COMP90049 Knowledge Technologies: Word Blending in Twitter

Anonymous

1 Introduction

This paper outlines the methods used to detect lexical blended words in the twitter data set [1]. The definition of Lexical blending is not concrete; however, it is defined as word blends created by two relatively novel words to convey combined meaning e.g. breakfast + lunch = brunch.

There are many challenges in detecting blended words since there is no formal formatting to how they are formed; e.g. Britain + exit = Brexit combined "Br" from Britain with all letters of "exit" however in the case of Brunch "Br" is used but only partial letters from "lunch" are used. Thereby the data from the twitter could be personalized by the user [2], leading to a different version of blended word carrying the same meaning.

This paper outlines the techniques used to detect lexical blends from candidates using string approximate matching techniques such as Jaro-Winkler Similarity, suffix and prefix similarity. Evaluating these techniques against the list of true lexical blends for accuracy, recall and precision, following with discussion of their respective performance and their suitability for this task.

2 Data Set

The data set used was a preprocessed twitter data and stored 16,684 tokens as potential blends denoted as candidates [1]. This data set contained lexical blend with addition to words with no meaning, spelling error.

Candidates words were checked against data set of dictionary words provided to find the lexical blends.

The algorithms were evaluated against a data set of true lexical blend manually extracted from the twitter data.

3 Models of Blends

3.1 Hypothesis

The lexical blends are formed using at least 2 letters from the prefix of 1st word and 2 from the suffix of the 2nd word.

3.2 Experiment

The way to test the hypothesis was to have one candidate word and having its prefix and suffix similarity checked with the dictionary data set. For a candidate to be a lexical blend, it must have prefix and suffix similarity with 2 different dictionary words (i.e. source words).

This experiment was carried out by two different approaches to detect lexical blending using approximate string match.

1. The first approach is to detect the lexical blends, using the approximate string-matching technique, of Jaro-Winkler best known for suffix and prefix similarity matching [3]. It was used for matching the source words.
2. The second method involved a combination of Jaro-Winkler with prefix and suffix length matching of the source words.

3.2.1 Assumptions for this data set

1. Candidate words are standalone lexical words and do not require contextual information to be identified as a blend.

2. All words from true blend data set are present within the candidate data set.

4 Methodology for Experiment

4.1 Filtering candidates

The candidate data contained words such as *'aaaaaaa'* and '*ahaha*' and spelling mistakes. Filtering these types of the candidate which do not contain any form of word blending were removed through an algorithm designed specifically for this. Filtering was done to eliminate the chances of non-lexical blend words being detected as blends and decrease computation time.

4.2 Method One

The Jaro-Winkler similarity is predominantly used for prefix and suffix similarity measure. A subset of the true blend data was taken to check the similarity of the candidate (blended word) with its source words; prefix source and suffix source.

As per the hypothesis, these lexical blends are a combination of a prefix and a suffix from two different source word. There should be a strong correlation between the candidate and its sources.

Statistical analysis was run to check the similarity value for the suffix and the prefix from the source word with the candidate, to gain better understanding composition of the lexical blend.

It was found from the subset of data the similarity for the prefix was between 0.75 – 0.85 and for suffix between 0.85 – 0.95. Suggesting that the suffix of a source word seems to be the dominating part, this observation is also agreeing with the study conducted on lexical blends [4].

These values were used to create thresholds for the algorithms detecting the lexical blends. The logic of the algorithm was if a candidate word has suffix similarity and prefix similarity within those bounds for 2 different source words from a dictionary, it is most likely to be a lexical blend.

4.3 Method Two

This method was developed build upon method one and used additional tools in further targeting lexical blends and to test the hypothesis.

With the addition of Jaro-Winkler, Suffix similarity and prefix similarity algorithms were used in detecting a lexical blend. The suffix and prefix similarity algorithms detect the length of the suffix and prefix matching characters. The typical constraints used to detect blends involved taking at least 2 characters from suffix and prefix of the source word [4].

