

# To Normalize or Not to Normalize: The Impact of Normalization on Part-Of-Speech Tagging

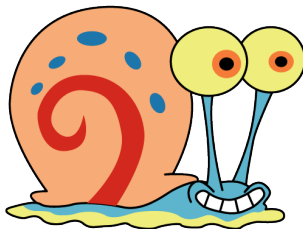
Rob van der Goot, Barbara Plank & Malvina Nissim  
University of Groningen

07-09-2017

# Problems

Gary did gd protectin SpongeBob house !

NNP VBD NN NN NN NN .



# Problems

Gary	did	gd	protectin	SpongeBob	house	!
NNP	VBD	NN	NN	NN	NN	.

Gary	did	good	protecting	SpongeBob's	house	!
NNP	VBD	JJ	VBG	POS	NN	.

# Problems

## Experiments

- Normalization for POS tagging
- Semi-supervised adaptation of a POS tagger
- Complementary

# Experimental setup

## Train data

Owoputi:

Test_O (549)	Dev (249)	Train (1576)

LexNorm:

Test_L (549)
-----------------

Data from: Chen Li, Yang Liu. Joint POS Tagging and Text Normalization for Informal Text. IJCAI 2015

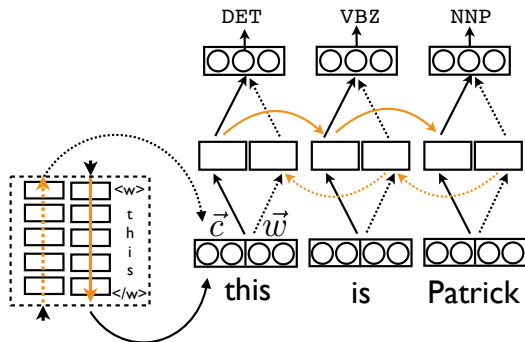
# Experimental setup

## Raw data

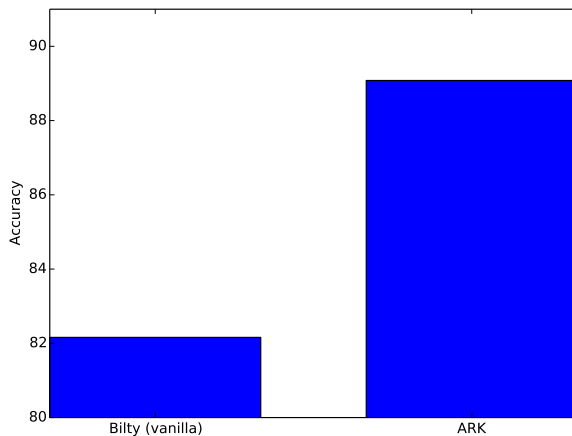
- Wikipedia
- Tweets
- No gazetteers, hard coded rules, etc.!

