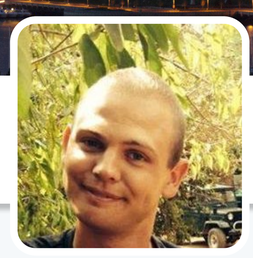
Parser Adaptation for Social Media by Integrating Normalization



TWEETS 513 | FOLLOWING 673 | FOLLOWERS 14,344

Follow

Rob van der Goot

@robvanderg

Abstract

This work explores normalization for parser adaptation. Traditionally, normalization is used as separate preprocessing step. We show that integrating the normalization model into the parsing algorithm is beneficial. To this end, we use a normalization model combined with the parsing as intersection algorithm. This way, multiple normalization candidates can be leveraged, which improves parsing performance on social media. We test this hypothesis by modifying the Berkeley parser; out-of-the-box it reaches an F1 score of 66.52. Our integrated approach performs significantly better, with an F1 score of 67.36, while using the best normalization sequence results in an F1 score of only 66.94.

Groningen

July 2017

@ r.van.der.goot@rug.nl

www.bitbucket.org/robvanderg/berkeleygraph

www.bitbucket.org/robvanderg/monoise

Previous photos and videos

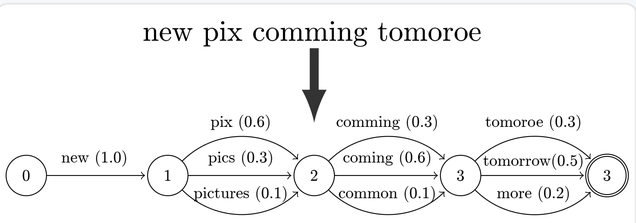


Figure 1: The output of the normalization model for the sentence 'new pix comming tomoroe'.

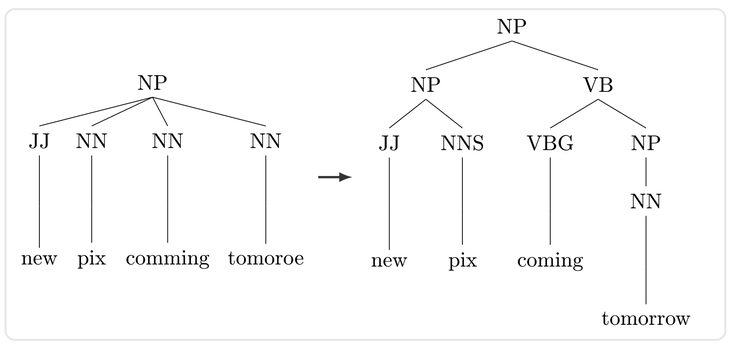
Corpus	Sents	Words/ sent	Unk%
WSJ (2-21)	39,832	23.9	4.4
EWT	16,520	15.3	3.7
Foster et al. (2011)	269	11.1	9.3
Li and Liu (2014)	2,577	15.7	14.1

Table 1: Some basic statistics for our training and development corpora. % of unknown words (Unk) calculated against the Aspell dictionary ignoring capitalization.

Tweets | Tweets & replies | Media

Rob van der Goot @robvanderg · Jan 10

The output of the Berkeley parser on a noisy sentence and its automatically normalized counterpart. #Interesting



45 14 43

Rob van der Goot Retweeted

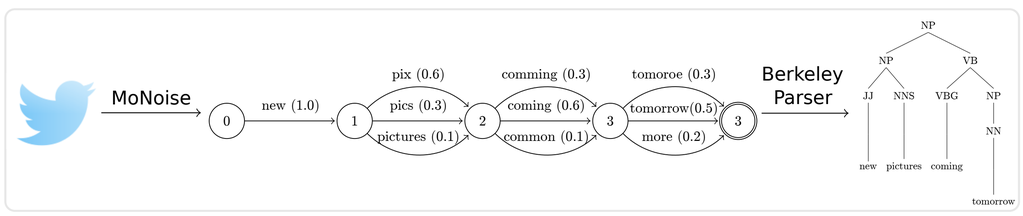
Gertjan van Noord @GJ · Jan 15

That is interesting!, maybe we can use the parsing as intersection algorithm to improve even further? 🙌🙌🙌

34 56 132

Rob van der Goot @robvanderg · Jan 20

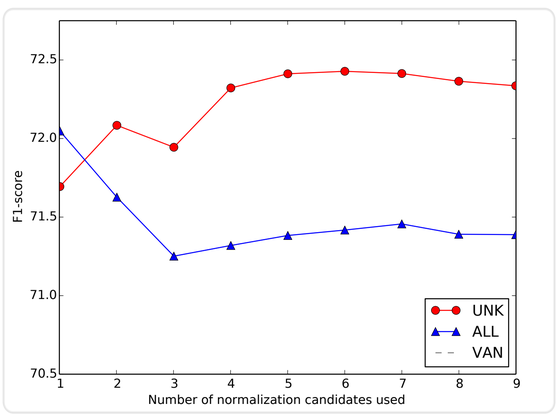
Overview of the model:



27 74 141

Rob van der Goot @robvanderg · Jan 22

@GJ F1 scores on the development data when integrating multiple candidates while normalizing ALL words or only the UNKnown words:



45 97 161

Rob van der Goot Retweeted

Gertjan van Noord @GJ · Jan 15

But Rob, is this #Significant?

34 56 132

Rob van der Goot @robvanderg · 3h

@GJ, it is! These are the F1 scores of our proposed models and previous work on the test set, trained on the EWT and WSJ, tested on a small Twitter treebank:

Parser	Dev	Test
Stanford parser	66.05	61.95
Berkeley parser	70.85	66.52
Best norm. seq.	72.04	66.94
Integrated norm.	**72.77**	**67.36***
Gold POS tags	74.98	71.80

*#StatisticalSignificant against Berkeley parser at P<0.01 and at P<0.05 against the best normalization sequence using a paired t-test.

1.1k 3.4k 7.5k

Back to top ↑

You may also like · Refresh

Yehoshua Bar-Hillel, Micha Perles...
On formal properties of simple phra...

Jennifer Foster, Ozlem C, etinoglu...
#hardtoparse: POS Tagging and pa...

Chen Li and Yang Liu
Joint POS tagging and text nomaliz...

Slav Petrov and Dan Klein
Improved inference for unlexicalized...

Worldwide Trends

#ParsingAsIntersection
33.9K Tweets

#ACL2017
152K Tweets

#normalization
35.1K Tweets

#NeuralNetworks
74.1K Tweets

#ConstituencyParsing
24.7K Tweets

#WordEmbeddings
57.3K Tweets

university of groningen

NUANCE FOUNDATION