Testing Features and Measures in Russian Paraphrasing Task

Natalia Loukachevitch (louk_nat@mail.ru),
Alexander Shevelev,
Valeria Mozharova
Lomonosov Moscow State University

Paraphrase detection

- A *paraphrase* is a restatement of the meaning of a text, passage or sentence using other words.
- Detection of paraphrases is important for
  - Information retrieval
  - Question answering
  - Text summarization
  - Document clustering
  - Plagiarism detection etc.

- Most research for English
- Other languages including Russian:
  - Much less research

Features proposed in previous work

- various measures of word and character similarities
  - length features, longest common sequence, n-gram overlap features, edit distances, machine translation similarities (BLUE, WER, TER, ROUGE-L etc.), information-retrieval measures (tf-idf, BM25), named entity similarity (Brychcın, Svoboda 2016);
- features of lexical differences between sentences
  - including parts of speech tags, named entities, meaningful words (Pronoza, Yagunova, 2015a);
- syntactic features based on similarity between dependency trees;
- semantic measures
  - based on WordNet conceptual structure (Mihalcea et al. 2006; Fernando, Stevenson, 2008);
- corpus-based similarities
  - using classical distributional vectors or distributed representations of words learned by neural networks on a large text corpus (Przybyla et al., 2016);
- last approaches (SemEval-2016):
  - combine neural networks, comparison of dependency trees and semantic measures based on WordNet similarity (Rychalska et al., 2016; Brychcın, Svoboda 2016).

Shared Task on Russian Paraphrase Detection (Pivovarova et al., 2016)

- Precise, loose and non-paraphrases
  - Sentences were extracted from news headlines
- Classifications tasks:
  - Binary (paraphrases vs. non-paraphrases) and three-class
- Collections
  - Train collection: about 7000 pairs
  - Test collection: 1924 pairs
- Type of runs
  - Standard: train data and manual resources
  - Non-standard: all types of resources

Examples from the Dataset

- Precise Paraphrase
  - *У Деми Мур украли одежду. (Demi Moor's clothes were stolen)*
  - *У Деми Мур похитили одежду. (Demi Moor's clothes were robbed)*

- Loose Paraphrase
  - *Названа причина смерти Уго Чавеса (The cause of Hugo Chavez's death is named*).
  - *Причиной смерти Чавеса назвали инфаркт (The cause of Chavez's death was a heart attack.*)

This Work: Features for Paraphrase Detection in Russian

- Semantic Similarity Features
  - Based on published version of RuThes thesaurus
  - http://www.labinform.ru/pub/ruthes/index.htm

- Combination of thesarus features with other features:
  - String-based Features
  - Information-retrieval features
  - Part-of-Speech Features

RuThes Linguistic Ontology

- Unified representation – single net of concepts
  - In WordNet there are nets of synsets divided into parts of speech
- Text entries of the same concept can include
  - Different parts of speech
  - (cf. WordNet: synsets contain only the same POS words)
  - Lexical units and domain terms
  - Words and multiword expressions
- RuThes-lite – published version
  - 115 thousand words and expressions

RuThes Relations

- Small set of relations
  - Class – subclass
    - Transitivity, inheritance
  - Part-whole
    - Transitivity of part-whole relations
  - External ontological dependence (Gangemi et al., 2001; Guarino, 2009)
    - Existence of *Car plant* depends on existence of *car*
    - *Inherited to sublasses and parts*
- Semantic similarity is usually calculated using the thesaurus paths
  - In RuThes paths are defined on the basis of relations' properties

САД - Mozilla Firefox

Главная О проекте Справка

А Б В Г Д Е Ж З И Й К Л М Н О П Р С Т У Ф Х Ц Ч Ш Щ Ы Э Ю Я

С Б... САМ... САН... САТ... СБЫ... СВИ... СГИ... СДИ...
СЕК... СЕМ... СЕС... СИМ... СИТ... СКЛ... СКУ... СЛИ...
СЛЫ... СМО... СНИ... СОВ... СОД... СОК... СОН... СОС...
СОУ... СОЧ... СПИ... СПР... СРИ... СТВ... СТЛ... СТР...
СТУ... СУЕ... СУТ... СХР... СЫК...

