Crosswalk: Aligned Crosslingual Word and Entity Embeddings

Alberto García Durán on behalf of EPFL dlab and INRIA magnet

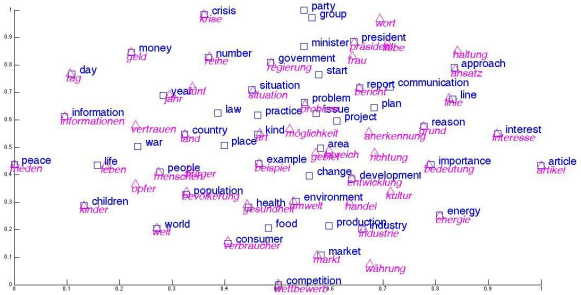
INRIA-EPFL workshop

Overview

- Aligned crosslingual word and entity embeddings
  - Related work
  - Applications
- Technical details
  - High-level idea
  - Challenges
- Personnel

Aligned crosslingual word and entity embeddings

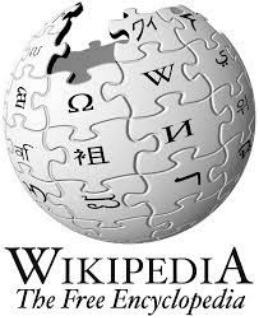
- Crosslingual word embeddings.

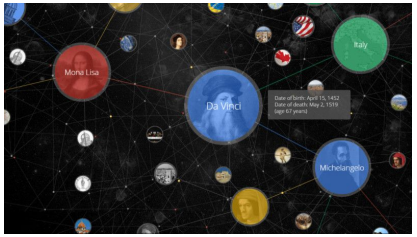


- Transfer knowledge from a high resource language to a low resource language.
  - Text Classification (HR: english, LR: sicilian)
  - Language modeling (HR: italian, LR: sicilian)
  - …
- Off-line vs Joint learning [Ormazabal et al. 2019]
  - Offline: Independent learned word embeddings are mapped to the same space
  - Joint: Word embeddings in multiple languages are learned jointly

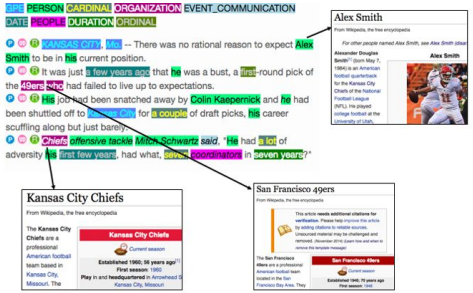
Aligned crosslingual word and entity embeddings

- Text linked to a knowledge graph





- Entity Linking: task of grounding mentions to entities in a knowledge graph



Aligned crosslingual word and entity embeddings

- Entity Linking - Candidate Generation

  Alex Smith = {alex_smith_singer, alex_smith_actor, alex_smith_player}

  49ers = {san_francisco_49ers_NFL, san_francisco_49ers_NBA}

Constructed by crawling the web!!!!

- Entity Linking - Disambiguation
  - Surrounding words are topically related to the right entity
  - Offline: Word and entity embeddings are learned independently
  - EL methods learn to align from **annotated data.**

- Vast majority of EL methods are evaluated in english

Aligned crosslingual word and entity embeddings

What if entity and crosslingual word embeddings are learned jointly?

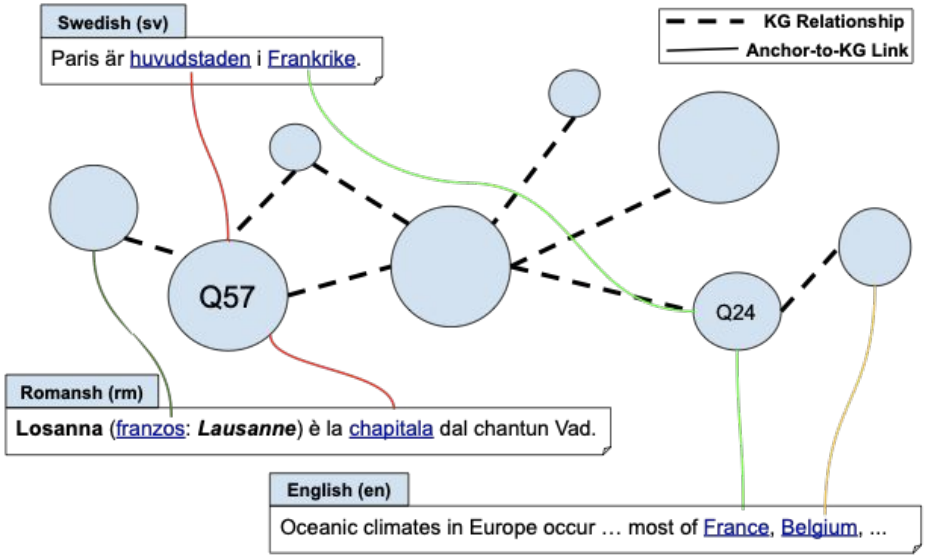
- Entity Linking - Candidate Generation
  - We can leverage the aligned embeddings to retrieve candidates
  - Learning dense representations for entity retrieval [Gillick et al. 2019]

Constructed by crawling the web!!!!

