Automated Thesaurus Enrichment For The Russian Language Using Self-Supervised Deep Learning Approach

Alexander Donets, Ivan Bondarenko, Tatyana Batura

Novosibirsk State University, December 2020

Contents

Introduction

- What is thesaurus?
- Where is can be used?
- The three most common-relations:
  hyponym-hypernym, holonym-meronym, concept-object
- The need for automation

Introduction

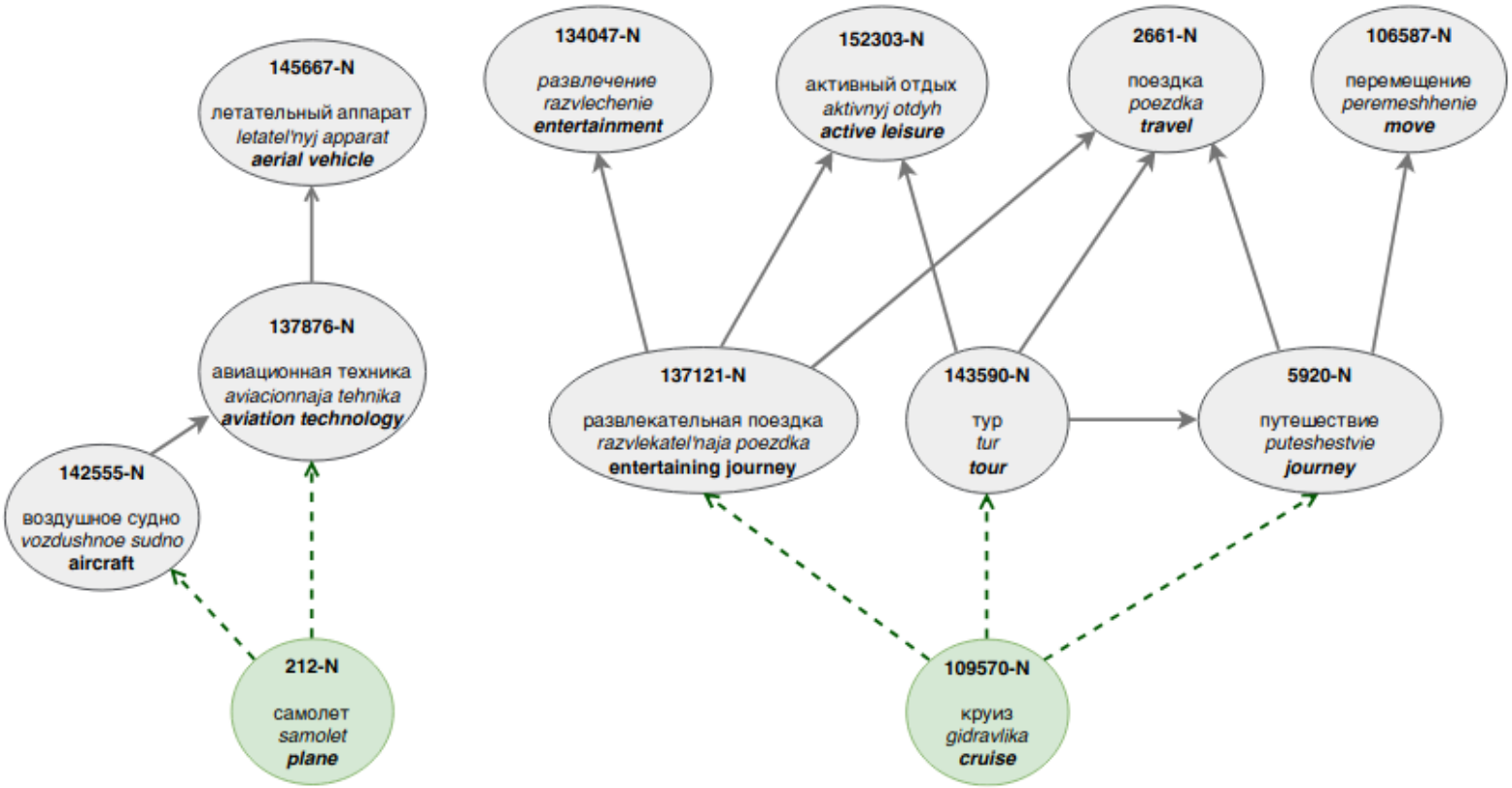


Figure 1: Examples of hypernym subgraphs from RuWordNet ground truth: direct and second-order hypernyms may be related in various ways motivating the evaluation metric based on connectivity components. While in (a) two parents lead to different senses, in (b, c, d) two parents lead to the same sense. Dashed lines indicate ground truth hypernyms.

Task

Task: given the thesaurus $\mathbb{T}$ and the hyponym $x \notin \mathbb{T}$, create a ranked list of best hypernym candidates
$h_1(x),\ h_2(x),\ ...,\ h_n(x) \in \mathbb{T}$.

