
Normalizing tweets with edit scripts and recurrent neural embeddings

Grzegorz Chrupała | Tilburg University

Normalizing tweets



Maully

@fvckoppz



Follow

S/o to my bro [@ti_pac7](#) fa putt n me on to
[@PhatGeez215](#)

Convert tweets to canonical form easy
to understand for downstream
applications

Examples

I will c wat i can do

I will see what I can do

imma jus start puttn it out there

I'm going to just start putting it out there

Approaches

- Noisy-channel-style
 - Finite-state transducers
 - Dictionary-based
 - Hand-crafted
 - Automatically constructed
-

Labeled vs unlabeled data

- Noisy-channel:

$$P(\text{target}|\text{source}) \propto P(\text{source}|\text{target}) \times P(\text{target})$$

labeled

unlabeled

- Dictionary lookup:
 - Induce dictionary from unlabeled data
 - Labeled data for parameter tuning
-

Discriminative model

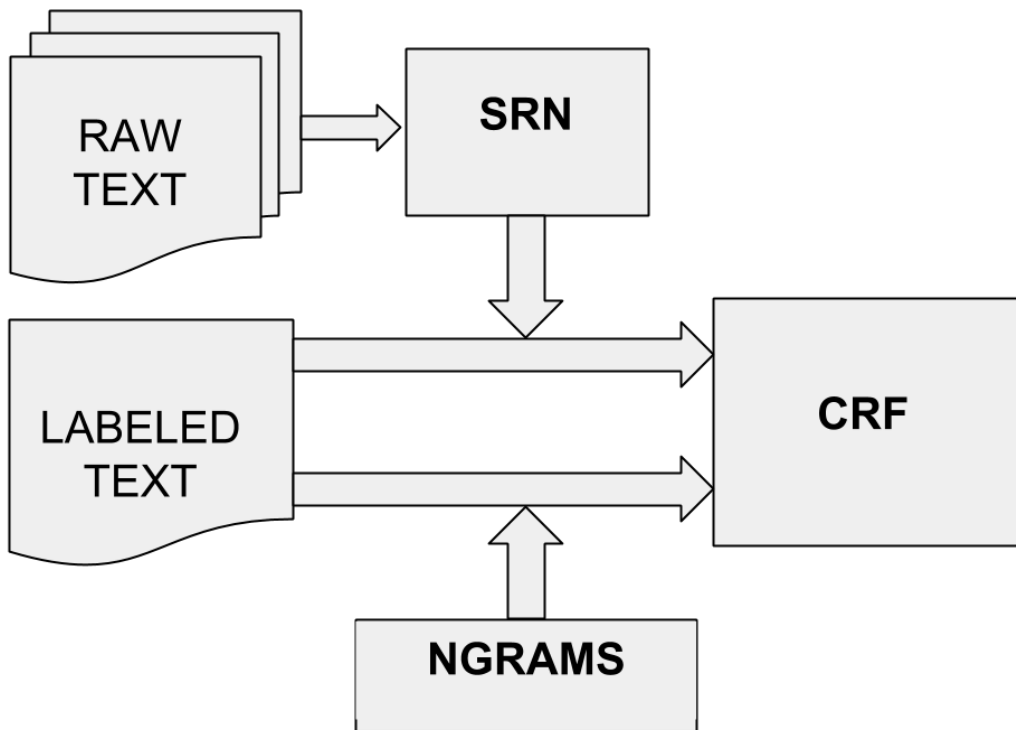
$$\underline{\text{target}}$$
$$=$$

$$\operatorname{argmax}_{\text{target}} \mathbf{P}(\mathbf{diff}(\text{source}, \text{target}) \mid \text{source})$$

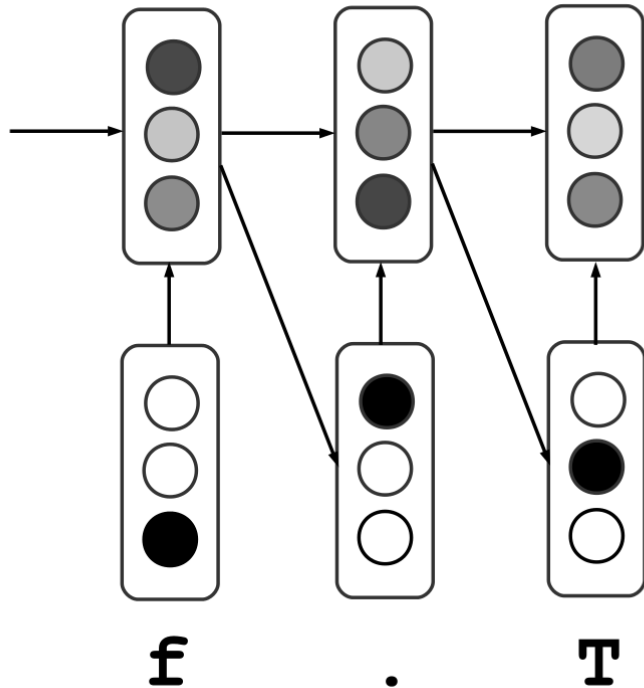
- $\mathbf{diff}(\cdot, \cdot)$ transforms source to target
 - $\mathbf{P}(\cdot)$ is a Conditional Random Field
-

