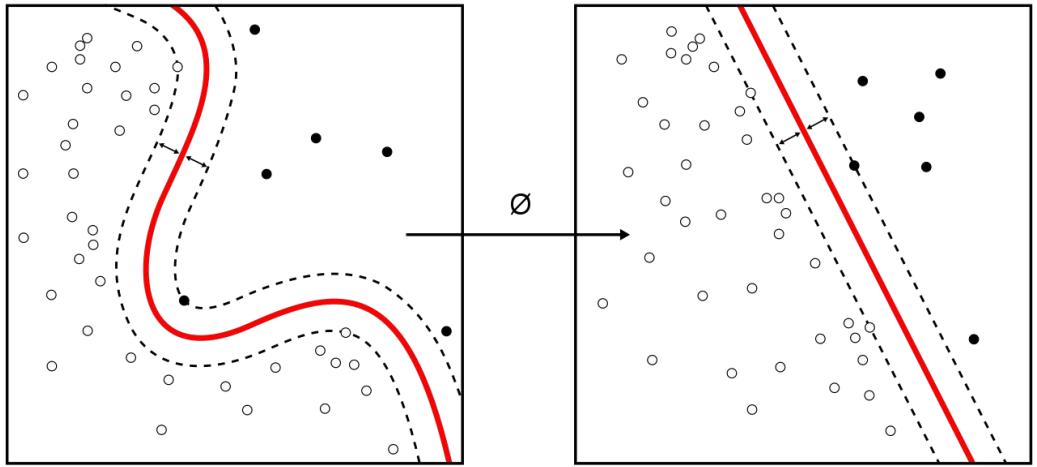
Poetry clustering algorithm based on lexical features and synonymy

Разработка алгоритма кластеризации поэтических текстов на основании лексических признаков с учетом синонимии

Student: Tagirova Elizaveta
Scientific advisor: Prof. Vladimir Borisovich Barahnin, D.Sc

Clustering

is a grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups.



Lexical features

is a set of words and phrases that are used in the text.

Synonymy

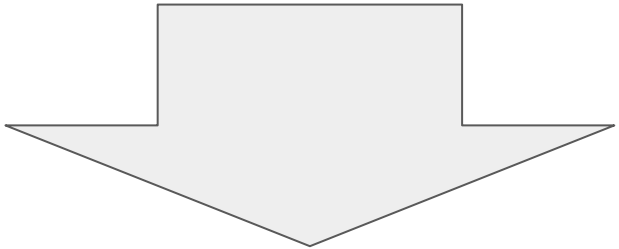
similarity of words in the meaning with the sound difference.

Lexical features

is a set of words and phrases that are used in the text.

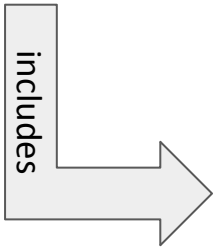
Synonymy

similarity of words in the meaning with the sound difference.



