# A SYNTAX-BASED DISTRIBUTIONAL MODEL FOR DISCRIMINATING BETWEEN SEMANTIC SIMILARITY AND ASSOCIATION

Trofimov I.V., Suleymanova E.A.
Program Systems Institute of RAS

# Terminology. **Relatedness**



**"similarity"**

- feature-based taxonomic relatedness

**"association"**

- thematic relatedness

# Related



**similar**

- sharing intrinsic features that account for membership in the same semantic category

**associated**

- frequently occurring together in space and language

# Related



## similar

- ***car*** and ***bike***
  - common physical features (wheels)
  - common function (transport)
  - fall within a clearly definable category (modes of transport)

## associated

- ***bee*** and ***honey***

## similar
## associated

- ***king*** and ***queen***

# State of the art



- With few exceptions, recent research in distributional semantics has focused on **quantitative** rather than **qualitative** aspects of word interaction within lexical semantic system.
- Such approaches **neglect the difference** between similarity and association: their focus is estimating the strength of the connection between two words in the semantic network, regardless of the relation type.

# Task



- Develop a distributional model aimed at recognizing semantic similarity—relations that are based on shared intrinsic features and common category membership

# Task



- Pairs of **similar** (and possibly associated) nouns

  should get higher scores than

- pairs of **pure associations** (relations that are based on thematic, or situational, co-occurrence and are not supported by taxonomical commonality)

# RuSim1000 dataset



- 1000 pairs of **related** nouns that are divided into two subsets
  - Positive examples are pairs of similar (and possibly associated) nouns
  - Negative examples are pairs of associated, but not similar nouns

# RuSim1000 dataset



- RuSim1000 was designed in such a way that it would be compatible with the RUSSE evaluation framework
  - Average Precision (AP) as evaluation measure

# RuSim1000 dataset. Positive subset



- Core of the positive subset:
  - synonyms (*имя-название*, *name-title*)
  - hyponym-hypernym (*питон-змея*, *python-snake*)
  - co-hyponyms (*писатель-поэт*, *writer-poet*).

# RuSim1000 dataset. Negative subset



- Core of the negative subset—pairs of nouns representing ontologically different entities:
  - part-whole (*шерсть-животное*, *fur-animal*)
  - element-set (*самолет-эскадрилья*, *airplane-squadron*)
  - functional (situational) relationship (*доктор-клиника*, *doctor-clinic*, *винтовка-выстрел*, *rifle-shot*)

# RuSim1000 dataset. Difficult and borderline cases



- **Antonyms**

  are taken to be similar (i.e. positive examples)

  - Assumption: their opposition holds within a certain category to which they both belong (*свет-тьма, light-darkness*)

# RuSim1000 dataset. Difficult and borderline cases



- **Roles**

  It was decided to qualify as positive (i.e. similar):

  - pairs of the kind “a type and its typical role” (*торф-топливо, peat-fuel*, but not *самолет-вооружение, airplane-armament*)
  - thematically related roles of the same holder type, including complementary roles (*врач-медсестра, doctor-nurse, врач-пациент, doctor-patient*)

# Dataset RuSim1000



| word 1 | word 2 | sim |
|---|---|---|
| лошадь (horse) | жеребец (stallion) | 1 |
| лошадь (horse) | кобыла (mare) | 1 |
| лошадь (horse) | пони (pony) | 1 |
| лошадь (horse) | кляча (jade) | 1 |
| лошадь (horse) | седло (saddle) | 0 |
| лошадь (horse) | конюх (groom) | 0 |
| лошадь (horse) | грива (mane) | 0 |
| лошадь (horse) | галоп (gallop) | 0 |

# Model



- similar objects tend to have more shared features than dissimilar
- similar objects tend to act in similar way
- similar objects tend to be exposed to similar actions

# Model



- The context vector is composed of
  - adjectives, for feature-based similarity measure
  - verbs—for behavioral similarity
- The length of vectors is not limited
- Positive pointwise mutual information (PPMI)
- Cosine similarity for measuring the distance between vectors

# Experiments and results



- Source of statistical data—RuWac corpus
- Evaluation on RuSim1000 (Average Precision)

| syntactic relation | | |
|---|---|---|
| **attributive** | **predicative** | **1-completive** |
| 0.907 | 0.846 | 0.882 |

| combination of syntactic relations | |
|---|---|
| **attributive + predicative** | **attributive + 1-completive** |
| 0.918 | 0.925 |

Thank you!