Towards Unsupervised Word Error Correction in Textual Big Data

Joao Paulo Carvalho and Sérgio Curto
INESC-ID, Instituto Superior Técnico, Universidade de Lisboa, Rua Alves Redol 9, Lisboa, Portugal

Keywords: Fuzzy Text Preprocessing, Medical Text Reports, Natural Language Processing, Word Similarity, MIMIC II.

Abstract: Large unedited technical textual databases might contain information that cannot be properly extracted using Natural Language Processing (NLP) tools due to the many existent word errors. A good example is the MIMIC II database, where medical text reports are a direct representation of experts' views on real time observable data. Such reports contain valuable information that can improve predictive medic decision making models based on physiological data, but have never been used with that goal so far. In this paper we propose a fuzzy based semi-automatic method to specifically address the large number of word errors contained in such databases that will allow the direct application of NLP techniques, such as Bag of Words, to the textual data.

1 INTRODUCTION

Since the invention of written language, textual information contained in documents has been the most commonly used form of expressing human knowledge. As such, textual information should be an important source for automatic knowledge representation. When the information contained in the texts has been properly edited, Natural Language Processing (NLP) tools can be used (with more or less success) to process it. However, in the case of unedited texts, such tools might not be reliable, since one of the most common NLP approaches consists in the use of the so-called "Bag-of-Words" model. This model essentially relies in word counts to extract information. Therefore, any word error, whether resulting from a typo, wrong transcription, or some cultural error, results on a model error (a misclassification, a miscount, etc.), that can have more or less serious consequences in what concerns knowledge representation.

In the present age of Big Data, this problem becomes very relevant, as can be exemplified by the MIMIC II database, a very large database of ICU patients admitted to the Beth Israel Deaconess Medical Center, that will be used here as a case study. The MIMIC II database includes both physiological and text data; However, the information contained in physicians' and nurses' text notes has never been used in any of the existing several models using the database (Cismondi, et. al, 2012; Fialho, et al. 2012; 2013), despite the fact that such notes contain rich information that is a direct representation of experts' views on the real time observable data.

This textual information has not been used before because of its size and the particularities of the documents:

- The reports are not structured as a typical written text – sentences are short, have many abbreviations, a reduced number of function words and most of the words are specific and relevant within the context;
- The reports have a large number of medical technical terms and specific technical abbreviations;
- There are many numerical values associated with physiological variables readings;
- Many different ways of expressing/representing the same information. E.g., dates (23-06-2014; 6/23/14; June, 23 2014, etc.), time (10:14PM; 22:04; 2204, etc.), etc.;
- Text contains a huge number of typographical and other word errors due to how the texts were collected (real time transcriptions from recordings; poor Optical Character Recognition of manuscripts; etc.);
- Text contains many other artifacts, such as misplaced control characters that break sentences into paragraphs, escape sequences,

Carvalho J. and Curto S..
Towards Unsupervised Word Error Correction in Textual Big Data.
DOI: 10.5220/0005140401810186
In *Proceedings of the International Conference on Fuzzy Computation Theory and Applications* (FCTA-2014), pages 181-186
ISBN: 978-989-758-053-6

etc.

These particularities prevent the effective use of common Natural Language Processing (NLP) techniques and hinder their use as “automatic information providers”, as shown in (Carvalho and Curto, 2014). Here we will focus in improving the process of automatic word error detection and correction in such databases.

Error detection and correction is something that everyone has become familiar with since the advent of word processors. Nowadays, most of the (computer / tablet / smartphone) written texts are automatically corrected in “real-time” (i.e., as they are being generated). However, everyone knows that, despite the large recent advances, not all corrections are proper, especially when less common words (e.g. technical terms, named entities, etc.) are being used. Also, most word correction tools are limited to one or two errors per word. The capability of humans to adapt very fast to new situations allows them to detect most unwanted corrections as they are proposed, and therefore react immediately. So, the problem of word error detection and consequent correction is basically non-existent when performed in “real time” (and as long as the used vocabulary is well known). However, if the texts have not been properly corrected while they were created, then we are facing a complex and expensive task that must usually be done manually, or, even if done automatically, demands a significant human intervention. This is especially relevant in unedited technical text. In the case of Big Data Text databases, this task must be somehow automated, since the size of the database would make the cost of manual offline text editing unbearably expensive.

