TAF

Journal of Advances in Humanities and Social Sciences

JAHSS

2020, **6**(3): 100-104

PRIMARY RESEARCH

Text correction algorithms for correct grammar and lexical errors in the English language

Abbas Hussein Tarish [1*], Humam K. Majeed Al-Chalabi [2]

[1] Field of Philology, West University of Timisoara, Timişoara, Romania

[2] Faculty of Automatics Computers and Electronics, University of Craiova, Craiova, Romania

Keywords

Algorithms
Text correction algorithms
Correct grammar
Lexical errors

Received: 11 January 2020
Accepted: 17 March 2020
Published: 29 September 2020

Abstract

The main objective of this study is to focus precisely on the discussion of the English language algorithms in the direction of grammar and lexical correction errors algorithms. The utilization of the language web-search engine has exploded in its popularity and its potential applications. The majority of search engine algorithms use error detection, which typically is not limited to a particular native language. Thus, the syntax of English web search queries, the problem was identified by automatic speech recognition systems. In an algorithm that takes on corrections for second language learners, a new approach was introduced. In this study, the algorithm proposed will take an input sentence with a preposition error and replaces it with the correct preposition that would relate to this specific sentence. This is based upon a rule-based and statistical approach, making it a hybrid made up of two different phases. This proposed algorithm will be useful for students, educators, and scholars.

INTRODUCTION

The value of search results has become a commodity, as the ability and ease of finding information, services, and locations has increased in importance over time. To date, web search engines are now becoming one of the most popular tools for identifying and researching information online (Stefanowski & Weiss, 2003). Authorities in the field have begun to understand that the search for data on the internet is in need of some guidance, which creates a problem concerning how a query will be able to retrieve documents or a subset of documents that are relevant to the inquiry (Weiss, 2001). This seems to bring to question how it is possible to find any relevant information about a particular topic online (Fletcher, 2004; Weiss, 2001). Automatic grammar learning is one of the technologies that have been utilized in aiding the ability for web search engines to identify lexicons of natural language, in addition to developing more accurate grammar (Copestake & Flickinger, 2000; Lee, Chang, & Hsieh, 2014; Ramanauskaite & Vaisnys, 2017). Even while there are automatic grammar learning paradigms being employed by software, still many of the computer applications are partially hand-built grammar or require manual work. A majority of researchers are now attempting to build more accurate algorithms that are less manual, more automated, and far more proficient in its implications (Porter, 2001). Many of the computational linguistics learning problems are related to research issues based on sentences, phrases, as well as words. Researchers want to resolve how to take in certain information and create an effective output system that will behave precisely on specified functions (Heinz, De la Higuera, & Van Zaanen, 2015; Heift & Nicholson, 2001). Many researchers attempt to improve algorithm enhancements for the overall progression of more accurate web search results (Karwa & Honmane, 2019; Medhat, Hassan, & Korashy, 2014; Sheela & Jayakumar, 2019). Web search engines allow for referrals, which make commercial websites easier to discover (Ntoulas, Najork, Manasse, & Fetterly, 2006). The end result is a boost in revenue and

*corresponding author: Abbas Hussein Tarish
†email: abbasamradu@yahoo.com

sales, and ultimately more profits. Since 2004, 1.9% of total US sales have been the outcome of web-based e-Commerce, which continues to grow at a rate of 7.8a% yearly (Ntoulas et al., 2006).

Research Structure

The research structure designed to discuss the linguistic structure of English web-search engine and what is the problems that faced. In the first part explains the main idea in general, in addition the general overview of the recent studies. The second part, we try to explain the English web-search queries to be clearer. The third part identifies the main problem that faced the web-search queries. The forth part discuss the solutions that have been achieved. Ultimately, in the last part, a discussion was given about what most of the established spell checking systems need.

THE LINGUISTIC STRUCTURE OF ENGLISH WEB-SEARCH QUERIES

Generally speaking, most web-search queries are short, and average about 2.8 words (Whitelaw, Hutchinson, Chung, & Ellis, 2009). They are constantly reformulated, and change with time, however not many people are able to understand the grammatical structure of theses search terms. Interpreting search queries is a crucial part of the information retrieval tasks. Identifying a grammatical structure of pattern for these search queries would vastly aid in developing a more beneficial information retrieval task system (Whitelaw et al., 2009). In order to disambiguate these search term queries when retrieving specific documents, one approach is the utilization of the part-of-speech tagging (Skoric & Kupresanin, 2018; Whitelaw et al., 2009). It aids in enhancing recall through query reformulation, and applies the use of part-of-speech tags to replace synonyms for words found in content, as well as part-of-speech tag for well-formed questions.