different lexical features can be close by the meaning

Complex analysis of poetic texts

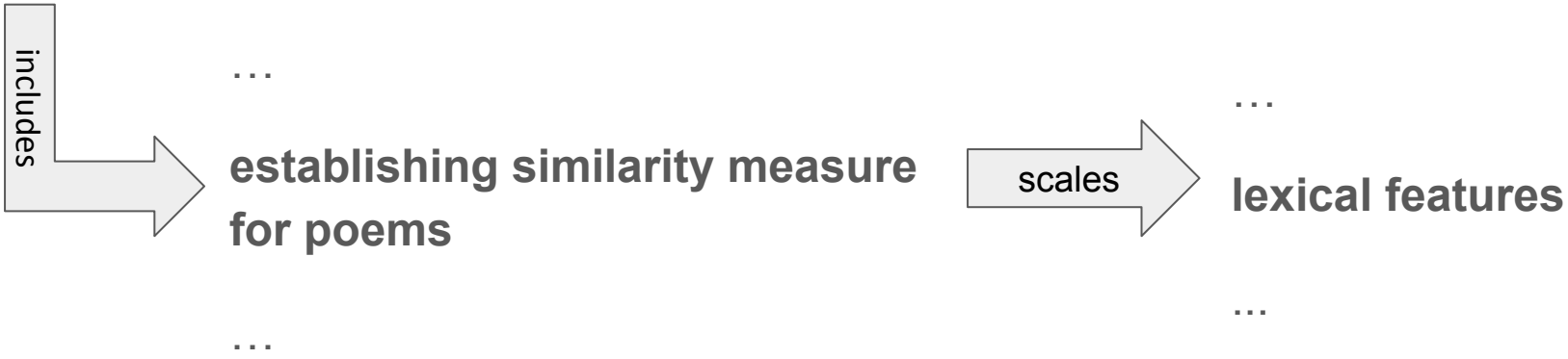


…

establishing similarity measure for poems

…

Complex analysis of poetic texts



Steps

1) extract lexical features - phrases
2) represent phrases in vector space
3) represent text as a vector of features
4) cluster texts

Step 1. Extracting phrases

1) build dependency tree using syntax parser

2) extract phrases from dependency tree [4]:
   a) phrase = subtree
   b) define stopwords → remove subtrees with stopword root
   c) consider not all dependencies

Step 1. Extracting phrases

- AOT
- Pullenti
- Text Chunking
- Tomita Parser
- MST Parser
- MaltParser
- DependencyParser (SpaCy)
- Syntaxnet
- UDPipe

Step 1. Extracting phrases

- AOT
- Pullenti
- Text Chunking
- Tomita Parser
- MST Parser
- MaltParser
- DependencyParser (SpaCy)
- Syntaxnet
- **UDPipe**

- Python lib
- linear complexity
- dependency trees
- not the best morph analyzer?

CoNLL 2018 Shared Task [1]

Step 1. Extracting phrases

- AOT
- Pullenti
- Text Chunking
- Tomita Parser
- MST Parser
- MaltParser
- DependencyParser (SpaCy)
- Syntaxnet
- **UDPipe**

- MorphoRuEval-2017 [2]

- Automatic morphological analysis for Russian: a comparative study [3]

Step 2. Phrases in vector space

Can sentence approaches be applied?

Step 2. Phrases in vector space

Can sentence approaches be applied?

Sentence clustering using continuous vector space representation [6]:

$$F(\mathbf{x}) = \sum_{w \in \mathbf{x}} f(w)$$

Step 2. Phrases in vector space

Can sentence approaches be applied?

Universal Sentence Encoder [5]:

- Transformer model
- Deep Averaging Network (DAN) model

Step 2. Phrases in vector space

Can sentence approaches be applied?

Transformer model

1) compute context aware representations of words (ordering and identity)

2) representations (1) are converted to a fixed length vector (sentence encoding): element-wise sum of the representations at each word position

Step 2. Phrases in vector space

Can sentence approaches be applied?

Deep Averaging Network (DAN) model

1) embeddings for words and bi-grams are averaged together

2) (1) passed through a feedforward deep neural network (DNN)

Step 2. Phrases in vector space

Can DisCoCat model be applied?

Step 2. Phrases in vector space

What about figurative meaning?

References

1. CoNLL 2018 Shared Task: https://universaldependencies.org/conll18/results.html
2. *Sorokin A, …* MorphoRuEval-2017: an Evaluation Track for the Automatic Morphological Analysis Methods for Russian
3. *O.V. Dereza, D.A. Kayutenko, A.S. Fenogenova*. Automatic morphological analysis for Russian: a comparative study
4. *В.И. Новицкий.* Подход к автоматическому поиску переводных словосочетаний на основе синтаксической информации и многоуровневой фильтрации

References

5. *Daniel Cer, Yinfei Yang,…* Universal Sentence Encoder (arXiv:1803.11175)
6. *Mara Chinea-Rios,…* Sentence clustering using continuous vector space representation

Thank you for attention!