Finding translation pairs from unordered internet text

Jenna Kanerva & Filip Ginter
TurkuNLP
turkunlp.github.io

Together with Jörg Tiedemann, Robert Östling

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

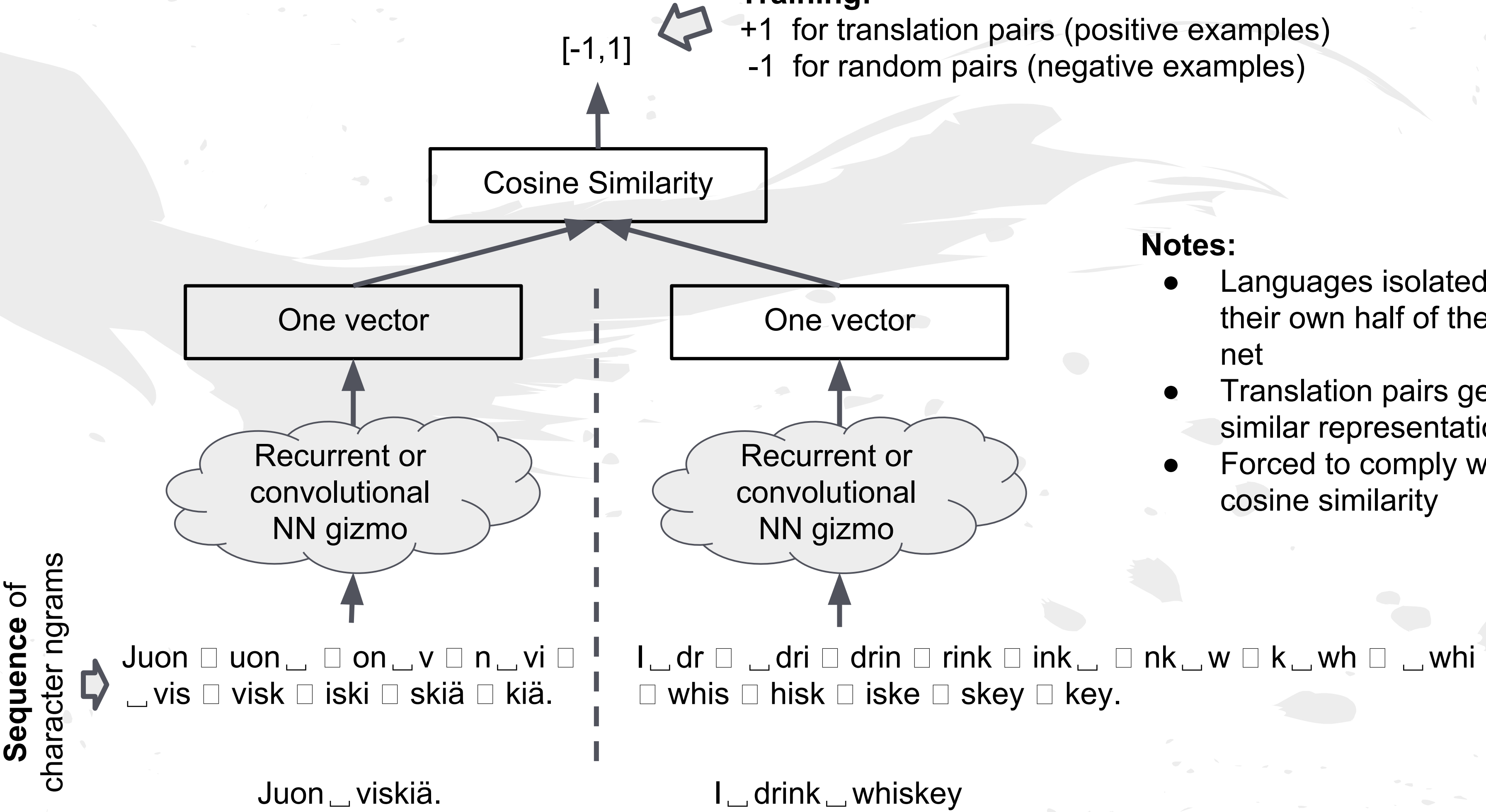
Motivation (cont.)

