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 (Brychcin, 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; Brychcin, 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

• Combination of thesaurus 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 subclasses and parts
• Semantic similarity is usually calculated using the thesaurus paths
  – In RuThes paths are defined on the basis of relations’ properties
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
АССОЦ₁ САДОВАЯ КУЛЬТУРА
АССОЦ₂ БЕСЕДКА
АССОЦ₂ САДОВНИК
АССОЦ₂ САДОВОДСТВО

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

\[ \text{sim}_{lch} = -\log_{2D} \frac{N_p}{2D} = 1 - \log_{2D} N_p \]

- where \( N_p \) is the distance between nodes
- \( D \) is the maximum depth in the taxonomy
- the distance between synonyms is equal 1

\[ \text{sim}_{path} = 1 - \frac{N_p}{2D} \]
Information Content (IC)

- IC (concept) = -\log(p(\text{concept})) \quad \text{(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 than 1 million news articles
Information Content
inherited frequency (if)

(tool
tf=105,000
if=76,255)

(power tool
tf=8,000
if=30,200)

(hand tool
tf=1,000
if=37,055)

(heat gun
tf=2,000)

(saw
tf=0
if=18,000)

(sander
tf=400
if=9,000)

(screwdriver
tf=15,000
if=30)

(hammer
tf=6,000
if=16,025)

(miter saw
tf=10,000)

(table saw
tf=8,000)

(disc sander
tf=3,000)

(belt sander
tf=6,000)

(Phillips-head screwdriver
tf=20)

(Flat-tip screwdriver
tf=10)

(ball peen hammer
tf=25)

(claw hammer
tf=16,000)

(Pedersen, 2013)
Measures based on information content

• Lin measure

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

• Jcn measure

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

• LCS – least common subsumer
• The smallest path is considered (for ambiguous words)
Calculating similarity measure between sentences

• Similarity matrix is calculated between words of two sentences Fernando, Stevenson, 2008)

\[
\text{sim} (\vec{a}, \vec{b}) = \frac{\vec{a}W\vec{b}}{||\vec{a}|| ||\vec{b}||}
\]

• If a word in the first 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 finding 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)

<table>
<thead>
<tr>
<th></th>
<th>Деми (Demi)</th>
<th>Мур (Moor)</th>
<th>Украдъ (steal)</th>
<th>Похитили (rob)</th>
<th>Одежда (Clothes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Деми (Demi)</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Мур (Moor)</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Украдъ (steal)</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.7941</td>
<td>0</td>
</tr>
<tr>
<td>Похитили (rob)</td>
<td>0</td>
<td>0</td>
<td>0.7941</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Одежда (Clothes)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
# Finding the Best Thesaurus Feature (F-measure)

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

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

\[
\text{feature}_1 = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|} \quad \text{feature}_2 = \frac{|S_1 \cap S_2|}{|S_1|} \quad \text{feature}_3 = \frac{|S_1 \cap S_2|}{|S_2|}
\]

- 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)

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>2-class task Acc/F1</th>
<th>3-class task Acc/F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best single thesaurus feature</td>
<td>- / 79.9</td>
<td>- / 56.4</td>
</tr>
<tr>
<td>1) String-based combination</td>
<td>73.80/79.00</td>
<td>60.03/57.90</td>
</tr>
<tr>
<td>2) 1)+BM-25</td>
<td>74.06/79.18</td>
<td>60.96/58.99</td>
</tr>
<tr>
<td>3) 2)+5-POS Features</td>
<td>74.42/79.32</td>
<td>61.07/59.03</td>
</tr>
<tr>
<td>3)+Best Thesaurus= 2 from lch (only hyper, hyper+whole)</td>
<td><strong>77.33/81.71</strong></td>
<td><strong>62.57/60.93</strong></td>
</tr>
<tr>
<td>Best res. of Shared Task Standard</td>
<td>74.59/80.14</td>
<td>59.01/56.92</td>
</tr>
<tr>
<td>Best res. of Shared Task Non-Standard</td>
<td>77.39/81.10</td>
<td>61.81/58.38</td>
</tr>
</tbody>
</table>
Experiments with other machine learning methods (three class task): scikit-learn

<table>
<thead>
<tr>
<th>Method</th>
<th>Default values</th>
<th>Grid tuning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear SVM</td>
<td>61.43/58.1</td>
<td>61.64/58.52</td>
</tr>
<tr>
<td>SVM with rbf kernel</td>
<td>60.49/57.62</td>
<td>59.61/57.32</td>
</tr>
<tr>
<td>Random forest</td>
<td>56.65/54.6</td>
<td>62.57/60.93</td>
</tr>
<tr>
<td>Gradient boosting</td>
<td>60.86/59.11</td>
<td>61.93/59.92</td>
</tr>
</tbody>
</table>
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