Rob van der Goot
(Natural Language Processing)

Bla bla bla ...

What do I work on?

- ~~Normalization and syntactic parsing of social media texts~~
- ~~Multi-task learning in NLP~~
- What are open challenges in NLP?

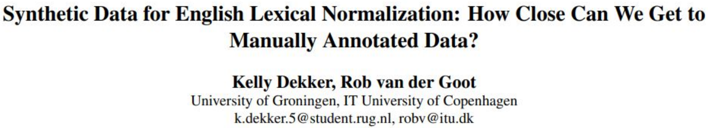
What do I work on?

- ~~Normalization and syntactic parsing of social media texts~~
- ~~Multi-task learning in NLP~~
- What are open challenges in NLP?
- Questioning assumptions

Projects I supervised

The following will be 1-2 minutes overviews of student projects, supervised by me.

🍒 🍒 🍒 Feel free to ask questions at any time!



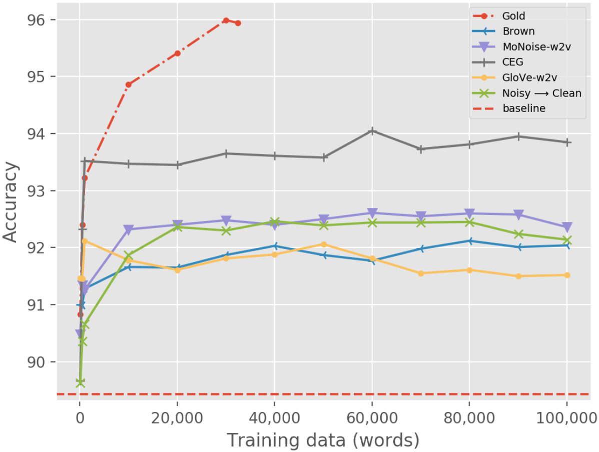
Synthetic Data for English Lexical Normalization: How Close Can We Get to Manually Annotated Data?

Kelly Dekker, Rob van der Goot
University of Groningen, IT University of Copenhagen
k.dekker.5@student.rug.nl, robv@itu.dk

wanting	to	make	a	yt	vid
wanting	to	make	a	Youtube	video

- Previous work: train-dev-test data
- Kelly: Only dev-test

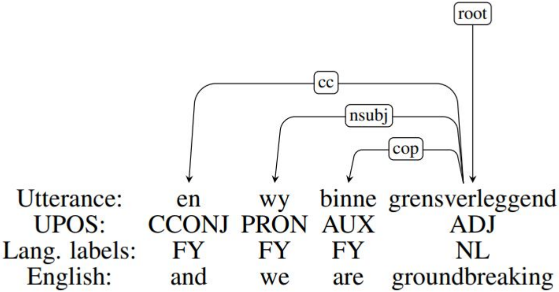
- Previous work: train-dev-test data
- Kelly: Only dev-test



Challenges in Annotating and Parsing Spoken, Code-switched, Frisian-Dutch Data

Anouck Braggaar
University of Groningen
a.r.y.braggaar@student.rug.nl

Rob van der Goot
IT University of Copenhagen
robv@itu.dk



Utterance:	en	wy	binne	grensverleggend
UPOS:	CCONJ	PRON	AUX	ADJ
Lang. labels:	FY	FY	FY	NL
English:	and	we	are	groundbreaking

- Annotated small dataset for evaluation, transfer from Dutch, German, or English
- Propose to select training data on the sentence level, instead of dataset level
- Also quite some negative results

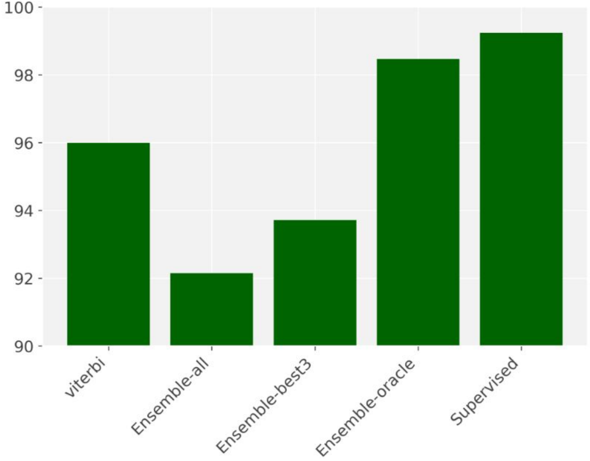
Much Gracias: Semi-supervised Code-switch Detection for Spanish-English: How far can we get?

