Proceedings of the Workshop on Hybrid Intelligence for Natural Language Processing Tasks HI4NLP (co-located at ECAI-2020)
Santiago de Compostela, August 29, 2020, published at http://ceur-ws.org

Syntactically Enriched Multilingual Sentiment Analysis

Carlos Gómez-Rodríguez
Universidade da Coruña, CITIC
Elviña, 15701 A Coruña, Spain
carlos.gomez@udc.es

Abstract

Sentiment analysis of natural language texts needs to deal with linguistic phenomena like negation, intensification or adversative clauses. In this talk, I present an approach to tackle such phenomena by means of syntactic information. Our approach combines machine learning and symbolic processing: the former is used to obtain dependency trees for input sentences, and the latter to obtain the sentiment polarity for each sentence using handwritten rules that traverse the tree. Thanks to universal guidelines for syntactic annotation, our approach is applicable to multiple languages without rewriting the rules. Additionally, very accurate parsing is not needed for our approach to be helpful: fast and simple parsers will do, even if they lag behind state-of-the-art accuracy.

The contributions presented in this talk are joint work with David Vilares, Miguel A. Alonso and Iago Alonso-Alonso.

1 Background

Polarity classification is a basic task in sentiment analysis of natural language texts, consisting on determining whether the expressed opinion in a piece of text is positive, negative or neutral. Since the seminal work by Pang et al. [3], many approaches have addressed the task by using machine learning models on features extracted from individual words or n-grams. However, a sentence is more than a set of words. For example, the following sentences contain the same words, but have opposite polarities:

> This phone is expensive, and not really very good. / This phone is good, and not really very expensive.

What sets them apart is word order, which in turn determines their syntactic structure, i.e., the way in which words interact with each other to form meaningful sentences. Namely, their polarities differ mainly due to the different roles played by the negation "not". In the first example, it modifies the positive word "good", making the phone "not good" and thus expressing a negative view, while in the second example, it modifies the negative word "expensive", making it "not expensive" which is positive. Since syntactic relations like these can happen at arbitrarily long distances, n-grams cannot always capture them – but a syntactic dependency tree can.

Syntax can be incorporated into a sentiment analysis model in various ways: for example, by decomposing trees into features for a supervised classifier [2] or by training a neural network on a syntactic treebank augmented with sentiment information [4]. Here, I present a different approach, where a dependency parse tree is used in conjunction with sentiment lexicons to propagate polarity according to a set of handwritten rules.

2 Sentiment analysis with syntactic rules

Our approach is described in detail in [6] (for Spanish) and [7] (for multiple languages). In brief, given a sentence, our sentiment analysis process has the following steps:

- The sentence is tokenized, part-of-speech tagged and parsed to a syntactic dependency tree. Any off-the-shelf dependency parser can be used for this purpose.
- Standard third-party sentiment lexicons are used to assign individual polarities to subjective words (for example, in a scale of -5 to +5, the word "excellent" could be assigned a polarity of +5, and "good" could be assigned +3).
- Handwritten rules are used to propagate the polarity in a top-down fashion, from individual words to the root of the tree (which yields the global polarity of the sentence). Rules are written to deal with syntactic phenomena that influence polarity. For example, given a negation, we use rules on the dependency tree to identify the scope of negation, and we modify the polarity of the negated element by subtracting -4 if it is positive, or +4 if it is negative (e.g. "not good" will be assigned $3 - 4 = -1$). Rules are also written to deal with adversative clauses, intensifiers and conditional statements.

Two conditioning factors need to be taken into account when writing the rules to ensure generalizability of the system across languages: first, rules are dependent on syntactic annotation criteria (but the appearance of universal annotation criteria that can be applied cross-linguistically, like Universal Dependencies, means that this is not an obstacle for multilinguality). Second, lists of negation words, intensifiers, adversative conjunctions and words introducing conditionals are needed to make the rules generalizable across languages, but these are available with standard sentiment lexicons.

Our experiments show that our approach outperforms existing unsupervised alternatives, as well as supervised models when evaluated outside of their corpus of origin. Using a high-accuracy parser is not vital [1], as we obtain good results even with modest parsing accuracy. Thus, very fast parsers, like those based on sequence labeling [5], can be used to obtain an overall efficient and green sentiment analysis system.

Acknowledgements

This work has received funding from the European Research Council (ERC), under the European Union's Horizon 2020 research and innovation programme (FASTPARSE, grant agreement No 714150), from the ANSWER-ASAP project (TIN2017-85160-C2-1-R) from MINECO, and from Xunta de Galicia and ERDF (ED431B 2017/01, ED431G 2019/01).

References

[1] Carlos Gómez-Rodríguez, Iago Alonso-Alonso, and David Vilares, 'How important is syntactic parsing accuracy? An empirical evaluation on rule-based sentiment analysis', *Artificial Intelligence Review*, **52**(3), 2081–2097, (Oct 2019).

[2] Mahesh Joshi and Carolyn Penstein-Rosé, 'Generalizing dependency features for opinion mining', in *Proc. of ACL-IJCNLPShort Papers*, pp. 313–316, Suntec, Singapore, (August 2009). ACL.

[3] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan, 'Thumbs up? sentiment classification using machine learning techniques', in *Proc. of EMNLP*, pp. 79–86. ACL, (July 2002).

[4] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts, 'Recursive deep models for semantic compositionality over a sentiment treebank', in *Proc. of EMNLP*, pp. 1631–1642, Seattle, Washington, USA, (October 2013). ACL.

[5] Michalina Strzyz, David Vilares, and Carlos Gómez-Rodríguez, 'Viable dependency parsing as sequence labeling', in *Proc. of NAACL*, pp. 717–723, Minneapolis, Minnesota, (June 2019). ACL.

[6] David Vilares, Miguel A. Alonso, and Carlos Gómez-Rodríguez, 'A syntactic approach for opinion mining on Spanish reviews', *Natural Language Engineering*, **21**(01), 139–163, (2015).

[7] David Vilares, Carlos Gómez-Rodríguez, and Miguel A. Alonso, 'Universal, unsupervised (rule-based), uncovered sentiment analysis', *Knowledge-Based Systems*, **118**, 45 – 55, (2017).