CroVeWA: Crosslingual Vector-Based Writing Assistance

Hubert Soyer *‡ — *hubert.soyer@gmail.com*

Goran Topić * — *goran_topic@nii.ac.jp*

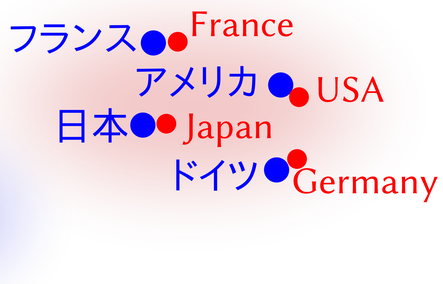
Pontus Stenetorp † — *pontus@stenetorp.se*

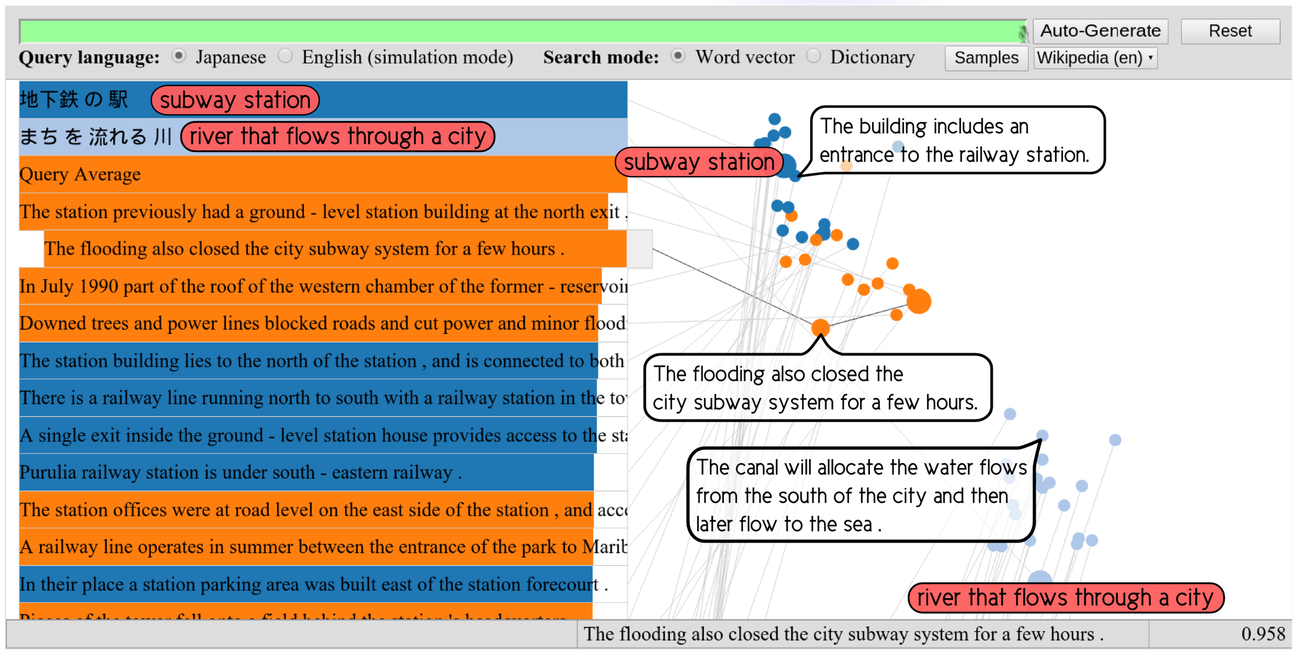
Akiko Aizawa * — *aizawa@nii.ac.jp*

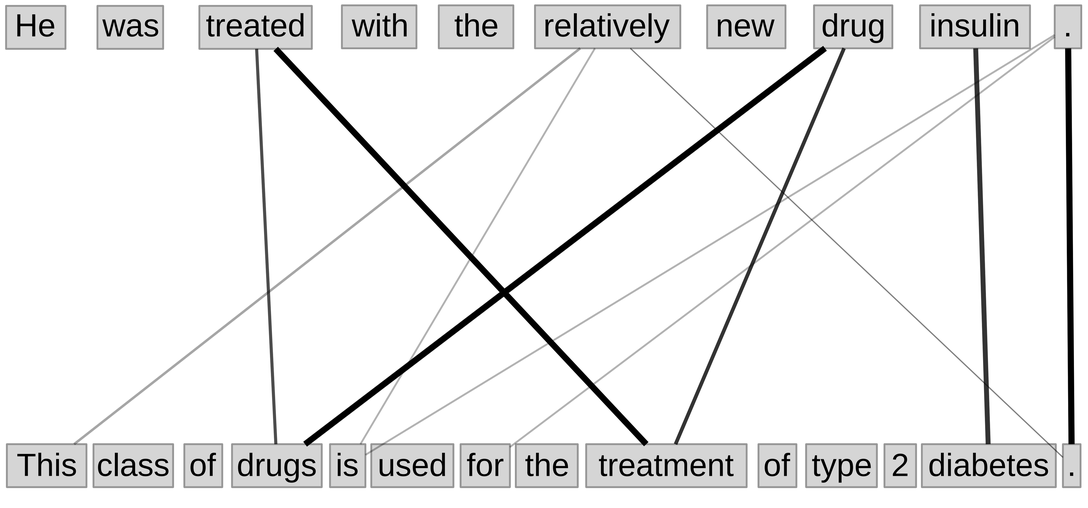
* National Institute of Informatics, Tokyo, Japan † University College London, London, UK ‡ now at Google DeepMind

CROSSLINGUAL WRITING ASSISTANCE

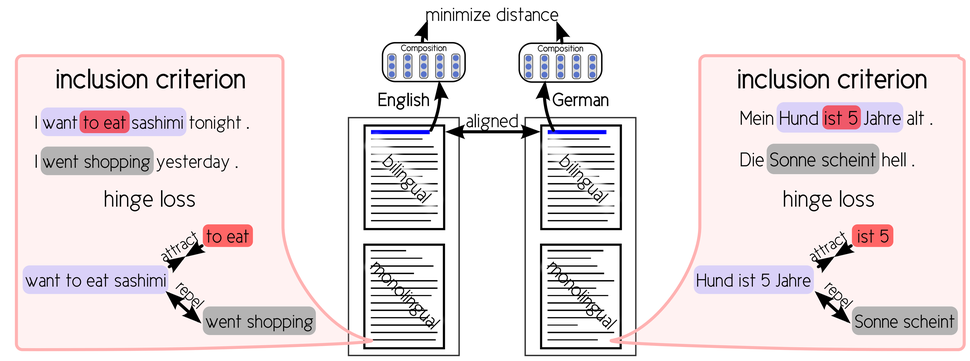
- Input: Japanese/English query
- Retrieve semantically related sentences from high-quality English corpora
- Based on Compositional Distributed Semantics
- Visualize phrase relationships
- Retrieve sentences from English-only corpora
- Word-to-word correspondences







WORD REPRESENTATIONS



Leveraging monolingual data for crosslingual compositional word representations

