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 libre license.
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.

The Approach

Principles

- Re-using the existing resources.
- Minimal efforts from the humans.
- Focusing on nouns, is-a relations, and domain ontologies.
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.
Construct an ego-network \( Ego(v) \) for \( v \in V \) and exclude \( v \).

Cluster \( Ego(v) \) using Chinese Whispers.

Reconstruct and disambiguate the global graph \( G \).

Cluster \( G \) using Chinese Whispers.
Gold Standard: RuThes-lite 2.0.

Metrics: pairwise IR metrics and V-measure.

<table>
<thead>
<tr>
<th>Method</th>
<th># sets</th>
<th>Pr</th>
<th>Re</th>
<th>F$_1$-score</th>
<th>V-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chinese Whispers</td>
<td>16 063</td>
<td>0.135</td>
<td>0.022</td>
<td>0.038</td>
<td>0.866</td>
</tr>
<tr>
<td>MaxMax</td>
<td>16 870</td>
<td>0.181</td>
<td>0.004</td>
<td>0.007</td>
<td>0.835</td>
</tr>
<tr>
<td>This</td>
<td>5 984</td>
<td>0.193</td>
<td>0.039</td>
<td>0.065</td>
<td>0.860</td>
</tr>
</tbody>
</table>

Examples

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

*Hyponymy* and *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* $\xrightarrow{is-a}$ *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.
**Embeddings:** 100 dimensions, skip-gram, 13 billion words corpus.

**Baseline (Fu et al., 2014)**

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

**Regularization (weighted by \(\lambda\))**

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

**Training set:** 21 997 pairs; **test set:** 10 811 pairs; \(k\)-means clustering; \(hit@10 \approx 0.37\).
So far, the relations correspond to individual words. However, now we have nearest neighbours $\text{NN}(\vec{x})$ for the embedding $x$ corresponding to the word $x$.

**Heuristic**

1. Compute the matchings $C(s) = \arg \max_{g \in |V| \setminus \{s\}} \left| g \cap \bigcup_{x \in s} \text{NN}(\vec{x} \Phi^*) \right|$ 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.
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.

<table>
<thead>
<tr>
<th>Method</th>
<th># candidates</th>
<th># correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Russian Wiktionary</td>
<td>1627</td>
<td>113</td>
</tr>
<tr>
<td>Projection Learning</td>
<td>3918</td>
<td>133</td>
</tr>
</tbody>
</table>

Examples

- \{атлет, силач, ...\} \rightarrow \{личность, человек\}
- \{преграда, препона, ...\} \rightarrow \{препятствие, трудность\}
- \{наводнение, потоп, ...\} \rightarrow \{злосчастье, катаклизм\}
Conclusion

- An ontology induction approach utilizing both dictionaries and word embeddings has been described and preliminary evaluated.
- Further studies should be primarily 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

LinkedIn: https://linkedin.com/in/ustalov
Email: 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.