A Statistical and Machine Learning Investigation into the Effect of Demographic Context on Dialogue Phenomena

David Hopes and Julian Hough[1]
[1]Cognitive Science Group
School of Electronic Engineering and Computer Science
Queen Mary University of London
London, UK
d.hopes@se17.qmul.ac.uk, j.hough@qmul.ac.uk

Abstract

We describe a study into how disfluencies are measures of communication efficacy and can indicate a demographic identity between participants in dyadic dialogue. The relative distribution of disfluency types are calculated for ternary (FF, FM, MM) gender context. Tests for association show that certain disfluency types are dependent on the demographic context, classification models using these features score well, predicting the gender context from the provided disfluency rates. The model is compared to common text classifiers and performs with similar accuracy at a faster training speed and the disfluency features are added to these classifiers to improve performance.

1 Introduction

The literature shows there is potential value in looking into disfluencies as indicators of a difference in communication efficacy and hence demographic context. The analysis of disfluency and dialogue annotation distribution for participants of differing demographics has not yet taken place using the Switchboard data-set and doing so will provide insight into differing demographic context and its effect on communication.

2 Background & Related Work

2.1 Disfluency

Disfluency, an involuntary disruption in the flow of speech, can be one way that dialogue efficacy is affected. Suggestions were made that disfluency rates may be correlated with gender (Lickley, 1994) and further studies elaborated on this by showing that filled pause rates are indeed correlated by gender, specifically that filled pauses that are more prevalent in males than females (Shriberg, 1994). Disfluency rates were shown to be higher in task-oriented speech for males with 4.35±1.62 disfluencies per 100 words than females with 3.76±1.42 (Branigan et al., 1999). The Switchboard Dialogue Act Corpus (SwDA) (Jurafsky et al., 1997) is composed of transcripts, tagged with demographic metadata, and utterances, annotated with disfluency markers.

3 Approach

3.1 Disfluency & Gender Context

The relative frequencies of disfluency types per word for ternary gender categories are calculated and plotted as a heat-map in Figure 1 and appears to show a marked difference between the three groups.

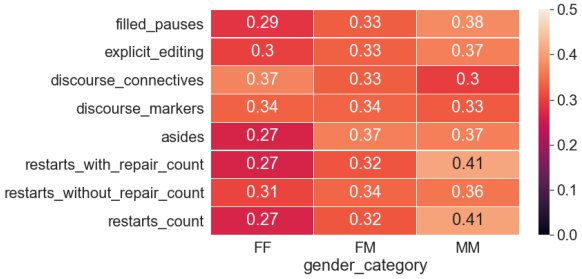

Figure 1: Heat-map showing relative frequency of disfluency per word rate for ternary gender context

4 Evaluation

4.1 Hypothesis Testing

Using chi-squared testing, the null hypothesis that there is no difference in the distribution of each disfluency type for female-only (FF), mixed (FM) or male-only (MM) gendered dialogue was rejected for all disfluency types, as shown in Figure 2.

4.2 Classification

A model was created to classify corpus utterances into the relevant labels. All classification was performed using 20-fold cross-validation on 80% of the corpus. The mean classification results for each

Proceedings of the 24th Workshop on the Semantics and Pragmatics of Dialogue, July 18-19, 2020, Online, hosted from Massachusetts, USA.

disfluency_types	chi_squared_score	p_value	null_rejected
restarts_with_repair_count	2191.327635	0.000000e+00	True
restarts_count	2190.436674	0.000000e+00	True
filled_pauses	1119.396647	8.435416e-244	True
explicit_editing	61.244858	5.021668e-14	True
discourse_connectives	59.624815	1.128849e-13	True
discourse_markers	57.108373	3.972595e-13	True
restarts_without_repair_count	37.288947	7.994810e-09	True
asides	13.653355	1.084455e-03	True

