Federal Research Center «Computer Science and Control»
of the Russian Academy of Sciences

# Cross-lingual similar document retrieval methods

Zubarev D.V.
Sochenkov I.V.

Moscow 2019

# Cross-language plagiarism detection task

- Source retrieval
  - given: a suspicious document and a large collection of sources
  - the task is to retrieve all plagiarized sources while minimizing retrieval costs
- Text alignment
  - given: a pair of documents
  - the task is to identify all contiguous maximal-length passages of reused text between them

# Cross-language plagiarism detection methods

- Machine translation approach: use a monolingual methods by translating query document (Barŕon-Cedẽno, 2012; Bakhteev, 2019)
- Cross-lingual similarity detection based on the word embeddings (Ferrero et al., 2017; Kutuzov et al., 2016)
- **Competitions**
  - Translated plagiarism cases were used in PAN 2011 competition as one of an obfuscation type (without Russian)

# Training cross-lingual word embeddings

- Russian-English parallel sentences
  - United Nations – first 2 million sentences (opus.nlpl.eu)
  - Wiki – 600K sentences (opus.nlpl.eu)
  - Yandex parallel corpus – 1M sentences
  - Tatoeba – 500k; collection of sentences for foreign language learners (opus.nlpl.eu)
  - QED – 600k; collection of subtitles for educational videos and lectures (opus.nlpl.eu)
  - JW300 – ~1M; texts from jw.org, mainly Watchtower and Awake! (opus.nlpl.eu)

# Text Preprocessing

- Tokenization, lemmatization and syntax analysis (aot, udpipe)
- Stop-words removal (conjunction, pronoun, etc.), and common words (be, the, a)
- syntactic phrases up to 4 words are treated as single “word”:

*martial_law*, *военное_положение*, *Организация_объединенных_наций*

- Result: pairs of parallel sentences, represented as sequences of lemmas

# Training corpus generation

- Generate multiple sentences from a sentence with a phrase:
- Russian_presidential_election ...
- Russian_election presidential_election …
- Russian presidential election ...

- 10 million sentences
- Dictionary size: 680k unique words/phrases (more than 10 occurrences in the corpus)

# Training embeddings

- Learning cross-lingual word embedding mapping
  - Using Vecmap framework (Artetxe et al. 2018)
    - Supervised with dictionary ~20k word pairs (from Muse project)
- Training word2vec model on bilingual corpus (Vulić et al. 2015)
  - Select pairs of sentences that differ in length by less than 5 words
  - Interleave two parallel sentences, e.g.

"Мама мыла раму" + "Mother washed the frame“ =
"мама mother мыла washed раму the frame".

# Retrieval-based approach: Indexing texts

- Tokenization and lemmatization, syntax analysis (aot, udpipe)
- Extract noun phrases up to 4 words
- Build inverted index of words, phrases:
  word_id -> doc_id, weight (LogTF)

# Retrieval-based approach: Search

- Represent the query document as a weighted vector
- Use LTF-IDF as the weighting scheme
- Select top words/phrases and map them to N other language keywords with cross-lingual embeddings

- Fetch documents from inverted index based on matched words/phrases
- Measure similarity (cosine or hamming) between query vector and other texts' vectors

# Approximate nearest neighbor search

- Represent each document as a vector by averaging vectors of the top K keywords of the document

- Index all vectors with ANN index (Faiss, IVF_SQ16)

- ANN-search at query time for a given document

# Cross-Lingual Explicit Semantic Analysis

- Represent a document as a weighted vector of concepts (Wikipedia articles):
  - We used 800k English articles that are aligned with Russian Wikipedia articles
  - Components of a vector are cosine similarity of a given document with each article
- At query time:
  - Document -> weighted vector of articles` identificators
  - Map identificators to articles in other language
  - Retrieve the most similar vectors in collection of other language

# Dataset

- Russian-English aligned Wikipedia articles (Wikipedia dump of June 2019)
- We sampled article pairs from 10 different groups
- Grouping by:
  - Amount of Russian sentences:
    *(9, 50], (50, 100], (100, 200], (200, 400], (400, 1000]*
  - Comparable/Non-comparable by size
- Sampled 100 document pairs from each group
- 1000 document pairs

# Evaluation results

- RBA – retrieval-based approach
- ANN – Approximate nearest neighbor search

| Method | Embeddings | DIM | Phrases | Recall | MAP | Rec@1 | Rec@10 | Rec@20 |
|---|---|---|---|---|---|---|---|---|
| RBA | Bilingual | 300 | No | 0.831 | 0.48 | 0.415 | 0.622 | 0.66 |
| RBA | Bilingual | 300 | 4-word phrases | <u>0.845</u> | <u>0.49</u> | <u>0.415</u> | **0.635** | <u>0.67</u> |
| RBA | Bilingual | 600 | 4-word phrases | **0.856** | **0.5** | **0.428** | <u>0.629</u> | **0.671** |
| RBA | Mapping | 300 | 4-word phrases | 0.767 | 0.336 | 0.263 | 0.478 | 0.533 |
| ANN | Bilingual | 300 | 2-word phrases | 0.728 | 0.398 | 0.337 | 0.508 | 0.548 |
| ANN | Bilingual | 600 | 2-word phrases | 0.724 | 0.433 | 0.374 | 0.527 | 0.577 |
| ANN | Mapping | 300 | 2-word phrases | 0.665 | 0.254 | 0.197 | 0.36 | 0.426 |
| CL-ESA | - | - | 4-word phrases | 0.833 | 0.318 | 0.254 | 0.453 | 0.501 |

# Evaluation results per each group for RBA

| No | Size in ru sents | Comparable by size? | MAP (Top=100) | Rec (Top=100) | MAP (Top=50) | Rec (Top=50) |
|---|---|---|---|---|---|---|
| 1 | (9, 50] | False | 0.346 | 0.82 | 0.346 | 0.82 |
| 2 | (9, 50] | True | 0.338 | 0.65 | 0.338 | 0.65 |
| 3 | (50, 100] | False | 0.419 | 0.79 | 0.419 | 0.8 |
| 4 | (50, 100] | True | 0.44 | 0.88 | 0.445 | 0.88 |
| 5 | (100, 200] | False | 0.461 | 0.81 | 0.453 | 0.82 |
| 6 | (100, 200] | True | 0.542 | 0.88 | 0.535 | 0.9 |
| 7 | (200, 400] | False | 0.451 | 0.79 | 0.473 | 0.8 |
| 8 | (200, 400] | True | 0.730 | 0.98 | 0.742 | 0.97 |
| 9 | (400, 1000] | False | 0.306 | 0.85 | 0.341 | 0.81 |
| 10 | (400, 1000] | True | 0.87 | 1 | 0.871 | 1 |

# Conclusions and future work

- Retrieval-based approach achieved best recall and MAP on Wiki dataset
- MAP/Recall are generally better for articles from the "Comparable by size" group

- Try combination of various embeddings Bilingual + Mapping
- Try other retrieval functions

Federal Research Center «Computer Science and Control»
of the Russian Academy of Sciences

# Cross-lingual similar document retrieval methods

Dataset - http://nlp.isa.ru/ru-en-src-retr-dataset/

**Zubarev D.V.**
**Sochenkov I.V.**

Moscow 2019