# Experimental setup

Bilty



# Experimental setup

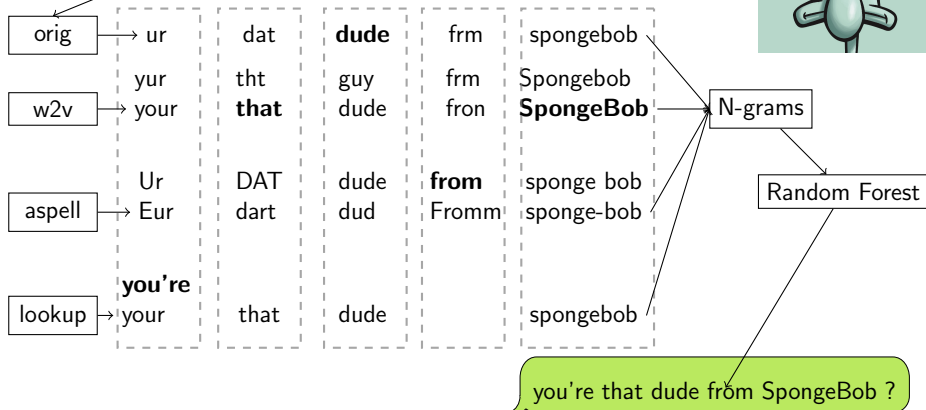




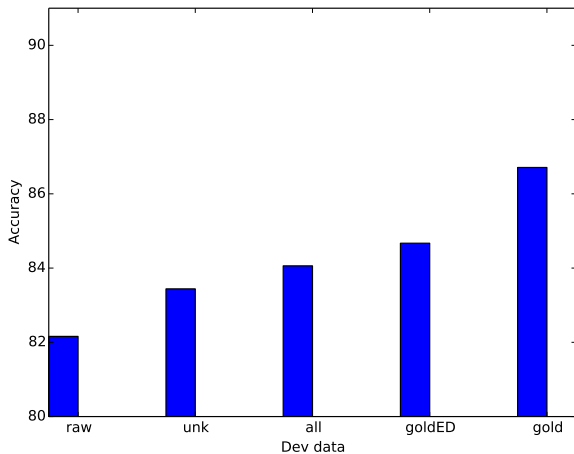
## To Normalize



ur dat dude frm spongebob ?



# To Normalize



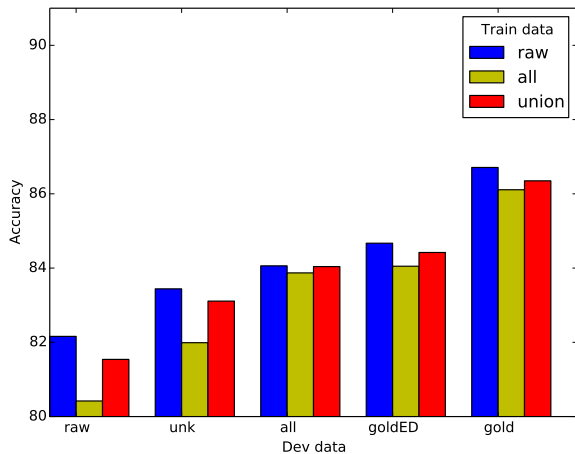
# To Normalize

non-canonical	canonical
Train	
Test	
Train	
Test	$\mapsto$

# To Normalize

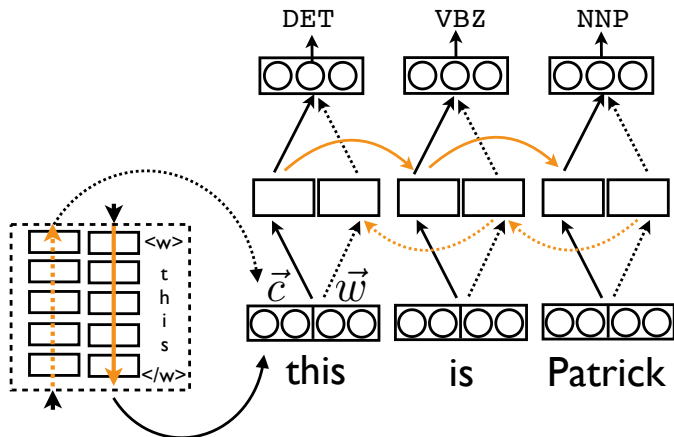
non-canonical	canonical
Train	
Test	
Train	
Test	$\mapsto$
Train	$\mapsto$
Test	$\mapsto$