These short queries are often ambiguous due to the differences in documents and their relevance to the part of the speech that is applied in search term queries. A single word can be a verb, a verb object, or the subject of a verb, where the intention of the word can have completely different outcomes (Barr, Jones, & Regelson, 2008). The ability to discern between the different intentions for word choices, some researchers proposed that observing the feedback from the user is necessary, by providing them different contexts for the search query terms.

In a study on the creation of a more accurate web search query system, researchers designed a syntactic parsing system. These applied rules that were sourced from the observed linguistic structure of search term queries, rather than natural language corpora (Barr et al., 2008). If the part-of-speech distribution and syntactic structure for search term queries are over tagged within indexed documents, this system creates an approach which can help resolve this problem applying the syntactic parsing system. This occurs through the use of a simple bijection mapping per categorically constructed query tags which relate to other tag sets.

The Automatic Speech Recognition System

In another approach to text correction algorithms, the automatic speech recognition system is thought to be one of the fastest evolving computing fields (Bassil & Alwani, 2012; Ferraresi, Bernardini, Picci, & Baroni, 2010; Shoeleh, Zahedi, & Farhoodi, 2017). This recognition system is applied for a huge number of uses, such as speech dictation systems, automatic speech-to-text systems, speech-driven home automation systems, voice user interfaces, voice-driven industrial control systems, and automated telephone services (Ferraresi et al., 2010; Gatpandan & Ambat, 2017). This is increasing in importance as there is a broadening amount of voice-search activities occurring over time as more individuals are using voice-to-text tools and mobile phones to search for information or locations.

The Problem

Unfortunately, automatic speech recognition systems are still plagued with errors, and can be incredibly inaccurate due to the fact that they are often utilized in the wrong environments (Bassil & Alwani, 2012). These are mainly due to lexical misspellings, linguistic mistakes found within the output text, and also the atmosphere which may have excessive noise within the environment. The noise is not the only problem, but also the dialect and how words are uttered, the speech quality, as well as the system's available vocabulary (Bassil & Alwani, 2012). Some new forms of error correction techniques were created for the purpose of reconfiguring how text is translated from voice scripts in order to improve the accuracy level. This includes some manual post-editing done on output transcripts where individuals correct misspellings, as well as the creation of acoustic mathematical models which help to enhance how the input waveforms are interpreted for the purpose of error-prevention.

The Solutions

To resolve this issue, the creation of automatic post-editing context-based real word error correction was invented (Bassil & Alwani, 2012). This was manifested from Bing's web search engine and its spelling technology for the pur-

ISSN: 2414-3111
DOI: 10.20474/jahss-6.3.2

pose of correcting lexical and linguistic errors that occur as a result of the automatic speech recognition system. These systems are based upon two main types of lexicons, one that is probabilistic of an n-gram collocation, and the other a phonetic of static pronunciations (Bassil & Alwani, 2012; Sidorov, 2013). With a larger vocabulary of lexicons, there are more accurate results with less errors found within the audio recognition process.

A majority of text-correction algorithms function by accumulating common errors, as well as their correct forms, in a practical way (Nelken & Yamangil, 2008). This also aids in creating models for theoretical linguistic studies. However, this isn't enough to prevent errors in grammar checking, in addition to context-sensitive spelling corrections. In a research study on how to improve these processes, researchers were able to temporally compare adjacent types of an article, and observe how users made corrections within these articles (Nelken & Yamangil, 2008). This approach was focused on correcting lexical errors that contain some semantic coherence and also has phonetic similarity. These can be mined automatically through observing Wikipedia's revision history, and can have future importance for the development of algorithms for linguist researchers (Nelken & Yamangil, 2008).

DISCUSSION

A majority of spelling systems these days need manually created language-specific resources. This includes rule bases, lexical, and a compiled list of common misspellings (Whitelaw et al., 2009). For those systems which apply the use of statistical models, an even larger compilation of annotated corpora of spelling errors is necessary for the training and machine learning. One group of researchers created statistical models which don't require any annotated data (Whitelaw et al., 2009). They rely mainly on the internet to provide large amounts of corpus data in order to help them observe frequently used terms for likely possible correction, for information dealing with misspellings based on terms used on the web, and token n-grams which are applied for the creation of a lexicon list to make context-appropriate corrections. The error model proposed in this study is developed from scored sub-strings where no fixed lexicon of correctly spelled words is used to determine misspellings. Properly spelled and misspelled words are allowed, which makes it unique from other systems. The n-gram application can also identify and correct real world substitutes for words as well.