- Use web crawls, assume no comparable data
- English data is shuffled in sentence level, no document level features available
- Learn embeddings of sentences
- Cross-product Finnish with English
  - Maybe fast enough to do at this scale?
  - How to learn the embeddings?

Method

- Vectorize each sentence so that similar English and Finnish sentences get similar vectors
- Similarity of two sentences is a simple dot product of these vectors

Training:
+1 for translation pairs (positive examples)
-1 for random pairs (negative examples)



[-1,1]

Cosine Similarity

One vector

One vector

Recurrent or convolutional NN gizmo

Recurrent 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)

Tweaks

- NN vectors work okayish but doesn’t cut it
- Lexical overlap not strong enough
- Combine with dictionary overlap
  - Combined <u>much</u> better than any of the two alone
- Reorder top candidates with word alignment scores

Fi-En experiment

- Finnish 270M sentences
- English 420M sentences
- Filter by length and remove duplicates: 170M + 300M left
- Only a meager

51,000,000,000,000,000

candidate pairs to consider

Is it doable?

- Yes, if you have a good cluster machine (taito), but it’s not practical

- How to make it faster?
    - Reducing dimensionality? ...numpy dot product is super fast, reducing dimensionality does not make it that much faster, we need to reduce candidate pairs...

Clustering

- Take a sample of English sentence vectors and calculate cluster centers
    - 5% of sentences (15M), 1000 clusters, < 1 hour
- If we calculate the clusters using both Eng and Fin sentences, it unfortunately learns to separate these languages...

Clustering

- Now we can compare only sentence pairs inside the same cluster
- Each Finnish sentence is compared to each cluster center + each English sentence inside the nearest cluster
    - One to 1000+300K (if distributed evenly), instead of one to 300M

Clustering, technical details

- How to make this fast?
  - We cannot load different files for each sentence...
  - It still makes sense to compare only sentences of similar lengths

Clustering, technical details

- Sort sentences by clusters, and inside a cluster, by sentence length
  - (and split to suitable sized files)
- Now we can slice Fin and Eng sides accordingly, and dot these slices
  - e.g. Finnish cluster 1, sentence length 10, English cluster 1, sentence lengths 10-16

Results

- What kind of translations did we find?
  - Common topics include e.g. factual sentences, religious sentences, traveling and hotels, cooking recipes, book and movie names, common phrases, hobby translations
- Also lot of machine translations…
- Near 1M quality starts to decrease (0.5% of 170M = 850K, success!)

Results

1) **Belgium has three official languages , Dutch , French , and German .**

Belgiassa on kolme virallista kieltä , hollanti , ranska ja saksa .

2) **The patient should also avoid tea , coffee , alcohol and tobacco .**

Potilaan pitäisi myös välttää teetä , kahvia , alkoholia ja tupakkaa.

3) **The West and Russia compete with economic offers , but identity is probably more important .**

Länsi ja Venäjä kilpailevat taloudellisilla tarjouksilla , mutta identiteetti on todennäköisesti tärkeämpi .

4) **HE IS NOT HERE : FOR HE IS RISEN , AS HE SAID .**

Hän ei ole täällä : hän on noussut , kuten hän sanoi .

Results

10K) It was the last time the two men saw each other .

Se oli viimeinen kerta , kun miehet näkivät toisiaan .

20K) Cornelius was expecting them and had called together his relatives and close friends .

Cornelius tietenkin odotti heitä ja oli kutsunut koolle sukulaisensa ja läheiset ystävänsä .

50K) " And that you serve Me .

" Ja sinä palvelet minua .

