Joining Dictionaries and Word Embeddings for Ontology Induction

Dmitry Ustalov

Krasovskii Institute of Mathematics and Mechanics

December 1, 2016

Introduction

Definition

*A **lexical ontology** (or a **thesaurus**) is a lexical database that groups the words into the sets of synonyms called synsets or concepts, and records a number of semantic relations between these concepts.*

Thesauri are widely used for addressing different NLP problems:

- word sense disambiguation;
- document classification;
- dialogue systems, etc.

Prominent thesauri: WordNet, BabelNet, RussNet, RuThes.

The Problem

Currently, there is no WordNet-like thesaurus for Russian being available under a libré license.

Goals and Objectives

The present study has been conducted within the **Yet Another RussNet** project.

The Goal

To develop means for ontology induction from unstructured data using both automatic methods and crowdsourcing.

Objectives:

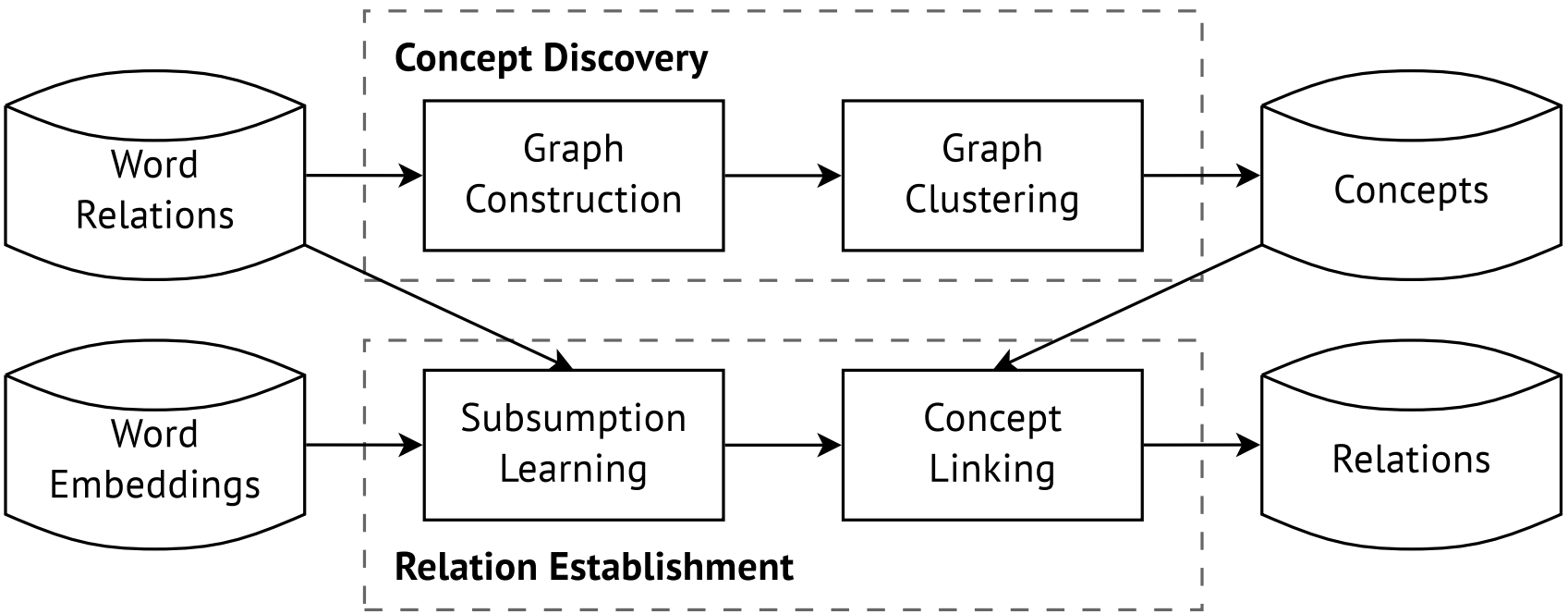
- to discover the concepts (also called the synsets);
- to establish relations between them;
- to evaluate them.

