Document Embedding Models on Environmental Legal Dataset

Samo Kralj, Živa Urbančič, Erik Novak, Klemen Kenda

Jožef Stefan Institute

October 7, 2019

Brief Overview

- **DATA:** sources, collecting process, metadata
- **METHODOLOGY:** word embedding models, construction of document embeddings
- **PRELIMINARY RESULTS**

Data

Data sources

ECOLEX

Led by *Food and Agriculture Organisation of the United Nations* (FAO), the *International Union for Conservation of Nature* (IUCN) and *United Nations Environment Programme* (UNEP).

EURLEX

European Union law.

Documents were collected using dedicated **web crawlers**.

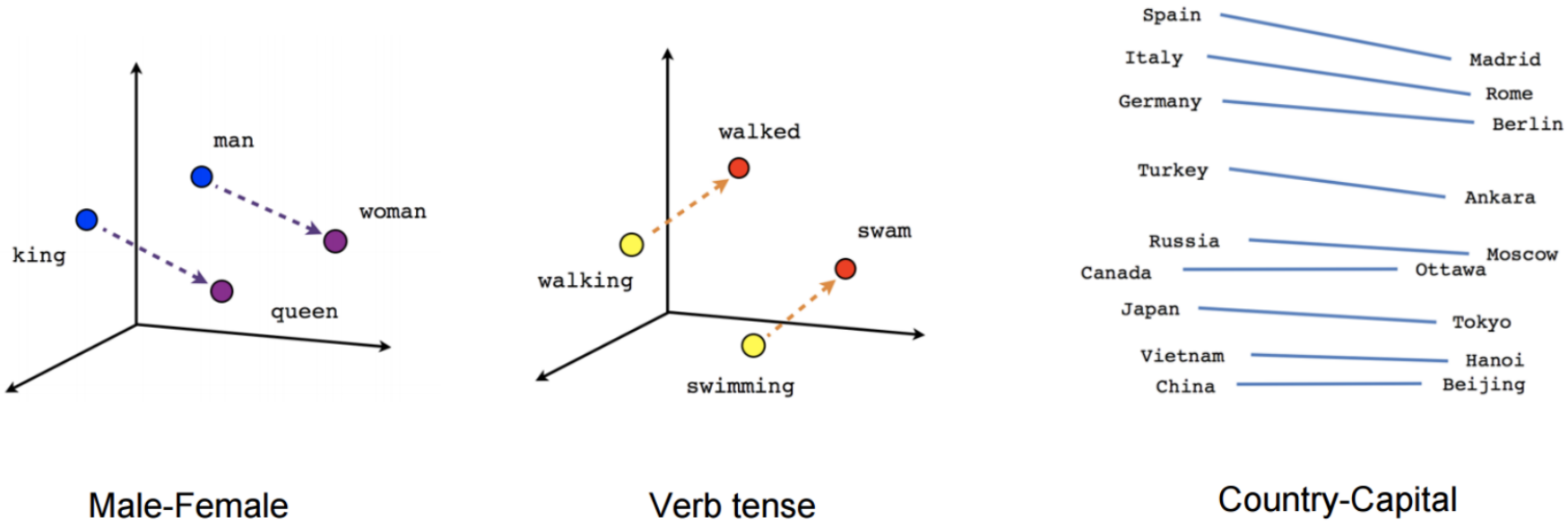
Data

Characteristics

- **220k** documents from ECOLEX, **800k** documents from EURLEX.
- EURLEX contains entire European Union law. **75k** documents were considered appropriate for our task.
- Metadata: title, authors, dates, subject, . . .
- Dataset specific metadata: **keywords** in ECOLEX dataset and **descriptors** in EURLEX dataset serve the same purpose.

Methodology

Word Embedding



Male-Female | Verb tense | Country-Capital

Figure: Word relationships captured by word embeddings.