In this paper we propose a fuzzy based semi-automatic method to specifically address the large number of word errors contained in Big Data unedited textual data, focusing specifically in the MIMIC II database.

2 THE MIMIC II DATABASE

The developed work uses data from the Multi-parameter Intelligent Monitoring for Intensive Care (MIMIC II) database (Saeed, 2002). This is a large database of ICU patients admitted to the Beth Israel Deaconess Medical Center, collected from 2001 to 2006, and that has been de-identified by removal of all Protected Health Information. The MIMIC II database is currently formed by 26,655 patients, of which 19,075 are adults (>15 years old at time admission). It includes high frequency sampled data of bedside monitors, clinical data (laboratory tests, physicians’ and nurses’ notes, imaging reports, medications and other input/output events related to the patient) and demographic data. From the available data, and for this particular problem, we are mainly interested on the physicians’ and nurses’ notes.

The MIMIC II text database contains a total of 156 million words with 3 or more characters, and 260180 distinct words. Of these 260180 distinct words, only 31527 (12%) appear in known word lists: 30828 appear on the SIL list of English known words (which contains 109582 distinct words) (SIL, 2014), and 429 appear on additional lists containing medical terms not common in English. The remaining 228923 words are simply unknown to dictionaries, and most are the result of typing or cultural errors!

As an example of the extent of such errors, here is a non-extensive list of the different misspelled variants of the word “abdomen” found in the MIMIC II database: abadomen, abdaomen, abndomen, badomen, abdaomen, abdeomen, abdcomen, abdemon, abdeom, abdoem, abdmoen, abdemon, abdiomen, abdman, abdmen, abdme, abddmen, abbomen, abdmn, abdme, abdmonen, abdonem, abdoben, abdodmen, abdoemen, abdoem, abdoem, abdomin. It should be noted that these errors are not isolated, e.g., the incorrect form “abdomin” appears 1968 times in the database.

3 RELATED WORK

A typographical error, colloquially known as “typo”, is a mistake made in the typing process. Most typographical errors consist of substitution, transposition, duplication or omission of a small number of characters. Damerau (1964) considered that a simple error consists in only one of these operations. Nowadays, many other types of errors can be found in text databases: Errors associated with smaller keyboards became relevant due to their effect in the increase of the number of word typos; Errors due to the widespread use of blogs, microblogs, instant messaging, etc.; Errors associated with real time voice transcriptions; Errors associated with poor Optical Character Recognition when digitizing manuscripts; etc. One must also mention linguistic errors, which are mostly due to lack of culture and or/education, and are usually the result of phonetic similarities.

As described previously, automatic word error correction is an expensive task when performed off line for which there is no current reliable automatic solution. The best performing methods are those that aim to find the word that is most probable to be the correct word in a given context. These methods are based in probability theory and what is the most likely word to follow a previous one (Jurafsky and Martin, 2009). However such methods demand many resources and are not the most adequate in texts containing many technical terms, many errors, and a very large vocabulary.

Independently from the degree of automatization, any word correction tool depends on a good similarity function to find the most likely correct word. Current research on string similarity offers a panoply of measures that can be used in this context, such as the ones based on edit distances (Levenshtein 1966) or on the length of the longest common subsequence of the strings. However, most of the existing measures have their own drawbacks. For instance, some do not take into consideration linguistically driven misspellings, others the phonetics of the string or the mistakes resulting from the input device. Moreover, the majority of the existing measures do not have a strong discriminative power, and, therefore, it is difficult to evaluate if the proposed suggestion is reasonable or not, which is a core issue in unsupervised spelling. Here we will use the Fuzzy Uke Word Similarity (FUWS) (Carvalho and Carola and Tomé 2006a; 2006b; Carvalho and Coheur 2013). This word similarity function combines the most interesting characteristics of the two main philosophies in word and string matching, and by integrating specific fuzzy based expert knowledge concerning typographical errors, can achieve a good discrimination.