Signal from raw tweets
included via
**learned text
representations.**

Architecture



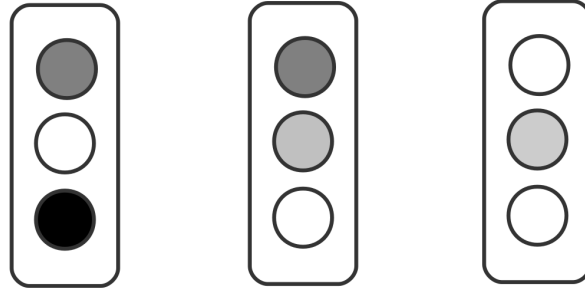
Simple Recurrent Networks



Elman, J. L. (1990). Finding structure in time. *Cognitive science*, 14(2), 179-211.

Recurrent neural embeddings

- SRN trained to predict next character
- Representation:



- Embed string (at each position) in low-dimensional space
-

Visualizing embeddings

String	Nearest neighbors in embedding space			
should h	should d	will s	will m	should a
@justth	@neenu	@raven_	@lanae	@despic
maybe	u maybe y	cause i	wen i	when i

diff - Edit script

Input	c	_	w	a	t
diff	DEL	INS(see)	NIL	INS(h)	NIL
Output		see_	w	ha	t

Each position in string labeled with edit op

Features

- Baseline n-gram features

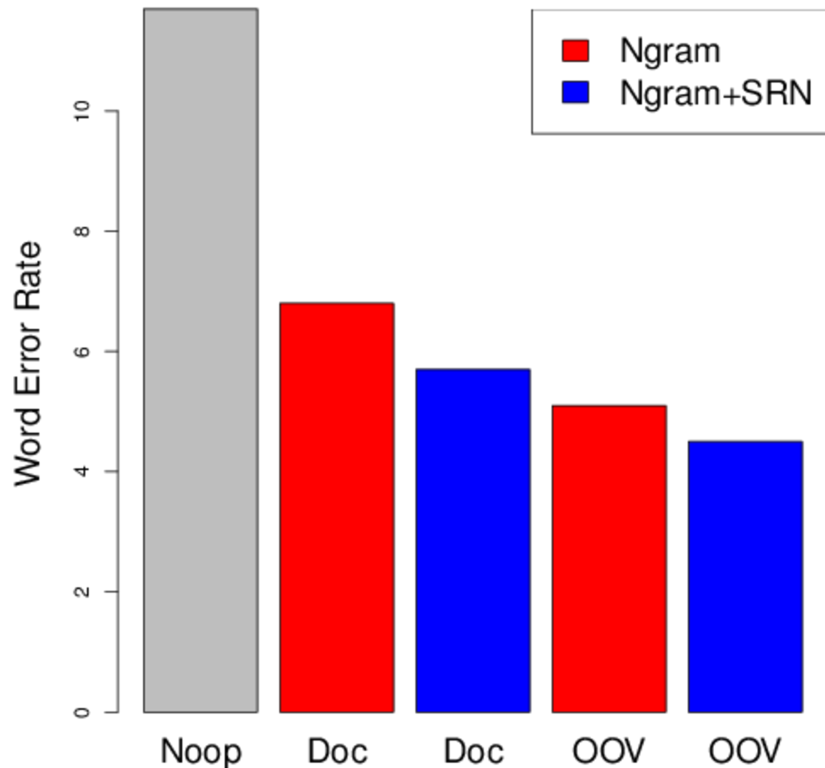
```
c _ w a t c _ _ w wa at c_w _wa  
wat c_wa _wat c_wat
```

- SRN features
 - 400 MB raw Twitter feed
 - 400 hidden units
 - Activations discretized
-

Dataset

- Han, B., & Baldwin, T. (2011). Lexical normalisation of short text messages: Makn sens a# twitter. In *ACL*.
 - 549 tweets, with normalized versions
 - Only lexical normalizations
-

Results



- **No-op**
make no changes
- **Doc**
train on and label whole tweets
- **OOV**
train on and label OOV-words

Compared to Han & Bo 2012

Method	WER (%)
No-op	11.2
S-dict	9.7
GHM-dict	7.6
HB-dict	6.6
Dict-combo	4.9
OOV NGRAM+SRN	4.7

Where SRN features helped

9 cont continued

5 gon gonna

4 bro brother

4 congrats congratulations

3 yall you

3 pic picture

2 wuz what's

2 mins minutes

2 juss just

2 fb facebook

Conclusion

- **Supervised discriminative** model performs at state-of-the-art with little training data
 - **Neural text embeddings** effectively incorporate signal from raw tweets
-