Institute for Systems Analysis
Federal Research Center «Computer Science and Control»
of the Russian Academy of Sciences

Paraphrased Plagiarism Detection Using Sentence Similarity

Zubarev D.V. – PhD student
Sochenkov I.V. – PhD

+7 (499) 135-04-63
117312, Moscow
pr. 60-letiya Oktyabrya, 9

Plagiarism detection task

- Two subtasks
  - 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

Related work

- Source retrieval
  - Querying search engines
  - Methods revolve around selecting keywords
  - Many heuristics for candidates filtering
- Text alignment
  - Many methods exist based on N-grams, skip N-grams, syntactic N-grams, stop words N-grams
  - Vector space models with cosine similarity are also widely used
- There was competition PAN (2009-2015)

Preprocessing of collection of sources

- Index all sources for future fast retrieval
- Store extra information about each word (PoS-tags, semantic roles, etc.)
- Some statistics of a source collection:
  - 5,7 million texts
  - 130 GB – raw size
  - 229 GB – size of indexes

Document preprocessing: Linguistic analysis

- Perform deep natural language processing of the uploaded text
  - POS-tagging
  - Syntactic parsing
  - Semantic role labeling
  - Semantic relation extraction

Shelmanov A. O., Smirnov I. V. Methods for semantic role labeling of Russian texts, Dialogue 2014

First stage: Candidates retrieval (source retrieval)

- Employ Vector Space Model and modified Hamming distance
- Use some noun phrases along with words for creating a vector
- Words and phrases are weighted (TF-IDF)
- Only top 100-200 are used
- The 600 most similar documents are retrieved on this stage
- They are called candidates

Second stage:
Suspicious sentences selection

- Filter sentences based on various criteria:
  - a TF-IDF weight of a sentence
  - a length of a sentence
  - an amount of non-alphanumeric symbols in a sentence
- TF-IDF weighting schema is used
- IDF weights are calculated based on word frequencies in all collections
- Top 10000 weighted sentences are selected

Second stage:
Fast filtering˚ (Text alignment)

- Intersect each selected sentence from the suspicious document with all other sentences from the candidates

- Apply fast algorithm for estimation of the size of intersection for filtering most irrelevant sentences with unmatched lexis

- Pairs of sentences that share at least 35 % of words are passed to the next stage

Third stage:
Sentence similarity (Text alignment)

- Calculate multiple similarities of each pair using different measures:
  - lexis similarity measure
  - syntactic similarity measure
  - semantic similarity measure
- Combine each obtained value into overall similarity
- Pairs that exceed predefined similarity threshold are considered to be incorrectly reused fragments

Tuning plagiarism detection method (Random search)

- 13 parameters to tune:
  - each of them has from 10 to 20 values
- Initialize each parameter with random value
- On each iteration
  - Slightly tweak each parameter by increasing/decreasing its value
  - Measure performance
  - Choose the best combination
  - Repeat
- Interrupt when the performance of the detection method is not changed for a while
- Repeat the whole search with a new seed

Evaluation corpus from PlagEvalRus 2017

- Source retrieval:
  - Sources collection: 5.7 million documents
  - training set: 671 suspicious documents
  - Test set: 10k suspicious documents

- Text alignment:
  - training set: 9k pairs
  - Test set:
    - ~10k pairs
    - available only on evaluation platform Tira

Evaluation corpus (2)

- Evaluation corpus includes plagiarism cases with various obfuscation types:
  - **Essay-1** – manually written essays with plagiarism; copy-paste and light/moderate modifications (only in training dataset)
  - **Essay-2** – manually written essays with plagiarism; moderate/heavy modifications
  - **Generated texts** – texts with randomly generated plagiarism; copy-paste or moderate modifications
  - **Academic texts** – real world examples of plagiarism; mostly copy-paste

Performance Measures (Source retrieval)

- Recall – the fraction of sources that are retrieved
- Precision – the fraction of retrieved documents that are true sources
- Mean average precision (MAP) – the higher the more sources are in the top of the result

Evaluation of source retrieval algorithm

- Results on the test data for source retrieval

	Recall	Mean average precision	Precision
Academic	0.978	0.61	0.003
Essays-2	0.989	0.39	0.009

Performance Measures (Text alignment)

- Recall – the fraction of a source text that is detected
- Precision – the fraction of detected text that is plagiarised
- Granularity reflects the consistency of detected text (the less the better)
- Plagdet – the combination of previous three measures

Evaluation of text alignment

- Results on the test data for source retrieval

	Recall	Precision	Granularity	Plagdet
Essays-2	**0.531**	0.82	**1.0016**	**0.644**
Baseline: Essays-2	0.076	**0.896**	1.141	0.128
Generated paraphrasing	**0.865**	**0.981**	**1.483**	**0.7**
Baseline: generated paraphrasing	0.833	0.97	3.464	0.416
Generated copy/paste	0.859	0.978	1.466	0.702
Baseline: generated copy/paste	**0.994**	**0.961**	**1.004**	**0.9744**

Most difficult obfuscation types for our method

- Training data was annotated with the type of obfuscation
- Recall per type for Essays-1 collection

	Description	Recall
CCT	concatenation of sentences	0.41
HPR	paraphrasing	0.44
SSP	splitting of sentences	0.65
LPR	moderate modifications (replacing/reordering of words)	0.78
ADD	addition of words	0.85
DEL	deletion of words	0.85
CPY	copy/paste	0.87

Evaluation of the plagiarism detection method

- Results on training data

	Source Retrieval		Text Alignment		
	Rec.	MAP	Rec.	Prec.	Plagdet
Essays-1	0.97	0.754	0.783	0.904	0.839
Essays-2	0.82	0.709	0.316	0.883	0.466

- Results on test data

	Source Retrieval		Text Alignment		
	Rec.	MAP	Rec.	Prec.	Plagdet
Essays-2	0.83	0.608	0.382	0.885	0.533

Future work

- Estimate current impact of semantic/syntactic similarity measures on recall
- Explore more possibilities to leverage them for detecting heavily disguised plagiarism
- Address weak points of detection some obfuscations (concatenation)

Denis Zubarev - zubarev@isa.ru

Demo - like.exactus.ru