4 TOWARDS AUTOMATIC WORD CORRECTION IN TEXTUAL BIG DATA

Given the above considerations concerning the extent of the word errors present in the MIMIC II text database, and the impact of such errors when considering any kind of text analysis, we propose a semi-automatic procedure to detect and correct typographical and other word errors in the MIMIC II text corpus, that improves and details the approach presented in (Carvalho and Curto, 2014), and can be extended to other Big Data Textual databases given the appropriate resources.

4.1 Resources – Known Words List (KWL)

The proposed procedure needs an extensive known-words list, and one or more technical words lists related with the subject of the textual database. We assume that, despite our best efforts, technical word lists might not be complete. The ordered set of all words contained in these lists, forms what we refer to as the "known-words list" (KWL).

It is also necessary to use a proper word similarity function, preferably fast (due to the size of the target databases) and that can achieve a good discrimination, i.e. has both a good precision and recall so that both false positives and false negatives are minimized.

4.2 Automatic Word Correction Steps

4.2.1 Corpus Word List (CWL)

The first step consists of creating a list containing all the different words present in the corpus, counting the frequency of each occurrence, and ordering the list. Words with less than 3 characters or more than 15 characters, and/or containing numerals are removed.

The removal of short words is due to the difficulty (or impossibility) of properly detecting and correcting such words. This is not an important issue since in NLP such words are often dismissed as they usually contain more noise than useful information.

Words with more than 15 characters are usually codes, concatenated words, sometimes chemical compounds, etc., and cannot or should not be corrected. Words containing numerals are removed since they usually consist of tokens that, as previously, should not (and cannot) be corrected.

4.2.2 High Threshold Filtering

In the second step we look very close matches between the CWL and the KWL in order to filter minor typos and to aggregate occurrences of very similar words. This is accomplished by selecting a very high threshold when testing for word similarity. When using the FUWS, the similarity threshold should be above 0.9 – note that in (Carvalho and Coheur, 2013), 0.67 is proposed as the standard similarity threshold. Here we want to guarantee that no false positives are generated, hence the much higher threshold.

It should be noted that this value will have more impact in errors occurring in longer words (more than 8 chars) than shorter ones.

After ordering the resulting list by word frequency we obtain what we refer to as filtered corpus word list (f-CWL). The f-CWL contains both known and unknown words and will be used as a corpus in the remaining procedure.

4.2.3 High Frequency (HFL) and Low Frequency Words List (LFL)

The next step consists in generating two different word lists based on the frequency of the words in the f-CWL: the High Frequency Word List (HFL), and the Low Frequency Word List (LFL).

The HFL will contain all the words in the f-CWL whose frequency is higher than a given High threshold *ht*. The HFL will be used as the known word list in the final word correction step.

The reasoning behind the creation of the HFL is based on the assumption that words that occur very frequently in the f-CWL should no longer contain errors (since common errors have been filtered in the previous step). As such, very frequent words in the f-CWL that are not present in the KWL should be considered new words instead of word errors, and should be used to correct errors that occur less frequently (this is consistent with the previous assumption that the available technical terms lists are probably not complete).

The LFL will consist of the unknown words of the f-CWL whose frequency is lower than a given Low threshold *lt*. It is important to note that words present in the KWL are removed from the LFL. The LFL will contain the words that should eventually be corrected in the final word correction step.

The LFL should consist mainly of: a) Very specific technical words; b) Unknown abbreviations; c) Unknown named entities (either individuals or organizations) and special non numerical codes; d) Words containing typing and/or other errors; e) Tokens formed by an undue lack of spacing separating proper words. Of these five different cases, only the word errors should be corrected. Note that errors occurring in some unknown named entities or in some abbreviations might also be corrected as long as the correct form is present in the HFL.

The definition of *ht* and *lt* values is obviously an important issue. Ideally they should be expressed as a percentage of the number of distinct words in the f-CWL, or of the joint word occurrence. Up to now the thresholds have been found empirically, but there are no indications if they can be generalized to other datasets.

4.2.4 LFL Correction

The final word correction step consists in attempting to correct the words in the LFL, while using the HFL as the known words list. Common sense would dictate using a normal word similarity threshold (in the case of the FUWS, the value would be 0.67). However, as it will be shown in the case study, best results were obtained using a 10% lower value.