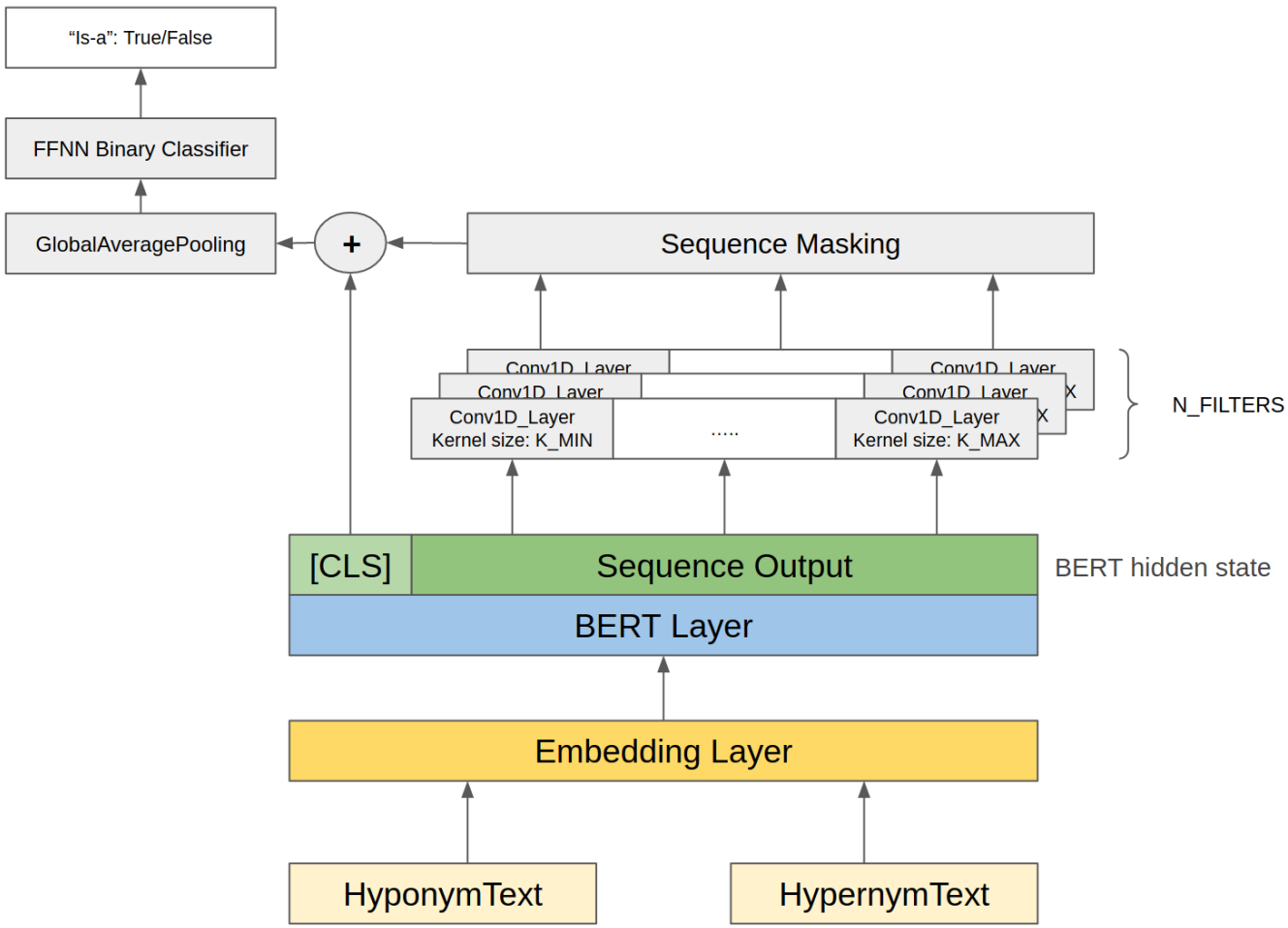
So we have the thesaurus, and an input hyponym. But what to with *homonymy*? A context needed.

Our Approach

Key points:

- Deep contextual embeddings from BERT
- Context mining (to overcome homonymy)
- Same text, different meaning

Architecture Overview



Datasets

1. RuWordNet thesaurus (total of 111500 utterances and words)
2. Train set provided by organizers of Dialogue-20 (12393 nouns, 2102 verbs)
3. Wikipedia dump (RU), 2017 (1.5GB of texts, 3 millions of tokens)
4. Next step: CommonCrawl (540TB of texts)

Evaluation Procedure

Metrics used:

- Mean Average Precision (main)
- Mean Reciprocal Rank (auxiliary)

As we have seen on the picture of hyponym-hypernym relations the hierarchy may be complicated. This leads to ambiguity during evaluation on which from TOP-N hypernym candidates to count as valid (when comparing them to TOP-N from thesaurus).

A curious note: if thesauri eventually will start to be enriched automatically, inevitably the amount of errors in them will be increased. This will affect our ground-truth thesaurus data.

Evaluation Procedure

"To avoid this hypernym ambiguity, we split all hypernyms of a word (both immediate and second-order) into groups. Each group corresponds to the connectivity component in the subgraph reconstructed from all hypernyms" (RUSSE'2020).

Baseline-1

RUSSE'2020 (no contextual transformer-like embdeddings):

For each new word — an orphan the participants should provide a ranked list of possible hypernyms (top-10).

1. Compute embeddings of all synsets in RuWordNet by averaging embeddings of all words from senses belonging to a synset.
2. Get embeddings for orphans. For multi-word orphans the embeddings are computed by averaging vectors for all words from which the orphan consist of.

Baseline-2

3. For each orphan compute the top K=10 closest synsets of the same part of speech as the orphan using the cosine similarity measure. An
4. Extract hypernyms for each of these closest synsets from the previous step. Take the first n=10 results (as each synset may have several hypernyms).

Baseline results: 0.42 (MAP), 0.4518 (MRR) (6-th place currently).

Work Plan

- Paper search and study (done)
- Data preparation (done)
- Implement algorithm for context search in Wikipedia (done)
- Implement simple BERT classifier (no CNNs) (finalizing)
- Evaluate simple BERT classifier (Dec 2020 - Jan 2021)
- Refactor code (Jan 2021)
- Implement and evaluate BERT-CNN classifier (Jan 2021)
- Summarizing results of experiments (Jan-Feb 2021)
- Preparing paper for publication (Feb-Mar 2021)

Thanks for your attention!