

University of Groningen

Normalization and parsing algorithms for uncertain input

van der Goot, Rob Matthijs

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version

Publisher's PDF, also known as Version of record

Publication date:

2019

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):

van der Goot, R. M. (2019). *Normalization and parsing algorithms for uncertain input*. [Thesis fully internal (DIV), University of Groningen]. University of Groningen.

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

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