Stellingen

behorende bij het proefschrift

Normalization and Parsing Algorithms for Uncertain Input

van

Rob van der Goot

---

1. The hardest part of the normalization problem, is knowing when to normalize.
2. When applying a POS tagger on social media data, normalizing the input before tagging it is beneficial. If the tagger is also trained on social media data, normalizing the training data leads to further improvements.
3. Current state-of-the-art syntactic parsers perform well on news texts ($> 90\%$ accuracy), but experience a huge performance drop when applied to social media texts ($\approx 65\%$ accuracy).
4. Using normalization as a pre-processing step is effective for constituency parsing and dependency parsing of tweets.
5. For a constituency parser, integrating the normalization leads to an even better performance compared to the direct use of normalization. This can be done by representing the top-n normalization candidates as a word graph, and then using this word graph as input to the parser.
6. When integrating normalization, paraphrasing certain words with incorrect normalizations leads to higher parser performance.
7. For a neural network parser, integration of normalization can be done by merging the vectors of the top-n normalization candidates, weighted by the probability from the normalization model.
8. Even when using gold normalization, parser performance on tweets is still far from what is achieved on news texts. Complementary methods are necessary.
9. lmao kause I kan 🤷‍♀️ it ain't English klass, its twittr 💀 — lia, 2018
10. The ability to speak does not make you intelligent. — Qui Gon Jin, 32 BBY