- Entity Linking - Disambiguation
  - EL method trained in a HR language with a lot of **annotated data**...
  - ... and evaluated in a LR language.

Technical details

- Crosswalk: Text (Wikipedia) + Graph (Wikidata)
  - Wikidata entities are language-agnostic
  - Q24: France, Francia, Frankreich...



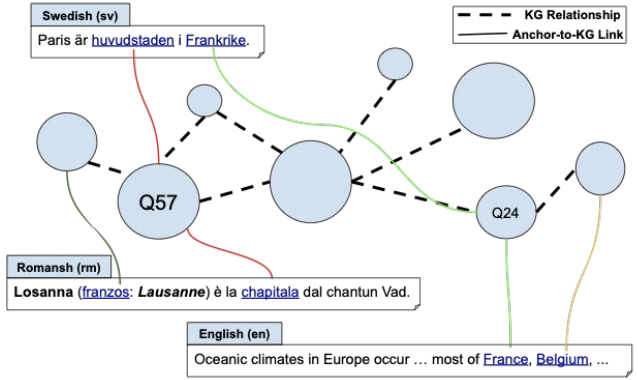
Technical details

- Entities are the connecting points across languages
- Generate “fake sentences” by jumping between the graph and the text

Losanna (franzos: Lausanne) e la chapitala del chantun Vad

Losanna (franzos: Lausanne) e la Q57 i Frankrike

Losanna (franzos: Lausanne) e la Q57 … Q24 Belgium



- Horizontal alignment + Vertical alignment

Technical details

- Apply standard techniques to representation learning

- Co-occurrence matrix + SVD
  - Improving distributional similarity with lessons learned from word embeddings [Levy et al. 2015]
  - Facebook SVD: FBPCA

- Challenges
  - A lot of the work in the background (redirect, disambiguation pages, identifier mapping)
  - Scalability. Only the knowledge graph already contains more than 20M entities
  - Number of fake sentences around an entity node?
  - Do not increase artificially co-occurrences of words

Personnel

- Crosswalk: NLP + Wikipedia + Graph
  - INRIA magnet - NLP
  - EPFL dlab (Robert West) - Wikipedia
  - EPFL dlab (Akhil Arora) - Scalability
  - EPFL dlab (me) - Knowledge Graph

THANK YOU ! QUESTIONS?

Data

- Wikidata graph
  - Full Wikidata graph has ~45M nodes
  - Version of Wikidata used for a previous project (~4M entities)
  - (FBpca) SVD takes < 10 minutes for the graph with ~4M entities and ~22M edges

- Wikipedia text
  - Apply wikipedia extractor, apply redirect, remove disambiguation pages, Wikipedia2Wikidata…

```
'Farben' chord  Farben chord
'Farmer' Burns  Martin Burns
'Farmingdale Post       Farmingdale, New York
'Fearless' Freddie Williams     Freddie Williams (businessman)
'Fessor Graham  Floyd Graham
'Fessor Graham Award    Floyd Graham
'Foghorn' Winslow       George Winslow
```

  - Tokenization, accent removal, punctuation removal
  - Vocabulary Size

Random walks

- Examples of edges: head \t tail \t distance \t doc_id \t position of the head

```
various factions          1       12      229
factions        within    1       12      230
within  the     1         12      231
the     ATXT:french_revolution    1       12      232
ATXT:french_revolution    Q6534   0       12      233
ATXT:french_revolution    labelled        1       12      233
Q6534   labelled          1       12      233
labelled        their     1       12      234
their   opponents         1       12      235
```

- Number of nodes proportional to vocab size & number of entities in Wikidata
- Edges maintain sequential information
  - Different edge types: w2w, w2e, e2w, e2e
- How many random walks to sample from each node?
  - Most of edges are w2w -> Deterministic random walks

Applications

- We will learn entity, (multi-lingual) word and anchor text embeddings that are aligned in the same embedding space.
- We want to showcase the utility of the crosswalk embeddings in applications where BOTH entity and word (anchor text) embeddings are required.
    - Entity linking. (Multi-lingual) EL systems need both word and entity embeddings.
    - Entity retrieval. Aiming to replace alias tables in the candidate generation module of EL.
        - Recently introduced in https://arxiv.org/pdf/1909.10506.pdf.
        - Different to that work:
            - Entities are not represented by their descriptions.
            - We can learn an entity retrieval system in a language (e.g. english), and apply it to a different language.
- While it might be interesting to evaluate cross-walk embeddings in applications where either only entity (e.g. entity relatedness) or only multi-lingual word embeddings (e.g. machine translation) are required, they will not be the primary use case. The main reason is that we think that dedicated models, such as BERT in the case of machine translation, are very hard to beat.