Methodology

Word Embedding Models

Chosen models:

- pre-trained fasttext model (2,5M tokens)
- newly trained fasttext model, trained using **gensim** library (500k tokens)

Methodology

Document Embedding Models

Document Embedding as the Average Word Embedding

Let $W = \{w_1, w_2, \ldots, w_n\}$ be a list of words that appear in a document, and let x_i be the embedding of the word w_i.

$$d = \frac{1}{|W|} \sum_{w_i \in W} x_i.$$

Approaches:

- Use all the words that appear in the document.
- Use the keywords/descriptors.
- A combination.

Methodology

Document Embedding

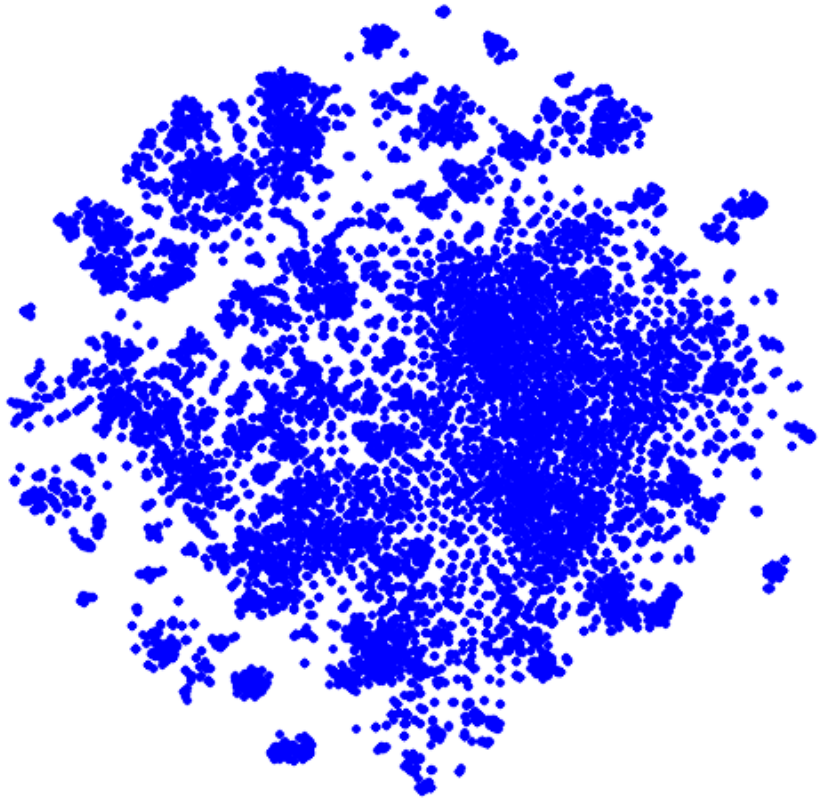


Figure: Planar projection of document embeddings of the first 15k English documents in the EURLEX corpus.

Preliminary Results

- Choosing arbitrary source document,
- performing k-nearest neighbors search,
- manually evaluating the results.

Preliminary Results

Example

1. **Source:** Agreement on fisheries between the European Economic Community and the Government of Canada . . .
2. Council Decision of 29 December 1981 on the conclusion of an Agreement on fisheries between the European Economic Community and the Government of Canada
3. Agreement on fisheries between the European Economic Community and the Government of the Commonwealth of Dominicia - Protocol on conditions relating to reciprocal access for fishing vessels of both Parties
4. Agreement on fisheries relations between the European Community and the Republic of Estonia
5. Agreement on fisheries relations between the European Community and the Republic of Latvia

Preliminary Results

Some observations:

- Even though the *pre-trained model performs better in general*, the newly trained model outperforms it in cases when source document is not particularly similar to other documents in the database.
- Length of the documents: when we use all words of the document to create the document embedding, documents with similar lengths will be closer. This is not as important when we only use document's keywords or descriptors.
- Documents with similar legal expressions and structure have closer embeddings.

Future Work

- Develop a service for searching through our database.
- Provide support for languages other than English.
- Evaluate and improve our model using the user's feedback.