Mining parallel data from incomparable corpora

Filip Ginter & Jenna Kanerva
TurkuNLP
bionlp.utu.fi

Motivation

- Can we take 200M Finnish sentences and 200M English sentences and get cheap parallel data?
  - For every Finnish sentence find its best English translation candidate
  - Sort by certainty
  - Keep top 0.5%, throw remaining 99.5% away

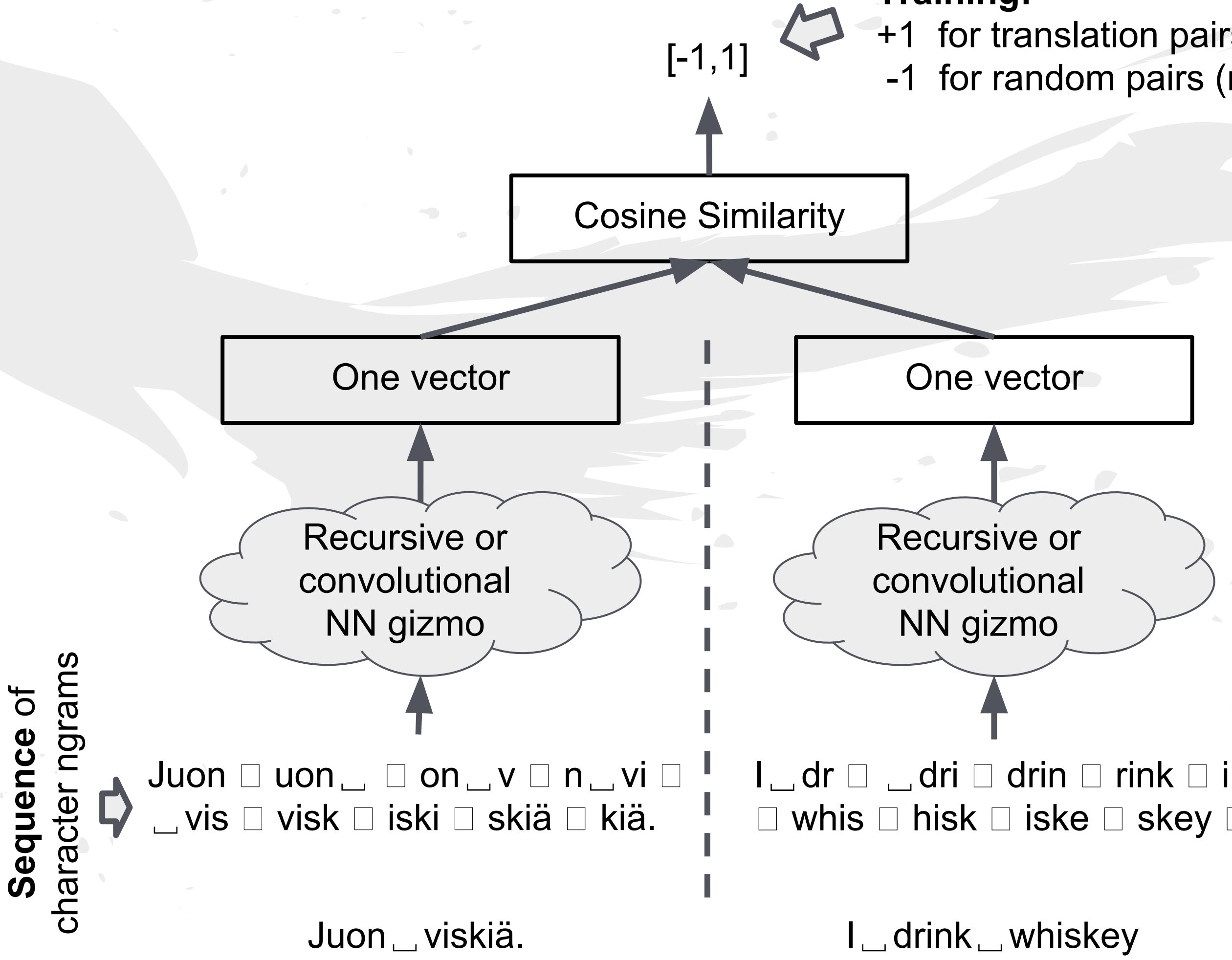
Motivation (cont.)

- Use web crawls, assume no comparable data
- Learn embeddings of sentences
- Cross-product Finnish with English
  - Maybe fast enough to do at this scale?
  - How to learn the embeddings?

