Study of methods for automatic taxonomy enrichment for the Russian language

Coursework

by Daria Pirozhkova

Advisors: Tatiana Batura, Ivan Bondarenko

NSU,

May 2020

Content

- Project goal and objective.
- Papers review.
- Data description.
- Data processing.
- References.

Project goal and objective

The main goal is to create methods that automatically enrich ruWordNet with new terms, and at the same time connect them with existing words using hypernym relations.

Hyponym	Hyponym id	Hypernym
cat	N-2584	animal
	N-2859	mammal
	..	...
duck	N-7245	waterfowl
	N-9743	bird
	..	...

The main task is to predict a ranked list of ten terms that are most likely to be hypernyms for a word not included in the thesaurus.

Papers review

- Different solutions to this task for the English language includes approaches based on the vectorization, classification, and clustering of the words that are hypernym relations.
- The less articles describe the solution with neural networks. But the values of precision and recall of the proposed approach are higher that values of the other methods.
- The results of the task "Taxonomy Extraction for selected four target domains" proved the solution for domain-specific data is better work that the one for language in general.

Papers review

- The same task for different languages indicated the translation problem. This leads to the importance of solving the problem specifically for the Russian language.
- The solution for Russian language [Karaeva et al. 2018] includes approach using the word embedding and calculating the distance between them.The obtained precision value is less than 65 percent.

Data description

Senses

```
<senses>
  <sense id="147272-N-712535"
   synset_id="147272-N" name="КРЕСТНЫЙ РОДИТЕЛЬ"
   lemma="КРЕСТНЫЙ РОДИТЕЛЬ"
   main_word="РОДИТЕЛЬ" synt_type="NG"
   poses="Adj N"
   meaning="1" part_of_speech="N"
   concept_id="147272" entry_id="712535"/>
 <sense id="7331-N-211359"
   synset_id="7331-N" name="ЗЛОУПОТРЕБЛЕНИЯ В ТОРГОВЛЕ"
   lemma="ЗЛОУПОТРЕБЛЕНИЕ В ТОРГОВЛЯ"
   main_word="ЗЛОУПОТРЕБЛЕНИЕ"
   synt_type="NG"
   poses="N Prep N"
   meaning="1"
   part_of_speech="N"
   concept_id="7331"
   entry_id="211359"/>
```

Synset_relations

```
<relation parent_id="147272-N" child_id="147272-A" name="POS-synonymy"/>
<relation parent_id="147272-N" child_id="4544-N" name="hypernym"/>
<relation parent_id="147272-N" child_id="126551-N" name="hyponym"/>
<relation parent_id="147272-N" child_id="2201-N" name="domain"/>
```

Synsets

```
<synsets>
<synset id="147272-N" ruthes_name="КРЕСТНЫЙ РОДИТЕЛЬ" definition="" part_of_speech="N">
<sense id="147272-N-712535">КРЕСТНЫЙ РОДИТЕЛЬ</sense>  </synset>
  <synset id="7331-N" ruthes_name="НАРУШЕНИЕ ПРАВИЛ ТОРГОВЛИ" definition="" part_of_speech="N">
<sense id="7331-N-211359">ЗЛОУПОТРЕБЛЕНИЕ В ТОРГОВЛЯ</sense>
<sense id="7331-N-213260">НАРУШЕНИЕ ПРАВИЛО ПРОДАЖА</sense>
<sense id="7331-N-107257">НАРУШЕНИЕ ПРАВИЛО ТОРГОВЛЯ</sense>
<sense id="7331-N-677543">НЕСОБЛЮДЕНИЕ ПРАВИЛО ТОРГОВЛЯ</sense>  </synset>
```

Derived_from

```
<sense name="НАРУШЕНИЕ" id="122242-N-143174" synset_id="122242-N"> <derived_from>
<sense name="НАРУШИТЕЛЬ" id="123618-N-166886" synset_id="123618-N"/>
<sense name="НАРУШИТЬ" id="122242-V-143176" synset_id="122242-V"/>
<sense name="НАРУШАТЬ" id="122242-V-143172" synset_id="122242-V"/>
<sense name="НАРУШИТЕЛЬНИЦА" id="123618-N-166887" synset_id="123618-N"/> </derived_from>
```

Composed_of

```
<sense name="НАРУШЕНИЕ ПРАВИЛ ТОРГОВЛИ" id="7331-N-107257" synset_id="7331-N"> <composed_of>
<sense name="НАРУШЕНИЕ" id="122242-N-143174" synset_id="122242-N"/>
<sense name="ТОРГОВЛЯ" id="1026-N-112958" synset_id="1026-N"/>    </composed_of>  </sense>
```

Data processing

Training data set 1

Couple of words are related by different relations (such as antonyms, synonyms, hyponym-hypernym) or are not related any relations. The classifier showing these relations are hypernym or not: 1 - hypernym, 0 - others.

child id	parent id	label
cat	animal	1
duck	tree	0

Data processing

Training data set 2

Data set consists of text pairs, which are a context for a hyponym. First text contains hyponym. Second text contains hypernym that replaces hyponym in corresponding form.

References

1. Gabriel Bernier-Colborne and Caroline Barriere. 2018. CRIM at SemEval-2018 Task 9: A Hybrid Approach to Hypernym Discovery.
2. William Held and Nizar Habash. 2019. The Effectiveness of Simple Hybrid Systems for Hypernym Discovery.
3. Mengyi Zhang, Tianxing Wu, Qiu Ji, Guilin Qi, Zhixin Sun. 2019. Mining Hypernym-Hyponym Relations from Social Tags via Tag Embedding.
4. Yifang Sun, Shifeng Liu, Yufei Wang, Wei Wang. 2019. Extracting Definitions and Hypernyms with a Two-Phase Framework.
5. aria Karaeva, Pavel Braslavski, Yury Kiselev. 2018. Extraction of Hypernyms from Dictionaries with a Little Help from Word Embeddings.
6. Georgeta Bordea, Paul Buitelaar, Stefano Faralli, Roberto Navigli. 2015. SemEval-2015 Task 17: Taxonomy Extraction Evaluation (TExEval).
7. Georgeta Bordea, Els Lefever, Paul Buitelaar. 2016. SemEval-2016 Task 13: Taxonomy Extraction Evaluation (TExEval-2).
8. David Jurgens, Mohammad Taher Pilehvar. 2016. SemEval-2016 Task 14: Semantic Taxonomy Enrichment.

Thank you for the attention.