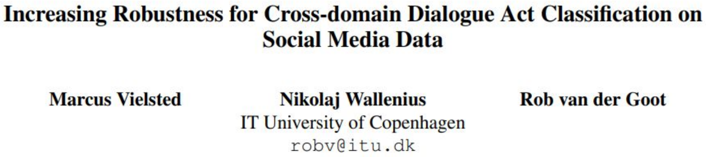
Dana-Maria Iliescu*, Rasmus Grand*, Sara Qirko*, Rob van der Goot
IT-University of Copenhagen
dail@itu.dk, gran@itu.dk, saqi@itu.dk, robv@itu.dk

El	online	exercise	de	hoy	:
ES	EN	EN	ES	ES	Other

- Current models: assume couple thousand sentences of training data, and massive training compute
- Dana et al: assume mono-lingual data, and a laptop for 10 minutes

- Current models: assume couple thousand sentences of training data, and massive training compute
- Dana et al: assume mono-lingual data, and a laptop for 10 minutes





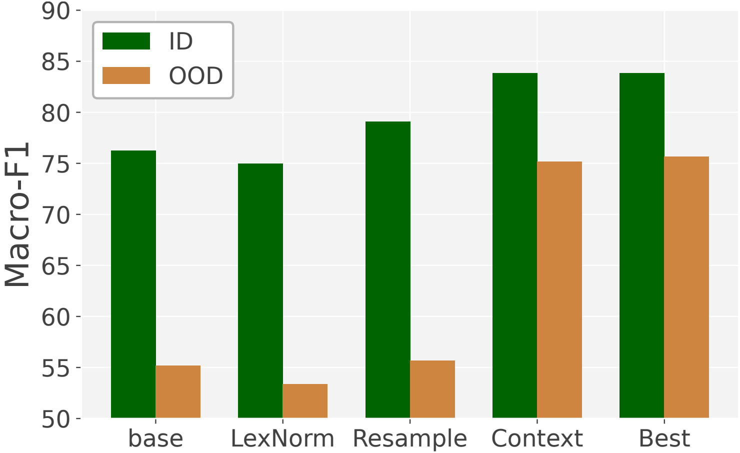
Increasing Robustness for Cross-domain Dialogue Act Classification on Social Media Data

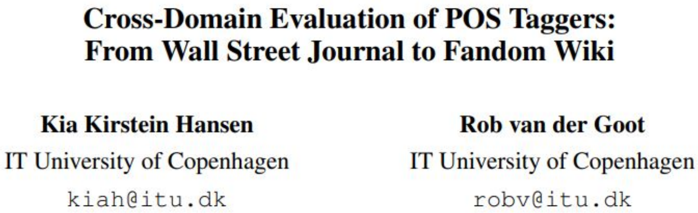
Marcus Vielsted **Nikolaj Wallenius** **Rob van der Goot**
IT University of Copenhagen
robv@itu.dk

Utterance	Label
"We are free tomorrow night, right?"	*propositional question*
"No, the final Grand Prix is on!"	*disagreement*

- Before: DAC only on online data from 2006 (train+test), or on phone conversations
- Marcus and Nikolaj: 2006 chat $\mapsto$ reddit
- Evaluate a variety of strategies to improve transfer with mixed results

- Before: DAC only on online data from 2006 (train+test), or on phone conversations
- Marcus and Nikolaj: 2006 chat ↦ reddit
- Evaluate a variety of strategies to improve transfer with mixed results

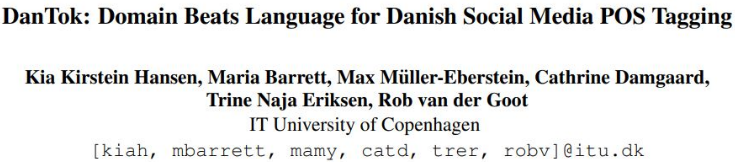




Cross-Domain Evaluation of POS Taggers: From Wall Street Journal to Fandom Wiki

Kia Kirstein Hansen
IT University of Copenhagen
kiah@itu.dk

Rob van der Goot
IT University of Copenhagen
robv@itu.dk



DanTok: Domain Beats Language for Danish Social Media POS Tagging

Kia Kirstein Hansen, Maria Barrett, Max Müller-Eberstein, Cathrine Damgaard, Trine Naja Eriksen, Rob van der Goot
IT University of Copenhagen
[kiah, mbarrett, mamy, catd, trer, robv]@itu.dk

Gentle$_{\texttt{NNP}}$ Star$_{\text{NNP}}$ of$_{\texttt{IN}}$ Stendarr$_{\texttt{NNP}}$ is$_{\texttt{VBZ}}$ a$_{\texttt{DT}}$ temple$_{\texttt{NN}}$ within$_{\texttt{IN}}$ the$_{\texttt{DT}}$ region$_{\texttt{NN}}$ of$_{\texttt{IN}}$ Alik'r$_{\texttt{NNP}}$ in$_{\texttt{IN}}$ Hammerfell$_{\texttt{NNP}}$.$_{\texttt{.}}$

- Transfer to relatively clean data with another topic is non-trivial: 97% $\mapsto$ 95%
- Better transfer from English social media compared to Danish news data!
- Lot of error analysis, many insights in remaining challenges