# To Normalize

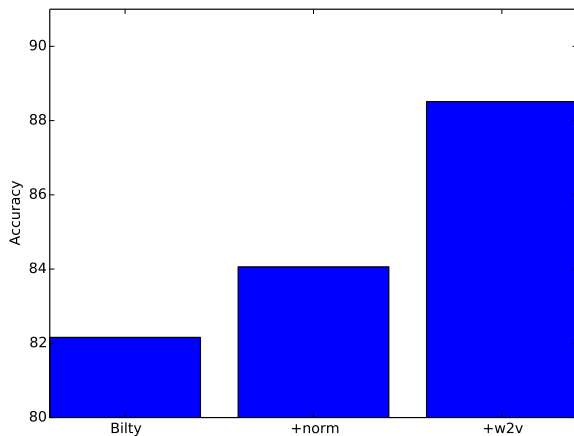


# Or Not to Normalize

## Word Embeddings



# Or Not to Normalize



# Or Not to Normalize

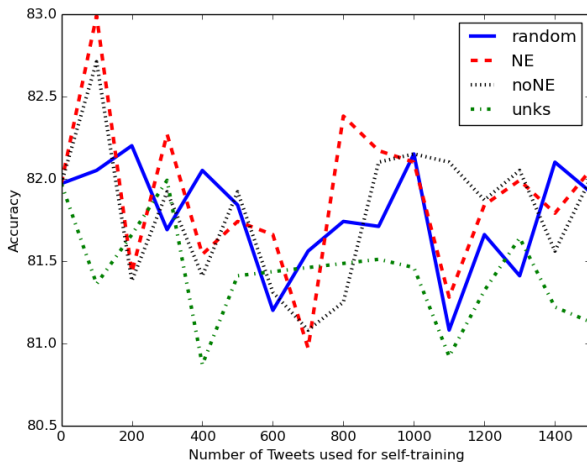
## Self training (Tweets)

- Random Tweets
- Tweets with NE
- Tweets without NE
- Tweets containing unknown words



# Or Not to Normalize

Self training (Tweets)



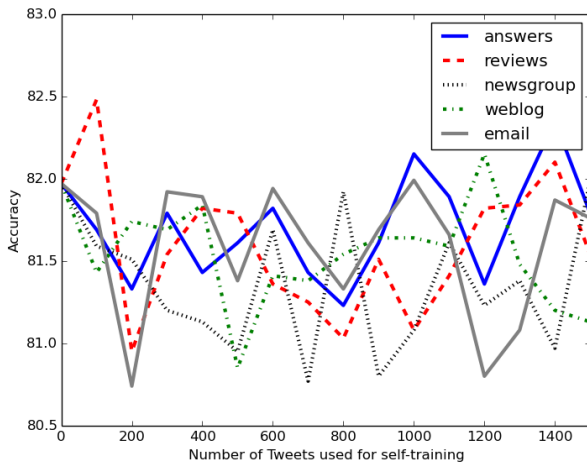
# Or Not to Normalize

Self training (EWT)

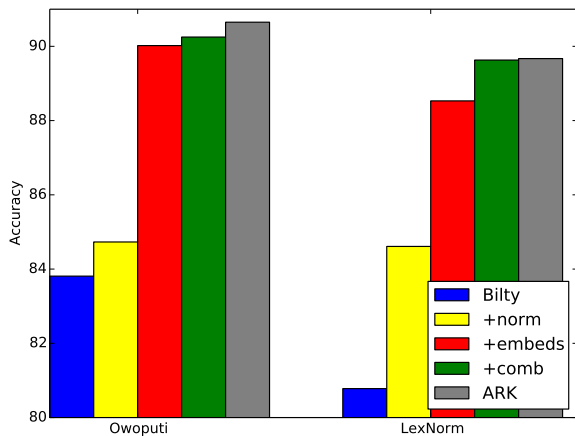
- Answers
- Reviews
- Newsgroups
- Weblog
- E-mail

# Or Not to Normalize

Self training (EWT)



# Combine



# Conclusions

- Normalization improves the baseline tagger
- Semi-supervised learning works even better
- Combining improves performance slightly
- Performance is close to ARK tagger

# Conclusions

## Negative results

- Do not normalize training data
- Self-training with pre-selection is not effective

# Conclusions

## Future work

- Self-training with post-selection
- Domain adaptation setup (train on canonical data)
- Joint/integrated approach

# Conclusions

