Shuhui Zhu
sz649

Critique 4: "Delete, Retrieve, Generate:
A Simple Approach to Sentiment and Style Transfer"

Juncen Li, Robin Jia, He He and Percy Liang

While the paper's proposed approach to detect attribute markers in reminiscent to Naïve Bayes with incorporation of neural generative models outperforms methods based on adversarial training, limitation lies in the process of separating attribute and content.

The current separation method calculates a relative frequency based on n-gram word count and sets a threshold to separate attribute and content. This approach poses a naïve assumption that words occur more often in positive (negative) corpus and less often in negative (positive) corpus are attitude words. Words with similar frequency in each corpus are content words. This assumption is flawed. First, the attitude of neutrality is not considered, which would share a similar frequency occurrence in both corpora. Second there are words appearing more often in positive (negative) corpus, but not attitude words. Depending on the sample, if one product or one restaurant, say McDonalds has more negative reviews rather than positive reviews, this product or restaurant would be considered as attitude word in the current separation strategy. However, in this case McDonalds is still content word. This imprecise assumption has led to a high error rate in mistagging, 16% in this paper's analysis. In addition, this high error rate is due to a disregard of the sentence structure.

Adopting sequence tagging task like fine-grained opinion extraction may reduce the mistagging rate. To obtain a training corpus, we could either search for existing corpus or utilize human resources. Given the current setup of the study, we could add one step in human reference, before asking Amazon Mechanical Turks to write gold outputs, tag the original text's attribute and content or more detailed five components of in opinion extraction, opinion trigger, polarity, strength/intensity, source, and target. They could choose to do a small amount as training data. Then HMM model could generate tags for the rest of sentences. First calculate the lexical generation probability, and the transition probability. Then use Viterbi algorithm to compute most likely tag given word. After obtaining tags, with the assumption that attribute corresponds to polarity and strength/intensity, content corresponds to opinion trigger, source, and target. We could delete attribute and proceeds to the next step proposed by the paper.

Another limit lies in the set-up of human evaluation. As pointed out, human evaluators' judgment of a sentence is largely relative to other sentences. The current setup is to have evaluators to finish evaluation in one batch. Since different evaluators may have different standards, it would yield a more consistent result across evaluators if we could provide a set of concrete examples corresponding to each scale under the three different criteria to benchmark against.

Despite of the limits, the paper's approach to disentangle attitudes and contents is relatively easy to implement. Meanwhile, the evaluation is comprehensive in conducting both human and automatic evaluation.