Hubert Soyer, Pontus Stenetorp, and Aizawa Akiko - ICLR 2015

Induce crosslingually constrained vector representations of words leveraging bilingual and monolingual resources

"Keep vectors of translations close" (bilingual corpus)

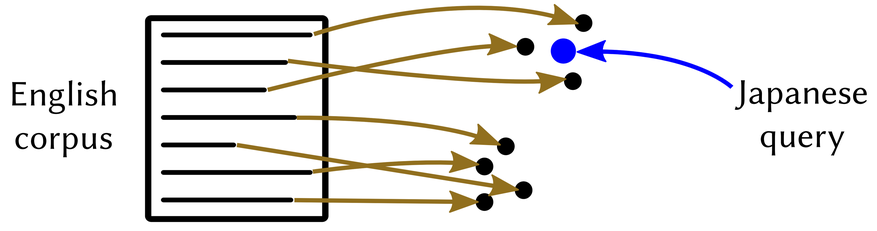
"Sub-phrases are usually closer to their mother phrase than to random other phrases" (monolingual corpora)

PRE-PROCESS (ENGLISH) CORPUS

For every phrase in high-quality English corpora

1.) Compose vector representation

2.) Insert into index of nearest neighbor search engine



(JAPANESE) QUERY LOOKUP

1.) Compose vector representation of query

2.) Find nearest neighbors among English corpora vectors

VISUALIZATION

Sentence-level visualization

- High-dimensional sentence and query vectors ⟶ 2D
  - Inspired by Multi Dimensional Scaling
  - Queries as points of reference

Word-to-word correspondences

- Sentence pair - visualize similarities between word-vectors