4.3 Correction Results

After the application of the previous steps we obtain a word list that will be used to replace the appropriate occurrences in the original textual database. As discussed in 4.2.3, not all unknown words are expected to (or should) be replaced, only the ones resulting from typing or other word errors.

5 CASE STUDY AND RESULTS: MIMIC II DATA BASE WORD ERROR CORRECTION

In this section we apply the previously presented procedure to the MIMIC II textual database.

To build the KWL we used the SIL English word list (SIL, 2014) and three medical terms lists publicly available online (mtherald, 2014) (Heymans, 2014) (e-medtools, 2014).

As a word similarity function we used the above mentioned FUWS, since it is fast, combines the most interesting characteristics of the two main philosophies in word and string matching, and by integrating specific fuzzy based expert knowledge concerning typographical errors, can achieve the intended good discrimination.

The original database contains 1 095 127 distinct words. After executing Step 1, the resulting CWL contains 260180 distinct words (corresponding to a joint occurrence of 177 446 957 words).

For the High threshold filtering operation a FUWS empirical threshold of 0.935 was chosen after several tests. This operation affected 15032 distinct words, which ended up being combined as 7805 distinct words. The number of errors is estimated (by sampling the results) to be much lower than 1%. Note that this reduction in the number of distinct words is quite significant if one considers that, in English, the necessary vocabulary to properly understand Academic textbooks ranges from 5000 to 10000 words. So we managed to reduce a similar number of words by using the high threshold filtering operation. After this step, the f-CWL size is 252 953.

In order to create the HFL, the LFL, and to apply the LFL correction, it was necessary to define the High and Low thresholds, and also choose the FUWS threshold. Several empirical tests were performed in order to find appropriate values. Even if the numbers are not yet fully optimized, words occurring more than 1400 times in the f-CWL appear to be good correction candidates for words occurring less than 800 times when using a FUWS threshold=0.6.

Therefore we are currently using *lt*=800 and *ht*=1400. The HFL consists of 6153 distinct words (with a joint occurrence of 171642123 words), i.e., the HFL contains the only the top 2.43% most frequent words, and yet contains 96.73% of the total number of words. The LFL has 200137 distinct words (79%) corresponding to a total of only 1585067 words of the database (0.8%).

The correction of the LFL using the HFL and a FUWS threshold of 0.6, resulted in the correction of 88867 distinct words (out of the 200137), reducing them to 5920 distinct words. I.e., the automatic procedure found 88867 different words containing typing errors that corresponded to only 5920 words. Those 88867 words have a joint occurrence of 59% of the LFL.

As expected, the uncorrected words fall mainly into the categories indicated in section 4.2.3. However not all the proposed corrections are correct. An estimation based on sampling the 25% more frequent corrections indicates the number of false positives to be around 5%. Overall this results in a very low number of errors when compared to the number of different words and joint occurrences. In the end, only 0.02% of the MIMIC II words are estimated to be incorrectly replaced, and only 0.48% are left uncorrected. Most of the unknown words that were not proposed for correction correspond to cases that should not be replaced or are indeed very difficult to correct without additional lengthy preprocessing. Following are some examples of the observed errors using the format "Unknown Word → Proposed Correction (# occurrences) (FUWS sim value); comments":

- Abbreviations:
  stg → gtts (209) (0,7500); stg may mean "superior temporal gyrus"
  ptsd → ptbd (317) (0,7500); ptsd may mean "post-traumatic stress disorder"
- Prefix variation not present in the known words:
  untolerated → tolerated (1) (0,7955); incorrect use of the prefix un-, proper correction should be "not tolerated"
  noincreased → increased (2) (0,7500); correct substitution should be "not increased"
- Terms aliasing due to the lack of spacing between words:
  remainslow → remains (1) (0,6750); should be "remains low")
  withinthe → within (5) (0,6389); should be "within the"
  parentsup → parents (1) 0,75; should be "parent support"
- Correct word is not the most similar:
  Weerk → were (1) (0,8125); probably the correct substitution should be "week" but it only has a 0,7 similarity
- Other special cases:
  gmother → mother (14) (0,8214); should be "grandmother"
  anormal → normal (12) (0,8214); correct substitution should be abnormal

6 CONCLUSIONS AND FUTURE WORK

In this paper we propose and describe a novel semi-automatic procedure to detect and correct errors in unedited textual Big Data based on a fuzzy word similarity function. The procedure is being applied to the MIMIC II database with very encouraging results.

