Idiom Token Classification with Distributed Semantics

Giancarlo D. Salton, Robert J. Ross, and John D. Kelleher

Applied Intelligence Research Centre
School of Computing
Dublin Institute of Technology
Ireland

Summary

Idiom token classification is the task of deciding for a set of potentially idiomatic phrases whether each occurrence of a phrase is a literal or idiomatic usage of the phrase. Identifying such usages of phrases is important for Natural language Processing (NLP) systems. For example, in Statistical Machine Translation (SMT) it has been shown that translations of sentences containing idioms receive lower scores than translations of sentences that do not contain idioms in SMT [2].

In this paper [3], we present an approach to idiom token classification based on features of distributed representations. We explore the capability of Skip-Thought Vectors [1] to encode features regarding the meaning of a sentence and show they are predictive with respect to idiom token classification.

We demonstrate that classifiers using these representations have competitive performance compared with the state of the art in idiom token classification. The current state of the art system uses specific features for each idiomatic phrase extracted from long discourse contexts. Importantly, however, our models use only the sentence containing the target phrase as input and are thus less dependent on a potentially inaccurate or incomplete model of discourse context.

We further demonstrate the feasibility of using these representations to train a competitive general idiom token classifier, i.e., a classifier that can take any idiomatic phrase as input. The ability to create a general idiom token classifier using these representations contrasts with previous work which focused on creating a separate models for each potentially idiomatic phrase.

References

1. Kiros, R., Zhu, Y., Salakhutdinov, R.R., Zemel, R., Urtasun, R., Torralba, A., Fidler, S.: Skip-thought vectors. In: Advances in Neural Information Processing Systems 28. pp. 3276–3284 (2015)
2. Salton, G.D., Ross, R.J., Kelleher, J.D.: An Empirical Study of the Impact of Idioms on Phrase Based Statistical Machine Translation of English to Brazilian-Portuguese. In: Third Workshop on Hybrid Approaches to Translation (HyTra). pp. 36–41 (2014)
3. Salton, G.D., Ross, R.J., Kelleher, J.D.: Idiom token classification using sentential distributed semantics. Proceedings of the 54th Annual Meeting on Association for Computational Linguistics pp. 194–204 (2016)