Related work: http://www.isprasopen.ru/files/conference.pdf, pp. 381–382.

The Approach

Principles

- Re-using the existing resources.
- Minimal efforts from the humans.
- Focusing on nouns, *is-a* relations, and domain ontologies.



Concept Discovery

Openly available synonym dictionaries:

- the Russian Wiktionary ($84\,625$ pairs);
- the Abramov's dictionary ($501\,612$ pairs);
- the Universal Dictionary of Concepts ($21\,657$ pairs).

Constructing an undirected graph $G = (V, E)$, where

- V is the set of the words;
- $(v, u) \in E \iff$ the words $v \in V$ and $u \in V$ are synonyms.

Assumption: cliques in G form the synsets.

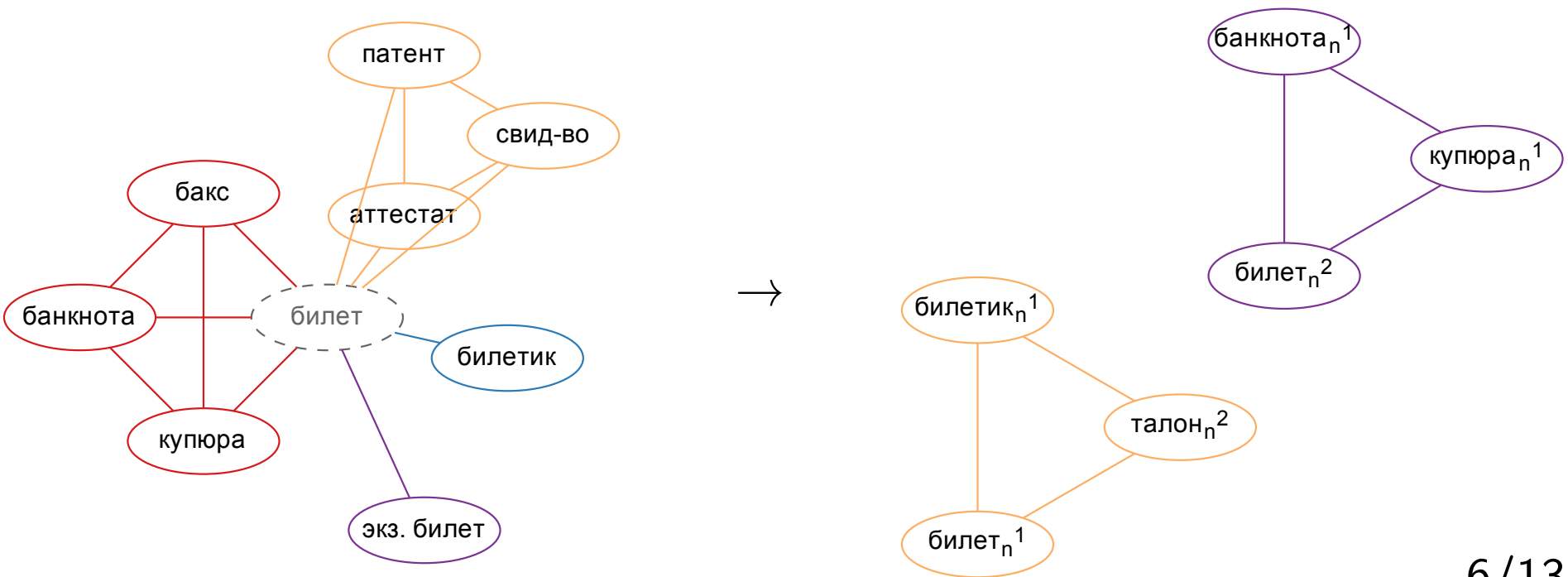
Challenges

- The clique problem is NP-complete.
- The phenomenon of polysemy.

Graph Clustering

Concept Discovery

1. Construct an ego-network $Ego(v)$ for $v \in V$ and exclude v.
2. Cluster $Ego(v)$ using Chinese Whispers.
3. Reconstruct and disambiguate the global graph G.
4. Cluster G using Chinese Whispers.