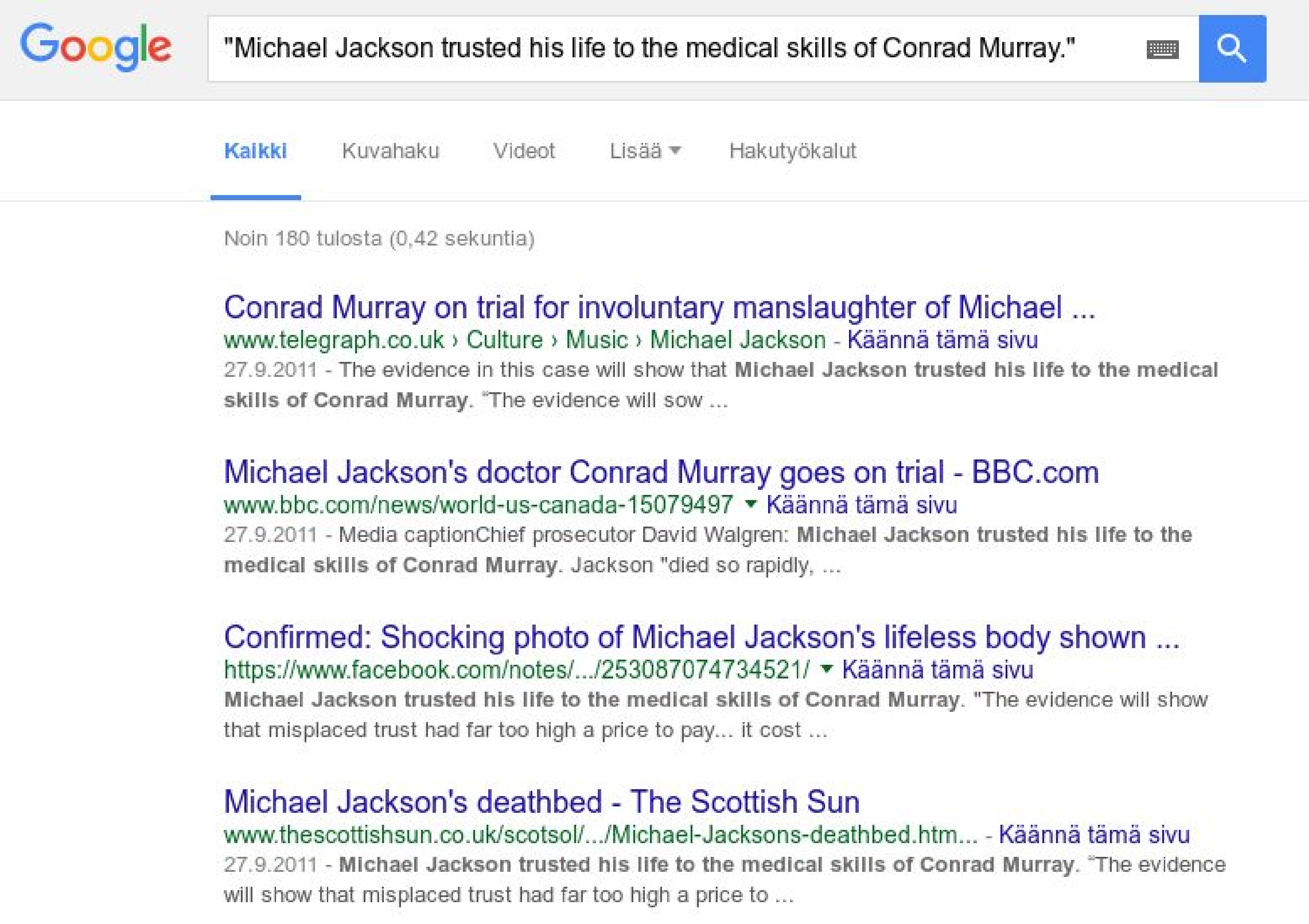
100K) Registration is easy and free .

Rekisteröityminen on helppoa ja ilmaista

Example findings

Fi: Michael Jackson luotti henkensä Conrad Murrayn lääkinnällisten taitojen varaan .

En: Michael Jackson trusted his life to the medical skills of Conrad Murray .



"Michael Jackson trusted his life to the medical skills of Conrad Murray."

