**COMP90094 Project 1 report: very tweetz. such non-canonickal. amayze**

**1 Introduction**

Lexical normalisation is the process of finding a canonical form of non-standard tokens. This process can take the form of approximate matching techniques used to predict standard forms based off a non-standard word.

Lexical normalisation in the context of social media data presents a unique challenge and one traditional natural language processing methods are not built for. Social media platforms present “language that is substantially different from any benchmark corpora” (Eisenstein, 2013). Shortenings, abbreviations and newly coined colloquialisms are some examples of difficult language to standardise.

This report will assess the performance of a number of approximate matching techniques on normalising a set of tokens from ‘tweets’ selected from the Twitter social media platform. These techniques compare the non-standard tokens to a provided dictionary of standard words and produce a list of candidate matches. These matches are then evaluated using a set of correctly normalised words.

**2 Methodology**

For all applied methods, the initial list of words was filtered for words that existed in a standard dictionary and the remaining, non-standard words were normalised.

**2.1 Global Edit Distance**

This method was employed to provide an alternative to the other techniques which employ some form of phonetic matching. The implementation of this method uses the Needleman-Wunsch Algorithm (Needleman and Wunsch, 1970). Two separate scoring parameters were used: Levenshtein Distance (Levenshtein, 1966) and a customized set. The customized set was derived to reward matches with +1, instead of 0 in the LD (i.e. 1,-1,-1,-1). This method is limited by the fact there is no ranking for when multiple possible candidates with the best edit distance are found.

**2.2 Soundex**

The Soundex algorithm, (Odell and Russell, 1918) generates four digit codes that correspond to a token’s phonetic profile. Comparing these codes allows us to perform lexical normalisation based off phonetics which in this context may be a powerful matching strategy; a common trait of language used in instant messaging or on social media is to use language informally, much the same as how it is spoken. This can mean using shortenings that bear a significant phonetic similarity to their standard form, but are different on a pure lexicology level. For example, “sux” being used to mean “sucks” will

For this report, the Soundex algorithm was applied to all dictionary entries and each non standard term was matched with the set of standard dictionary terms that have the same Soundex code. These candidates were then ranked using GED and the best candidate chosen.

**2.3 Editex**

The Editex algorithm, proposed by Zobel and Dart (Zobel and Dart, 1996), provides a variation to the standard edit distance measures, measuring the “phonetic distance” (Zobel and Dart, 1996) between two words. This method appears promising for similar reasons to the Soundex method.

**3 Evaluation criteria**

Relevance or “correctness” is defined in this experiment, as predicting an exact match to the canonical form found in the set of normalised words. To measure the effectiveness of the aforementioned methods, this report will include two evaluation metrics:

* Accuracy – the proportion of best matches for a set of queries that are deemed correct
* Recall – the proportion of correct matches that were found using a given method for a given query. In this case, as there is only one standard (“correct”) form for each query, the recall for each method applied to each query will be either 0 (did not find correct form) or 1 (did find correct form). For this experiment, Recall@10 will be used and the average recall over the query set will be taken. This metric is used to establish how accurate a method is at finding correct answer from a list of high ranking candidates.
* Time taken – run time for each method is also noted.

**4 Results**

The results listed in the table below show the accuracy (Acc.), recall@10 (Rec.) and run time (Time) in seconds for the three methods discussed. The methods were applied to the whole supplied tweet data set and compared with the normalized set. (NOTE: CS – Custom Set)

|  |  |  |  |
| --- | --- | --- | --- |
| **Method** | **Rec. (%)** | **Acc. (%)** | **Time (sec)** |
| GED - LD | 7.66 | 3.03 | 319.89 |
| GED - CS | 9.30 | 5.09 | 276.39 |
| Soundex | 5.76 | 3.44 | 15.69 |
| Editex | 6.78 | 2.62 | 1227.14 |

**5 Discussion**

GED has the highest measures of both recall@10 and accuracy on the data set. Editex performed surprisingly poorly, considering its focus on phonetic weighting which in this context is arguably seen as more appropriate. Soundex similarly performed well below the GED.

A possible theory to explain this outcome is the proportion of non-standard features in the tweets may favour the GED. For example, a term such as ‘dat’ meaning ‘that’ will not be assessed using the Soundex algorithm due to the different first letters; however, the GED between the two words will be relatively large (???), meaning it is more likely to be identified as a candidate.