Текстовый вход: *САД*

ДЕТСКИЙ САД

(*ДЕТСАД*, *ДЕТСАДИК*, *ДЕТСАДОВСКИЙ*, *ДЕТСКИЙ САД*, *САД*, *САДИК*, *САДОВСКИЙ*, *САД-ЯСЛИ*, *ЯСЛИ-САД*)

ВЫШЕ ДОШКОЛЬНОЕ УЧРЕЖДЕНИЕ
ЧАСТЬ ЯСЛИ

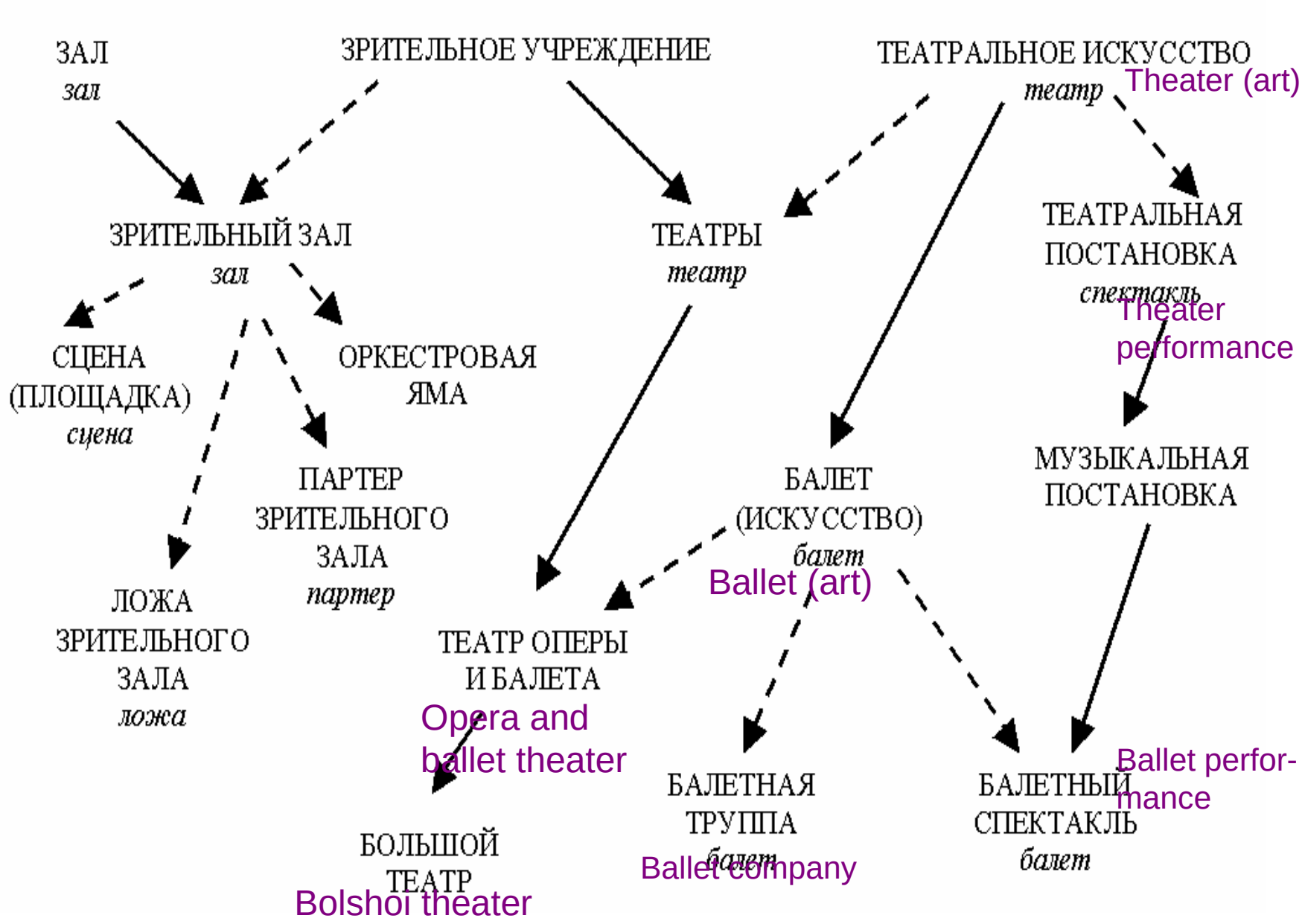
САД (УЧАСТОК ЗЕМЛИ)

(*САД*, *САДИК*, *САДОВЫЙ*)

ВЫШЕ ЗЕМЕЛЬНЫЙ УЧАСТОК
АССОЦ$_1$ САДОВАЯ КУЛЬТУРА
АССОЦ$_2$ БЕСЕДКА
АССОЦ$_2$ САДОВНИК
АССОЦ$_2$ САДОВОДСТВО

http://www.labinform.ru/pub/ruthes/index.htm

RuThes Fragment



Thesaurus-based Semantic Similarity Measures

- Well-known for WordNet
- We study:
  - Semantic measures for RuThes
  - Measures based on different types of concept paths
    - Only hypernyms
    - Hypernyms and wholes
    - All relations
  - Paths without length restriction vs. with additional restriction on the path length