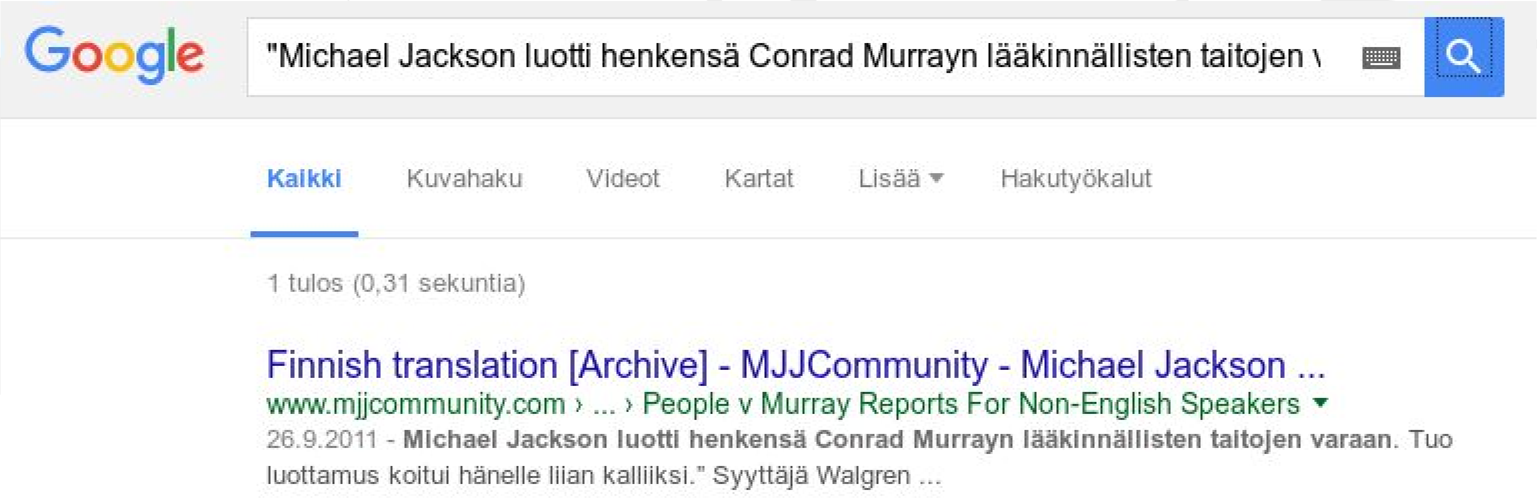
Noin 180 tulosta (0,42 sekuntia)

Conrad Murray on trial for involuntary manslaughter of Michael ...
www.telegraph.co.uk › Culture › Music › Michael Jackson - Käännä tämä sivu
27.9.2011 - The evidence in this case will show that **Michael Jackson trusted his life to the medical skills of Conrad Murray**. "The evidence will sow ...

Michael Jackson's doctor Conrad Murray goes on trial - BBC.com
www.bbc.com/news/world-us-canada-15079497 ▾ Käännä tämä sivu
27.9.2011 - Media captionChief prosecutor David Walgren: **Michael Jackson trusted his life to the medical skills of Conrad Murray**. Jackson "died so rapidly, ...

Confirmed: Shocking photo of Michael Jackson's lifeless body shown ...
https://www.facebook.com/notes/.../253087074734521/ ▾ Käännä tämä sivu
Michael Jackson trusted his life to the medical skills of Conrad Murray. "The evidence will show that misplaced trust had far too high a price to pay... it cost ...

Michael Jackson's deathbed - The Scottish Sun
www.thescottishsun.co.uk/scotsol/.../Michael-Jacksons-deathbed.htm... - Käännä tämä sivu
27.9.2011 - **Michael Jackson trusted his life to the medical skills of Conrad Murray**. "The evidence will show that misplaced trust had far too high a price to ...



"Michael Jackson luotti henkensä Conrad Murrayn lääkinnällisten taitojen v

1 tulos (0,31 sekuntia)

Finnish translation [Archive] - MJJCommunity - Michael Jackson ...

www.mjjcommunity.com › ... › People v Murray Reports For Non-English Speakers

26.9.2011 - **Michael Jackson luotti henkensä Conrad Murrayn lääkinnällisten taitojen varaan**. Tuo luottamus koitui hänelle liian kalliiksi." Syyttäjä Walgren ...

http://www.mjjcommunity.com/forum/archive/index.php/t-119226.html