Concept Discovery: The Results

Gold Standard: RuThes-lite 2.0.
Metrics: pairwise IR metrics and V-measure.

Method	# sets	Pr	Re	F_1-score	V-measure
Chinese Whispers	16 063	0.135	0.022	0.038	0.866
MaxMax	16 870	0.181	0.004	0.007	0.835
This	5 984	0.193	0.039	0.065	0.860

Examples

- {*зелёный*, *неспелый*, *недозрелый*, ...}
- {*зелёный*, *юный*, *молодой*, ...}
- {*билет*, *купюра*, *банкнота*, ...}
- {*билет*, *свидетельство*, *удостоверение*, ...}

Relation Establishment

Definition

Hyponymy* и *hypernymy *are asymmetric semantic relations that connect the more specific term (the hyponym) to the more general term (the hypernym).*

The *is-a* relation: *cat* $\overset{is-a}{\rightarrow}$ *animal* (*genus* and *species* in biology).

Challenges

- Availability of dictionaries.
- Relations between the synsets needed.

Idea: transform the $\vec{x}$ embedding into its hypernym embedding $\vec{y}$ and use these projections for connecting the synsets.

Projection Learning

Embeddings: 100 dimensions, skip-gram, 13 billion words corpus.

Baseline (Fu et al., 2014)

$$\Phi^* = \arg\min_{\Phi} \frac{1}{N} \sum_{(\vec{x},\vec{y})} \|\vec{x}\Phi - \vec{y}\|^2$$

Regularization (weighted by λ)

- hyponym $\vec{x}$: $\lambda \sum_{\vec{x}} (\vec{x}\Phi\Phi \cdot \vec{x})^2$
- synonym $\vec{z}$ of $\vec{x}$: $\lambda \sum_{(\vec{x},\vec{z})} (\vec{x}\Phi\Phi \cdot \vec{z})^2$

Training set: $21\,997$ pairs; **test set:** $10\,811$ pairs; k-means clustering; *hit*@10 ≈ 0.37.

Concept Linking

So far, the relations correspond to individual words. However, now we have nearest neighbours $\mathrm{NN}(\vec{x})$ for the embedding x correspoding to the word x.

Heuristic

1. Compute the matchings $C(s) = \underset{g \in |V| \setminus \{s\}}{\arg\max} \left| g \cap \bigcup_{x \in s} \mathrm{NN}(\vec{x}\Phi^*) \right|$ for each synset s.
2. Connect the synset s with $C(s)$.

Looking ahead, the performance of this heuristic combined with projection learning is not impressive, but the baseline is still needed.

Relation Establishment: The Results

A candidate relation is said to be correct $\iff$ there exists a directed path from the hyponym concept to the hypernym concept in RuThes-lite 2.0.

Method	**# candidates**	**# correct**
Russian Wiktionary	1 627	113
Projection Learning	3 918	133

Examples

- {*атлет*, *силач*, ...} $\rightarrow$ {*личность*, *человек*}
- {*преграда*, *препона*, ...} $\rightarrow$ {*препятствие*, *трудность*}
- {*наводнение*, *потоп*, ...} $\rightarrow$ {*злосчастие*, *катаклизм*}

Conclusion

- An ontology induction approach utilizing both dictionaries and word embeddings has been described and prelimiary evaluated.
- Further studies should be primariliy focused on improving the relation establishment approach.

Open Source Software

- https://github.com/dustalov/concept-discovery
- https://github.com/dustalov/projlearn

Thank You!

Dmitry Ustalov

https://linkedin.com/in/ustalov

dau@imm.uran.ru

The reported study was funded by Russian Foundation for Basic Research according to the research project № 16-37-00354 мол_а "Adaptive Crowdsourcing Methods for Linguistic Resources". This work was supported by the Russian Foundation for the Humanities project № 13-04-12020 "New Open Electronic Thesaurus for Russian" and project № 16-04-12019 "RussNet and YARN thesauri integration". The present work is also supported by a short-term grant provided by the Deutscher Akademischer Austauschdienst.