Thesaurus Features: Leacock-Chodorow measure and its linear variant

$$sim_{lch} = -\log_{2D} \frac{N_p}{2D} = 1 - \log_{2D} N_p$$

- where *Np* is the distance between nodes
- *D* is the maximum depth in the taxonomy
- the distance between synonyms is equal 1

$$sim_{path} = 1 - \frac{N_p}{2D}$$

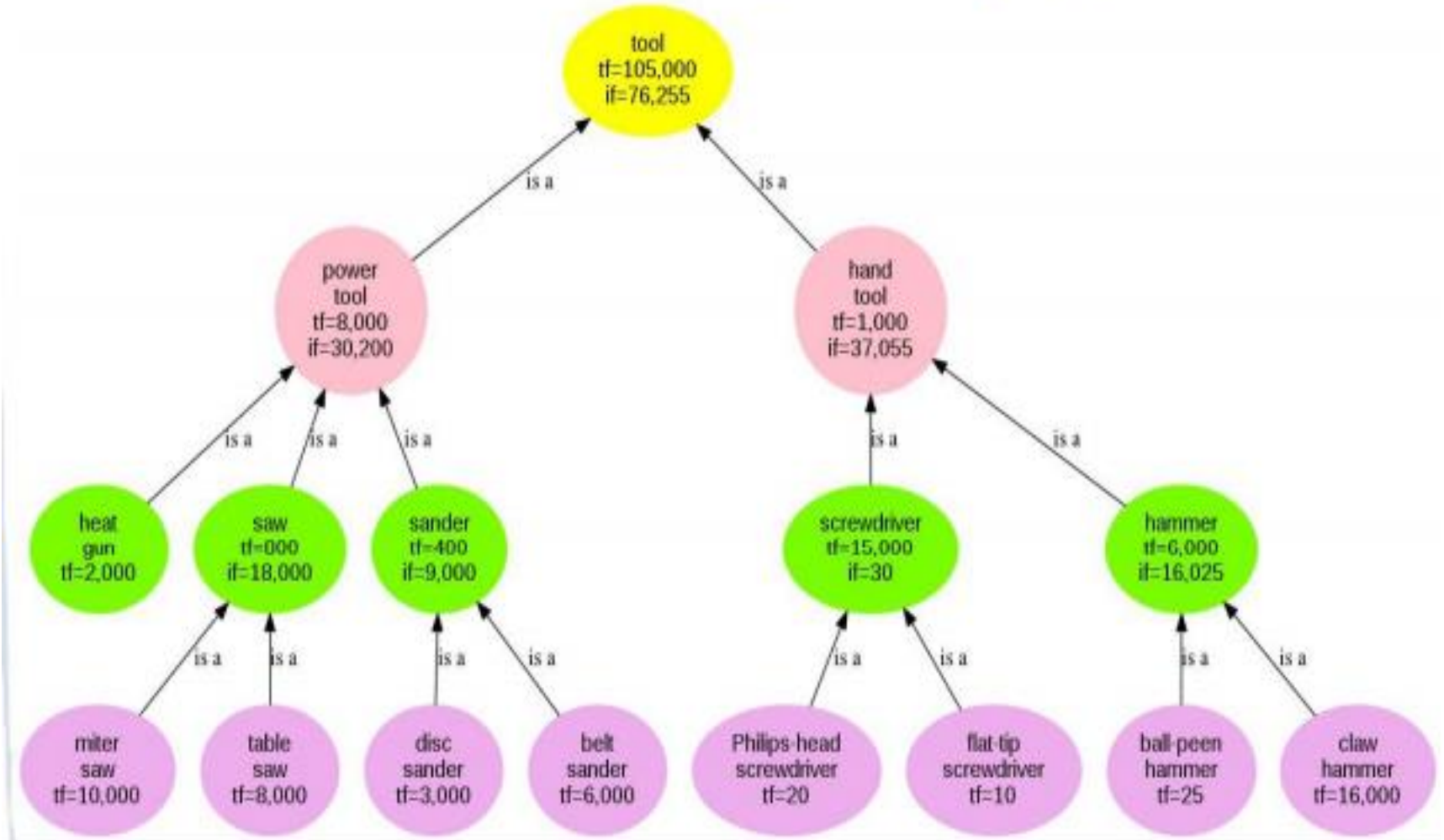
Information Content (IC)

- IC (concept)=-log(p(concept)) (Resnik, 1995)
- Counting IC
  - Term frequency + Inherited frequency
  - Inherited frequency=frequency of lower level concepts
- Low frequency concepts are often more specific than high frequent ones
  - IC – large positive value,
  - The more frequency of a concept is, the less IC is.

