Label Propagation for Tax Law Thesaurus Extension

Markus Müller, 09.11.2018, Master's Thesis Final Presentation

Chair of Software Engineering for Business Information Systems (sebis) 
Faculty of Informatics 
Technische Universität München 
www.matthes.in.tum.de

Prof. Dr. Stephan Günnemann (Group for Data Mining and Analytics) 
Jörg Landthaler, Elena Scepankova

Advisors
Outline

Motivation
- Problem: Thesauri in the Legal Context
- Base Technology: Word Embeddings
- Opportunity: Label Propagation on Graphs

Research Approach
- Research Questions
- Research Methods
- Thesaurus Extension Tool

Evaluation Results
- Quantitative Evaluation
- Qualitative Evaluation
- Baseline Comparison

Conclusion & Future Work
Problem: Thesauri in the Legal Context

Legal Content Providers
Provide their users with access to relevant legal documents

Leading Providers in Germany

Wolters Kluwer
DATEV
Haufe
ottoschmidt
GH-BECK

Thesauri enhance Information Retrieval via Synonym Sets

Search Query Expansion

Abwrackprämie

Also showing results for “Umweltprämie”

[...] Abwrackprämie, the colloquial term for Umweltprämie [...]

Creating and Maintaining Thesauri is hard

Mostly manual work, multiple domain-specific thesauri

Wolters Kluwer 2016 [1]
Focus: Thesaurus Extension as a Solution Approach

Existing Thesaurus → Suggest words from text corpus as synonym set (synset) additions → Text Corpus

Subject to research at this chair:
Landthaler et al. (2017) extended synsets starting from individual synset words
Potential Use-Cases for Thesaurus Extension

- **semi-automation, quality assurance**
  - Suggest synset adjustments

- **Thesaurus Creation**
  - Suggest additions to manually created synsets

- **Thesaurus Maintenance**
  - Thesaurus Usage

- **Thesaurus Enrichment**

- **thesauri linkage**
  - Identify relations between synsets of different thesauri
Word Embedding Technologies map similar words to similar vectors

⇒ Nearest Neighbors: Extend synset with words close to synset words

But then: Overall structure is not taken into account

⇒ Opportunity: Semi-Supervised Learning

Blue & Red: Words from different existing synsets
Green: Extension suggestion

A & B: Labeled with different synsets
Rest: Unlabeled

X would fit better to B than to A
Research Idea: Label Propagation for Thesaurus Extension

Label Propagation is used by Google in Combination with Word Embeddings for knowledge graph extension, e.g. for Emotion Association and Smart Replies.

RQ1: Can we apply Label Propagation to Word Embeddings to find new Synonyms?

Intuition

Text Corpus → Embedding Generation → Word Embeddings → Graph Construction → Sparsely-labeled Graph with Labels from existing Thesaurus → Label Propagation → Fully-Labeled Graph

Research Questions

1. How can we get **semantic & context information into a graph** for LP? (RQ2)
2. Can we **model the thesaurus extension problem** on the LP technology? (RQ3)
3. What LP **algorithms work best**? (RQ4)
4. Is LP a **suitable technology** for thesaurus extension in the legal domain? (RQ1)
5. How much **automation** for thesaurus creation is achievable with LP? (RQ5)
Research Approach

Build a Thesaurus Extension Tool for trying out many approaches

Can we model the thesaurus extension problem on the LP technology? (RQ3)

How can we get semantic & context information into a graph for LP? (RQ2)

What LP algorithms work best? (RQ4)

Is LP a suitable technology for thesaurus extension in the legal domain? (RQ1)

How much automation for thesaurus creation is achievable with LP? (RQ5)

Quantitative Evaluation
Automatic Parameter Studies

Qualitative Evaluation
Manual Studies

Comparison with Vanilla Word Embeddings Approach
Thesaurus Extension Tool: Architecture

