Are you happy, contented or jubilant? Machine learning for sentiment analysis of children's diaries with highly infrequent words.

Marjolein de Vries (Leiden University), Maya Sappelli (TNO), Ad Feelders (Utrecht University)

This study is executed within the The Personal Assistant for a healthy Lifestyle (PAL[1]) project, which develops a social robot for children with diabetes. Type 1 diabetes mellitus (T1DM) is one of the most common diseases among children and youngsters in the United States and Europe (Freeborn et al., 2013). Within the PAL project, it is studied whether a robotic companion could help children with the self-management of all the daily tasks, and whether it could increase children's knowledge on diabetes using ontologies (Neerincx et al., 2016).

One task of the robot is to automatically detect sentiments expressed by the children in the diary entries in order to monitor their emotional development and to improve the dialogue between robot and child. In this study, automatically detecting sentiment in Dutch children's diaries by means of machine learning is researched. The goal is to classify diary text entries by children as either positive, neutral or negative. This can help the avatar choose an appropriate response - for example 'I am sorry you are feeling bad', when a child expressed a negative sentiment.

The current dataset for training consists of 395 text entries. The dataset used in this study is sparse and contains highly infrequent words. Thus, when using machine learning techniques on a training and a test set, infrequent words that only occur in the test set do not have a weight learned for it, making it more difficult to predict the target value. Hence, when training a model on the whole dataset and using it on new input, this new input can also contain words that are highly infrequent and are not in the dataset.

Therefore, a new algorithm for semantic normalization on top of standard morphological normalization; stopword removal and stemming (Hollink et al., 2004) is introduced which maps highly infrequent words to more frequent words based on a combination of synonyms and either Part-Of-Speech-tags (POS-tags) or Word2Vec similarity scores. For every of the 692 words in our dataset after stopword removal and stemming, synonyms were extracted from synoniemen.net. Then, for every infrequent word in the test set, a more frequent synonym is chosen based on one of two methods: 1) the first synonym is chosen which has the same POS-tag as the infrequent original when placed in the sentence, 2) the synonym with the highest similarity score is chosen based on a Word2Vec model trained on a large corpus of child-directed text (Tellings, 2014).

[1] http://www.pal4u.eu

Results show that this newly created step consistently improves the performance in the current study. Semantic normalization with the Word2Vec approach performs consistently better than using only morphological normalization ($p = .031$), and also consistently better than using the POS-tag approach ($p = .036$). Moreover, results show that the Word2Vec approach is more selective than the POS-tag approach, as it replaces less entries and less words with a synonym. We also see that, when using the new semantic normalization algorithm on top of morphological normalization, the accuracy for neutral entries slightly decreases, but the accuracy for negative and positive entries increases. This effect is useful for the current application, as the task of the robot is to detect sentiments expressed by children, and a higher classification rate for negative and positive entries contributes more to this goal than a higher classification rate for neutral entries.

Altogether, the semantic normalization algorithm is an addition to the already existing set of morphological normalization techniques such as stopword removal and stemming. Semantic normalization can be a viable addition to an automatic text analysis pipeline in other research as well, especially for small and sparse datasets.

References

Freeborn, D., Dyches, T., Roper, S. O., & Mandleco, B. (2013). *Identifying challenges of living with type 1 diabetes: child and youth perspectives*. Journal of clinical nursing, 22, 1890–1898.

Hollink, V., Kamps, J., Monz, C., & De Rijke, M. (2004). *Monolingual document retrieval for european languages*. Information retrieval, 7, 33–52.

Neerincx, M., Kaptein, F. , Van Bekkum, M., Krieger, H., Kiefer, B., Peters, R., Broekens, J. Demiris, Y. & Sappelli, M. (2016). *Ontologies for social, cognitive and affective agent-based support of child's diabetes self-management*. In *Proc. ECAI* (Vol. 16, pp. 35-38)

Tellings, A., Hulsbosch, M., Vermeer, A., & van den Bosch, A. (2014). *Basilex: an 11.5 million words corpus of dutch texts written for children*.