The largest obstacle to the automatization, generalization and application of the method to other databases consists in the parameterization. Up to now, the definition of the three thresholds used in the proposed procedure, *lt*, *ht* and FUWS threshold, has been made empirically. An automatic procedure would be achieved if the values found up to now can be directly applied to other databases. However, that is not necessarily a likely situation, and we will not have an answer until the procedure is tested in other textual Big Data. Another option towards a more automatized process consists in using the obtained values as a starting point to an optimization procedure. Since only three parameters (and most likely only two) are involved, evolutionary algorithms could certainly be used towards this goal.

ACKNOWLEDGEMENTS

This work was supported by national funds through FCT – Fundação para a Ciência e a Tecnologia, under project PTDC/EMS-SIS/3220/2012 and project PEst- OE/EEI/LA0021/2013.

REFERENCES

A. S. Fialho, F. Cismondi, S. M. Vieira, S. R. Reti, J. M. C. Sousa, and S. N. Finkelstein. 2012. Data mining using clinical physiology at discharge to predict icu readmissions, *Expert Systems with Applications*, vol. 39, no. 18, pp. 13 158–13 165, December 2012.

A. S. Fialho, U. Kaymak, F. Cismondi, S. M. Vieira, S. R. Reti, J. M. C. Sousa, and S. N. Finkelstein. 2013. "Predicting intensive care unit readmissions using probabilistic fuzzy systems," Proc. of FUZZ-IEEE 2013, Hyderabad, India.

Carvalho, J. P., Coheur, L., 2013. Introducing UWS – A Fuzzy Based Word Similarity Function with Good Discrimination Capability: Preliminary results, *Proc. of the FUZZ-IEEE 2013*, Hyderabad, India.

Carvalho, J. P., Curto, S., 2014. Fuzzy Preprocessing of Medical Text Annotations of Intensive Care Units Patients, *Proc. of the IEEE 2014 Conference on Norbert Wiener in the 21st Century*, Boston, USA.

Carvalho, J. P., Carola, M., Tome,J.A., 2006. Using rule-based fuzzy cognitive maps to model dynamic cell behavior in Voronoi based cellular automata, *Proc. of the 2006 IEEE International Conference on Fuzzy Systems*, pp. 1687-1694, Vancouver, Canada.

Carvalho, J. P., Carola, M., Tome, J. A., 2006. Forest Fire Modelling using Rule-Based Fuzzy Cognitive Maps and Voronoi Based Cellular Automata, *Proceedings of the 25th International Conference of the North American Fuzzy Information Processing Society, NAFIPS 2006,* Montreal, Canada.

Damerau, F. J. A Technique for Computer Detection and Correction of Spelling Errors. *Communications of the ACM*, Março 1964, pp. 171-176.

e-medtools 2014. http://e-medtools.com/openmedspel.html, last accessed May 2014.

F. Cismondi, A. L. Horn, A. S. Fialho, S. M. Vieira, S. R. Reti, J. M. Sousa, and S. Finkelstein. 2012. Fuzzy multi-criteria decision making to improve survival prediction of icu septic shock patients, *Expert Systems with Applications*, vol. 39, no. 16, pp. 12 332–12 339.

Heymans 2014. http://users.ugent.be/~rvdstich/eugloss/EN/lijst.html, last accessed May 2014.

Jurafsky, D., Martin, J., 2009, Speech and Language Processing - An introduction to natural language processing, computational linguistics, and speech recognition, 2nd Edition, Prentice-Hall.

Levenshtein, V. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10:707–710.

M. Saeed, C. Lieu, and R. Mark, "Mimic ii: A massive temporal icu database to support research in intelligence patient monitoring," *Computers in Cardiology*, vol. 29, pp. 641–644, 2002.

mtherald 2014. http://mtherald.com/free-medical-spell-checker-for-microsoft-word-custom-dictionary/, last accessed May 2014.

SIL, 2014. http://www01.sil.org/linguistics/wordlists/english/, last accessed on February 2014.