In an algorithm which takes on corrections for second language learners, a new approach was introduced. In this study, the algorithm will take an input sentence with a preposition error and replaces it with the correct preposition that would relate to this specific sentence (Hermet, Desilets, & Szpakowicz, 2008). This is based upon a rule based and statistical approach, making it a hybrid made up of two different phases. The first phase consists of the rule-based approach, where the terms are processed in order to create a short expression within the context of the input sentence and its corresponding preposition. The second phase includes web searches for the exploration of the frequency that this expression occurs at, as well as the varieties of prepositions that are utilized as opposed to the known ones. Another approach incorporates the use of a multi-lingual semantic wiki which is based on controlling English and grammatical frameworks. In this study, a grammatical framework is investigated for its ability to program language in the application of building multilingual grammar (Kaljurand & Kuhn, 2013; Mala & Lobiyal, 2016). Each grammatical framework is based upon an abstract syntax, or a set of functions and their relative categories, as well as a set of concrete syntaxes which help to explain how the abstract categories and functions become linear within each language. There is a mapping between these concrete language strings, as well as their associated abstract trees made up of function name structures. The strings are developed into trees, and the trees can be linearized as strings, and mapped such that the abstract syntax can be automatically translated between one another. This aids in providing for more accurate systems to parse languages with the ability to manage more natural language features, such as long-distance dependencies, agreements, and morphological variations (Kaljurand & Kuhn, 2013). Similarly, the portable grammar format is applied in the use of natural language processing within web services (Barr et al., 2008). Developing these algorithms aid in improving grammar correction, and can ultimately build more impactful and interactive natural language web applications (Bringert, Angelov, & Ranta, 2009).

The creation of a statistical model can be used in a way which exploits the patterns of use within language. The collection of frequencies by which a word is utilized, and the word order in context, can help train an algorithm to identify the true category of the word in question (Heilman, Collins-Thompson, Callan, & Eskenazi, 2007). In a study on the creation of statistical language modeling, a statistical model was built for each predicted grade level. This approach is beenficial in that it provides better accuracy for web documents, as well as short passages (Heilman et al., 2007). This is different from the traditional readability for-

ISSN: 2414-3111
DOI: 10.20474/jahss-6.3.2

mulas which are based on a linear regression, which only includes two to three variables at a time. Additionally, unlike traditional language models, this one creates a probability distribution across the different grade model levels and not just a single prediction. Other benefits are the increase in information on the difficulty level of every word analyzed. This can be of great benefit in allowing more accurate vocabulary corrections (Heilman et al., 2007; Nava & Zubizarreta, 2009).

CONCLUSION

While search term recognition systems are increasing with popularity over time, the algorithms that are created for the purpose of parsing out web search inquiries are not up to par. Grammatical and text correction algorithms in the English language in general have a far way to come, as they have many contextual obstacles. For this reason, more research in this domain is encouraged. Specifically, improvised versions of existing algorithms must be introduced to cater for language translation and correction errors.

LIMITATIONS AND RECOMMENDATIONS

This study is limited in scope as an algorithm has been proposed but not developed and tested. Current research encourages scholar to design, implement and tests different algorithms which might solve the grammatical and text correction errors. Further research is necessary in order to help predict better algorithms for grammatical and lexical errors. Unfortunately, this can be complicated as grammatical errors can occur based on context, the way a noun or verb is used, and the intention behind its usage. Over time, more improved algorithms can be developed through experimentation and better prediction of word usage patterns.

REFERENCES

Barr, C., Jones, R., & Regelson, M. (2008). The linguistic structure of english web-search queries. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing,* Oxford, UK.

Bassil, Y., & Alwani, M. (2012). Post-editing error correction algorithm for speech recognition using bing spelling suggestion. *International Journal of Advanced Computer Science and Applications*, *3*(2), 34-45.

Bringert, B., Angelov, K., & Ranta, A. (2009). Grammatical framework web service. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Demonstrations Session,* New York, NY.

Copestake, A. A., & Flickinger, D. (2000). An open source grammar development environment and broad-coverage English grammar using HPSG. In *International Conference on Language Resources and Evaluation,* London, UK.

Ferraresi, A., Bernardini, S., Picci, G., & Baroni, M. (2010). *Web corpora for bilingual lexicography: A pilot study of english/french collocation extraction and translation.* Retrieved from https://bit.ly/33McSlp

Fletcher, W. H. (2004). Making the web more useful as a source for linguistic corpora. In, *Applied corpus linguistics*. Berlin, Germany: Brill Rodopi.

