

# Topics Rob

- 1 Diachronic/Bilingual word2vec for normalization
- 2 Multiword normalization
- 3 Emerging and Rare entity recognition
- 4 Make Berkeley/Stanford parser training available for all languages

What?

new	pix	comming	tomoroe
new	picture	coming	tomorrow

How?

- ...

How?

- ...
- Fast

# Diachronic/Bilingual word2vec for normalization

What?

cont continued

aite alright

uz use

bb blackberry

# Diachronic/Bilingual word2vec for normalization

How?

- Use raw data from 2 sources, Twitter and Conanical
- Compare to normal w2v model
- Diachronic approaches assume similarities

# Diachronic/Bilingual word2vec for normalization

How?

- Use raw data from 2 sources, Twitter and Conanical
- Compare to normal w2v model
- Diachronic approaches assume similarities
- Integrate in MoNoise:  
<https://bitbucket.org/robvandergerg/monoise>

# Multiword normalization

What?

iwill	i will
gonna	going to
ni ce	nice
finttna	??



# Multiword normalization

How?

- Split all possible positions: 'i will' 'iw ill' 'iwi ll' 'iwil l'
- Ngrams: 'Im gonna sleep' 'Im going to sleep'
- Word2vec: join all common cooccurrences, and find similars in vector space
- merge: ?

# Multiword normalization

How?

- Split all possible positions: 'i will' 'iw ill' 'iwi ll' 'iwil l'
- Ngrams: 'Im gonna sleep' 'Im going to sleep'
- Word2vec: join all common cooccurrences, and find similars in vector space
- merge: ?
- Integrate in MoNoise:  
<https://bitbucket.org/robvandergerg/monoise>

# Emerging and Rare entity recognition

What?

The image shows a screenshot of a Twitter interface. At the top, there are navigation tabs for Home, Notifications, and Messages. A search bar and a 'Tweet' button are also visible. The main content area features a tweet from Rob van der Goot (@robvander) with the text 'so.. kktny in 30 mins?'. The tweet is dated 1:40 AM - 21 Apr 2017 and includes a 'Translate from Indonesian' option. A light blue reply box is overlaid on the tweet, containing a profile icon and the text 'Tweet your reply'. To the right of the tweet, there is a 'Who to follow' section with several user profiles listed, including SciHi, Matteo Castagno, and Adam Falat. The bottom of the interface shows a video player with a scene from Star Wars featuring Yoda and R2-D2. The footer of the page contains navigation icons and a copyright notice for 2017 Twitter.

# Emerging and Rare entity recognition

How?

- Baseline: OOV words
- Exploit N-grams and/or word embeddings

# Make Berkeley/Stanford parser training available for all languages

What?

- Models are now available for English, Bulgarian, Chinese etc. (Baseline)
- Pos tags for unknown words are done by hand-made categories
- This is one thing that needs to be done by hand for other languages!

# Make Berkeley/Stanford parser training available for all languages

```
if (lowered.endsWith("ed")) {
    sb.append("-ed");
} else if (lowered.endsWith("ing")) {
    sb.append("-ing");
} else if (lowered.endsWith("ion")) {
    sb.append("-ion");
} else if (lowered.endsWith("er")) {
    sb.append("-er");
} else if (lowered.endsWith("est")) {
    sb.append("-est");
} else if (lowered.endsWith("ly")) {
    sb.append("-ly");
} else if (lowered.endsWith("ity")) {
    sb.append("-ity");
} else if (lowered.endsWith("y")) {
    sb.append("-y");
}
```

# Make Berkeley/Stanford parser training available for all languages

How?

- Why not use existing pos-taggers?
- DT1: 'this' 'that' DT2: 'the' 'a' DT3: 'some' 'these'
- Optimized in parsing process.
- Automatically detect common suffixes/prefixes etc.