Figure 2: Table showing disfluency type and chi-squared testing result

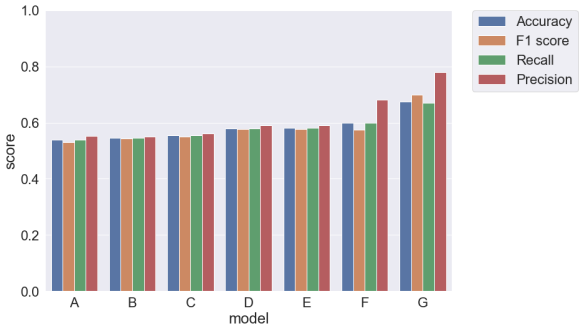

Figure 3: Chart showing classifier performance

fold were compared to standard text classifiers and the disfluency features were then combined to create more complex models.

4.2.1 Support Vector Classification

Using the disfluency types as features, a basic Support Vector Machine (SVM) classifier was trained resulting in an accuracy of 0.53±0.19 and F1 score of 0.524±0.18 (A). The parameters of the classifier were tuned with an L1 regularisation parameter with value of 1 resulting in an optimal accuracy of 0.538±0.18 and F1 score of 0.531±0.18 (B). The model training time was 384ms.

4.2.2 Comparison to Text Classifiers

To compare the performance of the disfluency classifier, standard text classification models were run, a bag-of-words classifier produced an accuracy of 0.55±0.12 and F1 score of 0.543±0.12 and took 7.03s (C). The TF-IDF classifier took 9.71s and produced a mean accuracy of 0.578±0.13 and F1 score of 0.576±0.13 (D).

4.2.3 Disfluency & Text Classification

The disfluency features were combined with the TF-IDF vector. The resulting classification took 10.5 seconds and produced an accuracy of 0.581±0.12 and F1 score of 0.576±0.12. This shows the accuracy of a standard TF-IDF classifier is improved upon by the inclusion of disfluency features (E).

An Adaboost classifier was implemented and the training time far exceeded the previous models at 8m16s, however the performance of the classifier increased with an accuracy of 0.600±0.11 and F1 score of 0.575±0.13 (F). The classifier was then trained on 80% of the corpus and tested on the remaining 20% providing an accuracy of 0.674 and F1 score of 0.700 (G).

4.3 Model Findings

The disfluency classifier worked well, achieving a mean accuracy of 53.8% on its own. Figure 3, shows the performance of the classifiers.

5 Conclusion

Studies of disfluency rates and context have been completed previously (Shriberg, 1996). However, they have only been analysed in the context of a single participant and their utterances, and not grouped participants and their utterances.

With the broader aim of investigating the relationship of how characteristics of dyads affect dialogue fluency and communicative efficacy, we found a dependency between participants' gender and disfluency rates. Previous research has been done into how to detect disfluency and similar phenomena, and models created for disfluency detection, but not the other way around, using disfluency to improve a language model. A number of machine learning models were successfully created as proof of concept that disfluency measures can be used to classify and improve classification of text. The model that used only disfluency features was much faster to train which could result in uses where training speed is valuable.

6 Limitations

Sparsity was a big limitation for the remaining forms of demographic context (age, dialect area and education) and getting reliable counts from a few transcripts that may be heavily influenced by a single caller reducing the ability to explore other types of context. The conclusions and test results from this analysis could differ when transferred to another data-set, therefore findings from this study could be SwDA specific.

References

Holly P Branigan, Robin J Lickley, and David McKelvie. 1999. Non-linguistic influences on rates of disfluency in spontaneous speech. *Proceedings of the 14th International Conference of Phonetic Sciences*, pages 387–390.

Dan Jurafsky, E Shriberg, and Debra Biasca. 1997. Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual.

Robin J Lickley. 1994. *Detecting Disfluency Spontaneous Speech*. Ph.D. thesis.

Elizabeth Shriberg. 1994. *Preliminaries to a theory of speech disfluencies*. Ph.D. thesis.

Elizabeth Shriberg. 1996. Disfluencies in Switchboard. *Proceedings of International Conference on Spoken Language Processing*.