Extendable & Open Source on sebischair@GitHub

Special Character Handling

- word2vec, fastText, GloVe
- Corpus Preprocessing
- Embedding Generation
- Graph Construction
- k-nearest-neighbors
- ε-neighborhood graph
- Thesaurus Preprocessing
- Thesaurus Sampling
- Graph Labeling
- Label Propagation
- Evaluation Results
- Many variants incl. parameters implemented for later evaluation

via Group for Data Mining and Analytics

Buschmann et al. (1996)
Quantitative Evaluation: Set-up

Tax Law Data Set by DATEV (in German)
- text corpus: 132,581 legal documents
- handcrafted existing thesaurus: 12,288 synsets

Evaluation Thesaurus (Subset):
- 2,552 thesaurus synsets
  - Training Set: 3,277 words
  - Test Set: 2,887 words

Hyper-Parameter Studies on these Phases

Goal: Find hyper-parameter configuration with highest accuracy
⇒ as input for Qualitative Evaluation

Challenge: Lots of possible configurations (> 1,000 runs)
Quantitative Evaluation: Lessons Learned & Final Result

Greatest performance impact: Word Embeddings Choice

High performance through hyper-parameter optimization

Optimized Configuration Results

But: Also good suggestions outside of the existing thesaurus?

Configuration:
Pre-Processing: Keep letters & hyphens, muß⇒muss, single line saving
Embedding Generation: 400 dimensions, 40 iterations
Graph Construction: k-nearest neighbors, k=12, weighted undirected edges, no self-references allowed
Label Propagation: LabelSpreading, α=0.2, 15 iterations
Qualitative Evaluation: Set-up

Show synset suggestions to humans & get ratings

Pre-Study

Identify influence factors for good suggestions

Main Study (2x)

Rate suggestions of best configurations

Scores
0: Not similar to predicted synset
1: Same semantic area
2: Should be added to synset

Rated 54 synsets per study, 10 suggestions per synset ⇒ 540 ratings/study

• Originally planned with legal experts
• In the end, conducted by Jörg Landthaler & Markus Müller, supported by Text Corpus via ElasticSearch instance
Qualitative Evaluation: Pre-Study Lessons Learned

High confidence, high synset training number and low synset prediction number lead to better rating.

E.g. correlation between prediction confidence and score

Scores
0: Not similar to predicted synset
1: Same semantic area
2: Should be added to synset

Prediction confidence of algorithm vs. Human Score

Scores:
0: Not similar to predicted synset
1: Same semantic area
2: Should be added to synset

E.g. correlation between prediction confidence and score.
Qualitative Evaluation: Main Study Lessons Learned

fastText again considerably better than word2vec

But: Why does fastText perform better?

Ratings

Scores
0: Not similar to predicted synset
1: Same semantic area
2: Should be added to synset

fastText

word2vec

<table>
<thead>
<tr>
<th>Scores</th>
<th>fastText</th>
<th>word2vec</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>blue</td>
<td>blue</td>
</tr>
<tr>
<td>1</td>
<td>blue</td>
<td>blue</td>
</tr>
<tr>
<td>2</td>
<td>blue</td>
<td>blue</td>
</tr>
</tbody>
</table>

© sebis
Qualitative Evaluation: Interpretation

fastText predominantly suggests **syntactically** similar words, word2vec suggests really different words (⇒ more interesting)

Our evaluations favored syntactically similar words

<table>
<thead>
<tr>
<th>Existing Synset Words</th>
<th>fastText Propagation (Top 5)</th>
<th>word2vec Propagation (Top 5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>kst-bescheid</td>
<td>körperschaftsteuer-bescheids</td>
<td>erstattungsjahre</td>
</tr>
<tr>
<td>kst-bescheide</td>
<td>kst-bescheiden</td>
<td>leistungsgebote</td>
</tr>
<tr>
<td>körperschaftsteuer-bescheid</td>
<td>körperschaftsteuer-bescheide</td>
<td>vek-bescheide</td>
</tr>
<tr>
<td>körperschaftsteuerbescheid</td>
<td>körperschaftsteuerbescheide</td>
<td>zuwendungsbestätigungsempfänger</td>
</tr>
<tr>
<td></td>
<td>körperschaftsteuerbescheiden</td>
<td>umsatzsteuervorauszahlungsbescheide</td>
</tr>
</tbody>
</table>

Example

We compiled a list of common challenges around Thesaurus Extension
„Synset Vector“ Baseline: Approach

- Nearest neighbors approach, operates directly on word embeddings
- Self-designed, inspired by Rothe and Schütze (2016) [4]
Baseline performs equal or better than label propagation approach, while being less complex.