Overall, these results are fairly poor. The task itself is difficult and the highly variable language of social media platforms makes usage of traditional matching techniques somewhat ineffective. For example, there are a whole range of non-standard features found in this context that would require alterations to the current techniques or a new matching technique altogether:

* Abbreviations (such as ‘ofc’, meaning ‘of course’) could not be resolved using any of the above methods. This is due to the standardised form containing multiple words and there being no lexical or phonetic resemblance to the correct form besides the starting letters.
* Different languages are used across social media (eg. Spanish, ‘gusta’, in the dataset) and standardising this would require translation.
* Proper nouns (such as names and places) are commonly used in social media. They are technically already normalized, however in the context of this experiment will not be resolved as they are not in the dictionary.
* Numbers will in some cases require normalisation (eg. ‘2’ to ‘to’), but in other cases should be left in the current form.
* Emojis (eg. ‘: d”) are commonly used across social media, but would need to be ignored in the normalisation process.

Taking these common features into consideration the scope of using approximate matching techniques is more obvious. Lexemes with standardised forms that match their usage (eg. Proper nouns, emojis, different languages and some use of numbers), but are not in the dictionary, would require a more elaborate approach in lexical normalisation. This may include an expansion of the dictionary to include references to emojis, common abbreviations, proper nouns and other hardcoded translations to resolve these cases. This is a somewhat dangerous path to go down as it involves extensive human contribution which goes against the original need for a computer algorithm for lexical normalisation.

The normalisation that is more achievable in this context are the lexemes that have undergone more slight changes (eg. Shortenings).

Considering the above discussion, a more accurate assessment was created that removes any predictions where the query is the same as its correct form (i.e. words that required no change, but were not in the dictionary):

|  |  |  |  |
| --- | --- | --- | --- |
| **Method** | **Rec. (%)** | **Acc. (%)** | **Time (sec)** |
| GED - LD | 39.84 | 15.78 | 35.21 |
| GED - CS | 48.40 | 26.47 | 29.86 |
| Soundex | 29.95 | 17.91 | 3.29 |
| Editex | 35.29 | 13.64 | 201.54 |

These results are for discussion purposes only as they use the correct set to filter the queries; a real method would require a more comprehensive dictionary that dealt with more of these issues (eg. Included more proper nouns and instant-messaging terms, such as emojis) or some other filtering method that looked for these words and left them unaltered.

With GED showing the greatest promise, the main drawback of no ranking for queries that produce multiple best candidates was addressed; using N-Gram distances to rank the candidates with the same GED to the query, the following results were produced:

|  |  |  |  |
| --- | --- | --- | --- |
| **Method** | **Rec. (%)** | **Acc. (%)** | **Time (sec)** |
| GED - LD | 39.84 | 18.45 | 35.21 |
| GED - CS | 48.40 | 26.47 | 64.855 |

The results using the refined method were similar showing a minor improvement in accuracy.

**6 Recommendations**

To further the above investigation, a few points need to be considered:

* Expanding the dictionary – inclusion of non-standard lexemes and a more complete, up-to-date index of internet lingo that have no standard form.
* Refinement of current methods – more exploration into common non-standard features in social media data and refinement of methods based off of this
* Pre-processing of dictionary – as the complexity of the matching method increases, pre-processing could be used to reduce time and computational cost.

**7 Conclusion**

The application of a number of standard approximate matching techniques in the context of social media data produced varied results. This report discussed the shortcomings of these techniques and briefly outlined how a more comprehensive method of matching should be approached.

**8 References**

Baldwin, T., de Marneffe, M. C., Han, B., Kim, Y.-B., Ritter, A., & Xu, W. (2015). Shared Tasks of the 2015 Workshop on Noisy User-generated Text: Twitter Lexical Normalization and Named Entity Recognition. *Proceedings of the ACL 2015 Workshop on Noisy User-generated Text*, 126-135.

Eisenstein, J. (2013). What to do about bad language on the internet. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 359-369.

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

Needleman, S., & Wunsch, C. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 443-453.

Russell, R., & Odell, M. (1918). *US Patent No. 1261167.*

Zobel, J., & Dart, P. (1996). Phonetic String Matching: Lessons from Information Retrieval. *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, 166-172.