Training:

+1 for translation pairs (positive examples)

-1 for random pairs (negative examples)



[-1,1]

Cosine Similarity

One vector

One vector

Recursive or convolutional NN gizmo

Recursive or convolutional NN gizmo

Sequence of character ngrams

Juon □ uon␣ □ on␣v □ n␣vi □ ␣vis □ visk □ iski □ skiä □ kiä.

I␣dr □ ␣dri □ drin □ rink □ ink␣ □ nk␣w □ k␣wh □ ␣whi □ whis □ hisk □ iske □ skey □ key.

Juon␣viskiä.

I␣drink␣whiskey

Notes:

- Languages isolated in their own half of the net
- Translation pairs get similar representation
- Forced to comply with cosine similarity

Training data

- OPUS
    - FI-EN
    - Some 17M pairs fully de-duplicated (no sentence repeated twice)
- Negative examples: random sentence pairs
- Trains in a day or so (GPU, Keras)

Example output (NN)

source: Tiedät , etten tunne niin .

reference: You would know that I do n't feel it .

- You would know that I do **n't** feel it . (sim:0.606)
- I do **n't** know . That 's what worries me . (sim:0.577)
- Okay . Look , I know it 's **not** my place to say it , but ... (sim:0.555)
- **No**body 's telling me nothing . (sim:0.537)
- It 's **not** like me at all . (sim:0.532)
- Are you sure **no** one has seen me ? (sim:0.502)
- I 'm glad I did **n't** take your advice about not coming along . (sim:0.494)
- I WAS AN ASSHOLE , OK ? BUT I CA **N'T** GO BACK AND CHANGE THAT . (sim:0.488)
- I do **n't** know why you 're not taking this `` I 'm out of here '' seriously , but I am out of here , seriously . (sim:0.481)

source: Tämä asetus tulee voimaan 10 päivänä helmikuuta 2011 .

reference: This Regulation shall enter into force on 10 February 2011 .

- It shall expire on 28 May 2010 . (sim:0.619)
- This Regulation shall enter into force on 10 February 2011 . (sim:0.615)
- The Annex to Decision 2008/377/EC is replaced by the text in the Annex to this Decision . (sim:0.594)
- FLAVOURING SUBSTANCES REFERRED TO IN ARTICLE 9 -LRB- a -RRB- OF REGULATION -LRB- EC -RRB- No 1334/2008 (sim:0.579)
- Commission Regulation -LRB- EC -RRB- No 1156/2009 (sim:0.575)
- Member States referred to in Article 5 -LRB- 4 -RRB- of Regulation -LRB- EC -RRB- No 479/2008 shall not have an obligation to fill point C and F. (sim:0.572)
- OJ L 261 , 6.8.2004 , p. 28 . (sim:0.553)

Tweaks

- Works okay(-ish) but doesn’t cut it
- Lexical overlap not strong enough
- Combine with dictionary overlap
  - Combined much better than any of the two alone

WMT 2016 test set

...just 3000 sentences!!!

Source: Tampereella karkuteillä ollut 9-vuotias poika löytyi

Reference: A 9-year-old boy missing at Tampere was found

1) 9-Year-old missing boy found at Tampere (sim: 0.550)
2) A 9-year-old boy missing at Tampere was found (sim: 0.504)
3) The police organised a search at Tampere on Tuesday evening because a 9-year old boy ran away . (sim: 0.337)
4) The boy went outside to play around 4 o'clock in the afternoon . (sim: 0.314)
5) Knowles said Prentiss , who had a hound dog named Lightning , had been dating Lamb for about three years . (sim: 0.310)

WMT 2016 test set

Source: Hän on kunnossa ja nyt aloitamme jälkiselvittelyt .

Reference: He is fine , and now we begin to settle the situation .

1) He is fine , and now we begin to settle the situation . (sim: 0.443)
2) He has taken steps in this regard . (sim: 0.284)
3) She 's in jail now . (sim: 0.247)
4) He is a fading force too . (sim: 0.227)
5) He added that he wants the Corporation to `` find a hit from our own in-house stable '' next time it launches an entertainment series . (sim: 0.224)

Real, big scale run

- we are preparing/testing it, no results yet

... we already found one example, yet some millions to go :)

What have we learned?

- Translation pairs on input and dot product on output is a good way to learn sentence embeddings
- Character n-gram sequences are surprisingly good
    - Switching over to words makes results clearly worse

Does this really work?

…we don’t know yet…