**Quantitative Results** with baseline k=200

**Qualitative Results** with baseline k=30

*Scores*

0: Not similar to predicted synset
1: Same semantic area
2: Should be added to synset
Label Propagation approach was not better than Baseline, but overall results were promising.

fastText and word2vec predictions could be used in a semi-automated way for Thesaurus Extension.

And: We contributed to the problem area.
Conclusion: Contributions & Future Work

Contributions

- Created Open Source "ThesaurusLabelPropagation" tool
  - Found implementation issues around label propagation in "scikit-learn" (32,000 stars)
  - Significantly optimized performance for graph construction on word embeddings
- Conducted multiple hyper-parameter studies (>1000 individual runs) & optimized configurations
- Rated configurations within 5 qualitative evaluations (overall 2,500 suggest. manually rated)
  - Identification of influence factors for quality of suggestion results
  - Classification of typical thesaurus challenges
- Introduced & evaluated new baseline approach

Future Work with regards to Label Propagation

- Evaluation with a corpus in a different language and/or more training data?
- Evaluation within a different application area besides tax law?
- Augment word embeddings with other semantic knowledge, e.g. Wikidata, Wikipedia, Freebase
References


Hyper-Parameter Study on Word Embeddings

![Graph 1: Accuracy vs. Dimensions](chart1.png)

- Blue line: word2vec
- Orange line: fasttext

![Graph 2: Accuracy vs. Iterations](chart2.png)

- Blue line: word2vec
- Orange line: fasttext
Hyper-Parameter Study on Graph Construction

![Graphs showing accuracy vs. k and accuracy vs. ε (Radius)]
Backup

Qualitative Evaluation: Correlations

- Spearman \(r = 0.4; p = 3.8e-21\)
- Spearman \(r = 0.12; p = 0.0074\)
- Spearman \(r = -0.2; p = 8.2e-06\)
## Challenges around Thesaurus Extension

<table>
<thead>
<tr>
<th>Category</th>
<th>Type</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Semantic Challenges</strong></td>
<td>Context-dependent word meaning</td>
<td>leiter (ladder vs. manager)</td>
</tr>
<tr>
<td></td>
<td>Identification of defining word parts</td>
<td>milchwirtschaft (“milch” is more defining)</td>
</tr>
<tr>
<td></td>
<td>Broader or more specific terms</td>
<td>steuerrecht, einkommenssteuerrecht</td>
</tr>
<tr>
<td><strong>Syntactic Challenges</strong></td>
<td>Inflected words</td>
<td>zeitungsträgern, zeitungsträger</td>
</tr>
<tr>
<td></td>
<td>Same word stem</td>
<td>stornierung, stornieren</td>
</tr>
<tr>
<td></td>
<td>Word splits</td>
<td>eigentümerehegatten, eigentümer ehergatten</td>
</tr>
<tr>
<td></td>
<td>Hyphenation</td>
<td>zwölfmonatszeitraum, zwölfmonats-zeitraum</td>
</tr>
<tr>
<td></td>
<td>Old spellings/Misspellings</td>
<td>fitneß-studios, fitness-studio</td>
</tr>
<tr>
<td></td>
<td>Abbreviations</td>
<td>ustk, ust-kartei</td>
</tr>
<tr>
<td></td>
<td>Numbers</td>
<td>12-monatsfrist, zwölfmonatsfrist</td>
</tr>
</tbody>
</table>
Backup
Possible Reasons and Future Work

Language & Training Data
Evaluation with a corpus in a different language and/or more training data?

Context of Tax Law
Evaluation within a different application area?

Graph Type
Augment word embeddings with other semantic knowledge, e.g. Wikidata, Wikipedia, Freebase [3]
Supervised learning: Learn on labeled training instances, perform prediction on unknown test data.

Inductive semi-supervised learning: Learn on labeled training instances and unlabeled training instances, perform prediction on unknown test data.

Transductive semi-supervised learning: Learn on labeled training instances and unlabeled training instances, perform prediction on known test [training] data.