracy
1	0.386	0.386		0.533	0.533
2	0.567	0.674		0.563	0.634
3	0.568	0.710		0.591	0.695
4	0.506	0.689		0.571	0.670
5	0.567	0.736		0.564	0.702
6	0.596	0.761		0.0	0.0
>=7	0.448	0.687		0.0	0.0

Table 2: Precision and accuracy for each different query length.



	Accuracy	Precision	Recall
Hunspell	0.636	0.527	0.654
WikiSpell	0.631	0.559	0.656

Table 3: Overall average accuracy, precision and recall from 2,000 results.

Conclusion

This research proposed the WikiSpell algorithm for the automatic detection and correction of spelling errors in a sequence of terms. Results from this research have shown that the Wikipedia Search API has shown to be effective as a source of candidate terms for spelling correction due to the variety of topics covered in the English Wikipedia. The utilization of Wikipedia English corpus search statistics such as the number of search results as a source for identifying term sequences has shown to be useful for spelling error identification and replacement.

The algorithm proposed in this research, WikiSpell, has shown comparable results to the Hunspell algorithm when used for automatic spelling identification and correction. When working with short sequences of terms, an IR-based approach to spelling correction was highly successful due to the dependency on context.

The sample code and dataset for this project is available on GitHub.

Contact

E-mail: Kyle.Goslin@tudublin.ie