Gatpandan, P. H., & Ambat, S. C. (2017, jun). Implementing knowledge discovery in enhancing university student services portfolio management in higher education institutions. *Journal of Advanced Research in Social Sciences and Humanities*, *2*(3). doi:https://doi.org/10.26500/jarssh-02-2017-0306

Heift, T., & Nicholson, D. (2001). Web delivery of adaptive and interactive language tutoring. *International Journal of Artificial Intelligence in Education*, *12*(4), 310-325.

Heilman, M., Collins-Thompson, K., Callan, J., & Eskenazi, M. (2007). Combining lexical and grammatical features to improve readability measures for first and second language texts. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics,* California, CA.

Heinz, J., De la Higuera, C., & Van Zaanen, M. (2015). Grammatical inference for computational linguistics. *Synthesis Lectures on Human Language Technologies*, *8*(4), 1-139. doi:https://doi.org/10.2200/S00643ED1V01Y201504HLT028

Hermet, M., Desilets, A., & Szpakowicz, S. (2008). Using the web as a linguistic resource to automatically correct lexico-syntactic errors. In *Language Resources and Evaluation,* London, UK.

Kaljurand, K., & Kuhn, T. (2013). *A multilingual semantic wiki based on attempto controlled english and grammatical framework*. Berlin, Germany: Springer.

Karwa, R., & Honmane, V. (2019). Building search engine using machine learning technique. In *International Conference on Intelligent Computing and Control Systems (ICCS),* Bangkok, Thailand.

Lee, M. C., Chang, J. W., & Hsieh, T. C. (2014). A grammar-based semantic similarity algorithm for natural language sentences. *The Scientific World Journal*, *6*(8), 56-70. doi:https://doi.org/10.1155/2014/437162

ISSN: 2414-3111
DOI: 10.20474/jahss-6.3.2

Mala, V., & Lobiyal, D. K. (2016). Semantic and keyword based web techniques in information retrieval. In *International Conference on Computing, Communication and Automation (ICCCA),* Istanbul, Turkey.

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, *5*(4), 1093-1113. doi:https://doi.org/10.1016/j.asej.2014.04.011

Nava, E., & Zubizarreta, M. L. (2009). Order of l2 acquisition of prosodic prominence patterns: Evidence from L1 Spanish/L2 English speech. In *Proceedings of the 3rd Conference on Generative Approaches to Language Acquisition North America,* Florida, FL.

Nelken, R., & Yamangil, E. (2008). Mining wikipedia's article revision history for training computational linguistics algorithms. In *Proceedings of the AAAI Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy,* Seoul, South Korea.

Ntoulas, A., Najork, M., Manasse, M., & Fetterly, D. (2006). Detecting spam web pages through content analysis. In *Proceedings of the 15th International conference on World Wide Web,* Kulalumpur, Malaysia.

Porter, M. F. (2001). *Snowball: A language for stemming algorithms*. New York, NY: Sage Publications.

Ramanauskaite, E., & Vaisnys, J. R. (2017). Qualitative longitudinal research on lithuanian student migration. *Journal of Advances in Humanities and Social Sciences*, *3*(4), 193-205. doi:https://doi.org/10.20474/jahss-3.4.1

Sheela, A. C. S., & Jayakumar, C. (2019). Comparative study of syntactic search engine and semantic search engine: A survey. In *Fifth International Conference on Science Technology Engineering and Mathematics (ICONSTEM),* California, CA.

Shoeleh, F., Zahedi, M. S., & Farhoodi, M. (2017). Search engine pictures: Empirical analysis of a web search engine query log. In *International Conference on Web Research (ICWR),* Rome, Italy.

Sidorov, G. (2013). Syntactic dependency based n-grams in rule based automatic English as second language grammar correction. *International Journal of Computational Linguistics and Applications*, *4*(2), 169-188.

Skoric, J., & Kupresanin, J. (2018). Social work in educational system of the balkans – is social worker needeed in schools? *International Journal of Humanities, Arts and Social Sciences*, *4*(6), 245-252. doi:https://dx.doi.org/10.20469/ijhss.4.10003-6

Stefanowski, J., & Weiss, D. (2003). Carrot 2 and language properties in web search results clustering. In *International Atlantic Web Intelligence Conference,* Berlin, Heidelberg.

Weiss, D. (2001). *A clustering interface for web search results in Polish and English* (Unpublished master's thesis). Poznan University of Technology, Poznan, Poland.

Whitelaw, C., Hutchinson, B., Chung, G. Y., & Ellis, G. (2009). Using the web for language independent spellchecking and autocorrection. In *Proceedings oF Conference on Empirical Methods in Natural Language Processing,* New Dehli, India.

ISSN: 2414-3111
DOI: 10.20474/jahss-6.3.2