- We used news corpus, more then 1 million news articles

Information Content inherited frequency (if)

(Pedersen, 2013)



Measures based on information content

- Lin measure

$$sim_{lin} = \frac{2 \cdot IC(LCS(C_1, C_2))}{IC(C_1) + IC(C_2)}$$

- Jcn measure

$$sim_{jcn} = \frac{1}{IC(C_1) + IC(C_2) - 2 \cdot IC(LCS(C_1, C_2))}$$

- LCS – least commom subsumer
- The smallest path is considered (for ambiguous words)

Calculating similarity measure between sentences

- Similarity matrix is calcualted between words of two sentences Fernando, Stevenson, 2008)

$$sim\ (\vec{a}, \vec{b}) = \frac{\vec{a} W \vec{b}}{|\ \vec{a}\ ||\ \vec{b}\ |}$$

- If a word in the fist sentence is similar to several words in other sentences, this similarities are summed up
- In our work: word similarity – not more than 1

- One-feature classifier (linear SVM) was trained
- It allows finging optimal thresholds between classes

Example of similarity matrix (Lch measure)

- (s1) *У Деми Мур украли одежду. (Demi Moor's clothes were stolen)*
- *(s2) У Деми Мур похитили одежду. (Demi Moor's clothes were robbed)*

	Деми (Demi)	***Мур (Moor)***	***Украсть (steal)***	***Похитить (rob)***	***Одежда (Clothes)***
Деми (Demi)	1	0	0	0	0
Мур (Moor)	0	1	0	0	0
Украсть (steal)	0	0	1	0.7941	0
Похитить (rob)	0	0	0.7941	1	0
Одежда (Clothes)	0	0	0	0	1

Finding the Best Thesaurus Feature (F-measure)

Feat.	Relations	2-class Best Results/Full	3-class Best Results/Full
Lch	Only Hypernyms Hypernyms and Wholes All relations	78.4 (6)/ 78.8 (5) 78.9 (5)	54.1 (3) 54.5 (5) 54.9 (5)
Path	Only Hypernyms Hypernyms and Wholes All relations	78.4 (3) 78.8 (4) 78.8 (5)	54.2 (5) 54.3 (4) 54.2 (2)
Lin	Only Hypernyms Hypernyms and Wholes All Relations	**79.5 (2)/**74.7 **79.4 (2)/**74.9 **79.9 (2)/**75.0	54.5 (2)/35.8 55.5 (2)/34.5 55.1 (2)/34.6
Jcn	Only Hypernyms Hypernyms and Wholes All relations	**79.6 (3)/**79.09 **79.5 (2)/**78.7 **79.6 (2)/**78.7	**56.2 (2)/**55.4 **56.0 (3)/**54.0 **56.4 (3)/**54.2

Combining with Other Features

- String Features in form of intersections
  - 2- and 3-symbol Ngrams, 1-3 word Ngrams

$$feature_1 = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|} \qquad feature_2 = \frac{|S_1 \cap S_2|}{|S_1|} \qquad feature_3 = \frac{|S_1 \cap S_2|}{|S_2|}$$

- Information-Retrieval features
  - BM25
  - Idf of words in difference set between sentences

- POS features of words in difference set between sentences

Results of machine learning
(Random Forest classifier, grid parameter tuning)

Feature Set	**2-class task Acc/F1**	**3-class task Acc/F1**
Best single thesaurus feature	- / 79.9	- /56.4
1) String-based combination	73.80/79.00	60.03/57.90
2) 1)+BM-25	74.06/79.18	60.96/58.99
3) 2)+5-POS Features	74.42/79.32	61.07/59.03
3)+Best Thesaurus= 2 from lch (only hyper, hyper+whole)	**77.33/81.71**	**62.57/60.93**
Best res. of Shared Task Standard Non-Standard	 74.59/80.14 77.39/81.10	 59.01/56.92 61.81/58.38

Experiments with other machine learning methods (three class task): scikit-learn

Method	**Default values**	**Grid tuning**
Linear SVM	**61.43/**58.1	61.64/58.52
SVM with rbf kernel	60.49/57.62	59.61/57.32
Random forest	56.65/54.6	**62.57/60.93**
Gradient boosting	60.86/**59.11**	61.93/59.92

Conclusion

- We studies Ruthes similarity measures for Russian paraphrase task
- Semantic features
  - Proposed for WordNet
  - Use of all relations are usually slightly better than to utilize only hypernyms relations
  - Restriction of length path improves the measures significantly
  - The best thesaurus features as addition to other features were two features lcg (without accounting IC)
- The best method: random forest
  - scikit-learn with grid tuning