This allows, in pinpointing the possible prefix and suffix source word. Jaro-Winkler is then applied in the same manner as described in method 1 using those thresholds.

Method 2 does not depend on floating values as method 1; it filters the words with at least 2 characters matching further applying Jaro-Winkler to determine the candidate's suitability of being a lexical blend.

5 Evaluation

Evaluation matric was developed to evaluate the suitability of the two methods. The used to evaluate the methods are Accuracy, Precision and Recall.

Accuracy: the fraction of correct detection of blends among a total number of candidates.
Precision: the fraction of correct detection of blends among attempted responses
Recall: the proportion of correct detection of blends among true items (blends)

Metric (%)	Method 1	Method 2
Accuracy	1.1784	0.14984
Precision	0.9963	1.8939
Recall	14.883	13.6612

6 Discussion

Given the goals of this paper is to detect the blends from the twitter data set. Method 2 performed better compared to method 1, this is determined by observing the value of Precision. Precision describes the algorithms output reliability; high precision refers high proportion

of predicted outputs to be correct; thereby method 2 has better detection capabilities.

Method 1 has a lower score due to` sole use of Jaro-Winkler. Despite it is best known for prefix/suffix matching; this technique alone did not produce a good result. As it focuses on the multiple variables such as the length of the string and cross-matching of letters and does not focus strictly on prefix/suffix components when compared with another term. For example, Californication = California + fornication; when Jaro-Winkler is utilized in this case, the similarity out for prefix source word is 0.93 and suffix is 0.77, only provides floating values but no absolute information regarding suffix and prefix match. Thereby Jaro-Winkler needs to be used with another measure to output correctly as done in method 2.

The method 2 return supports the hypothesis made, with regards to the prefix and suffix length matching leading to higher precision i.e. more reliable outputs.

6.1 Improvements

The algorithm certainly fails at many levels for correctly identifying a lexical blend. Observing the true blends list, it was noted, using Jaro-Winkler with many other techniques could potentially improve the precession.

6.1.1 Context information

Some lexical words such as Hijack = Highway + Jacker; given contextual information an algorithm could have been developed in detecting the lexical blends of such.

6.1.2 Personalized

Lexical blending is changing and is not uniform [5], there are words being blended together to form a new blend. Since there is no formal formatting for a blending word, it is susceptible to personalization by the user. Thereby, an AI generative learning [2] [5] could be used to train the model and apply reinforcement learning such that it models the lexical blends formatting outputs the correct blend.

6.1.3 Phonetics

Words such as amtrak = amphibious + tractor could be detected if, phonetics were considered. Developing an algorithm, by modelling the phonetics would increase the precession.

7 Conclusion

The hypothesis of the lexical blends consisting of least 2 letters from the suffix of source and 2 from the prefix is not always true, this constraint does not always work as phonetics contains was also needed to be taken into account [5]. Only using Jora-Winkler as a matching method will not always provide complete information to improve the detection of lexical blends.

An AI algorithm, learning in real time the development of the lexical blends and modeling will certainly increase the precision of the algorithm.

8 References

[1] B. O. N. A. S. E. P. X. Jacob Eisenstein, "A latent variable model forgeographic lexical variation.," 2010.

[2] K. a. G. S. Das, "A Neural Network Ensemble Model for Lexical Blends.," in *The 8th International Joint Conference*, 2017.

[3] M. H. a. F. C. Schadd, "A Generalization of the Winkler Extension and its Application for Ontology Mapping".

[4] P. a. S. S. Cook, "Automatically Identifying the Source Words of Lexical Blends in English.," *In Computational Linguistics.*

[5] A. a. K. K. Deri, "How to Make a Frenemy: Multitape FSTs for Portmanteau Generation," in *The 2015 Annual Conference of the North American Chapter of the ACL,*, 2015.