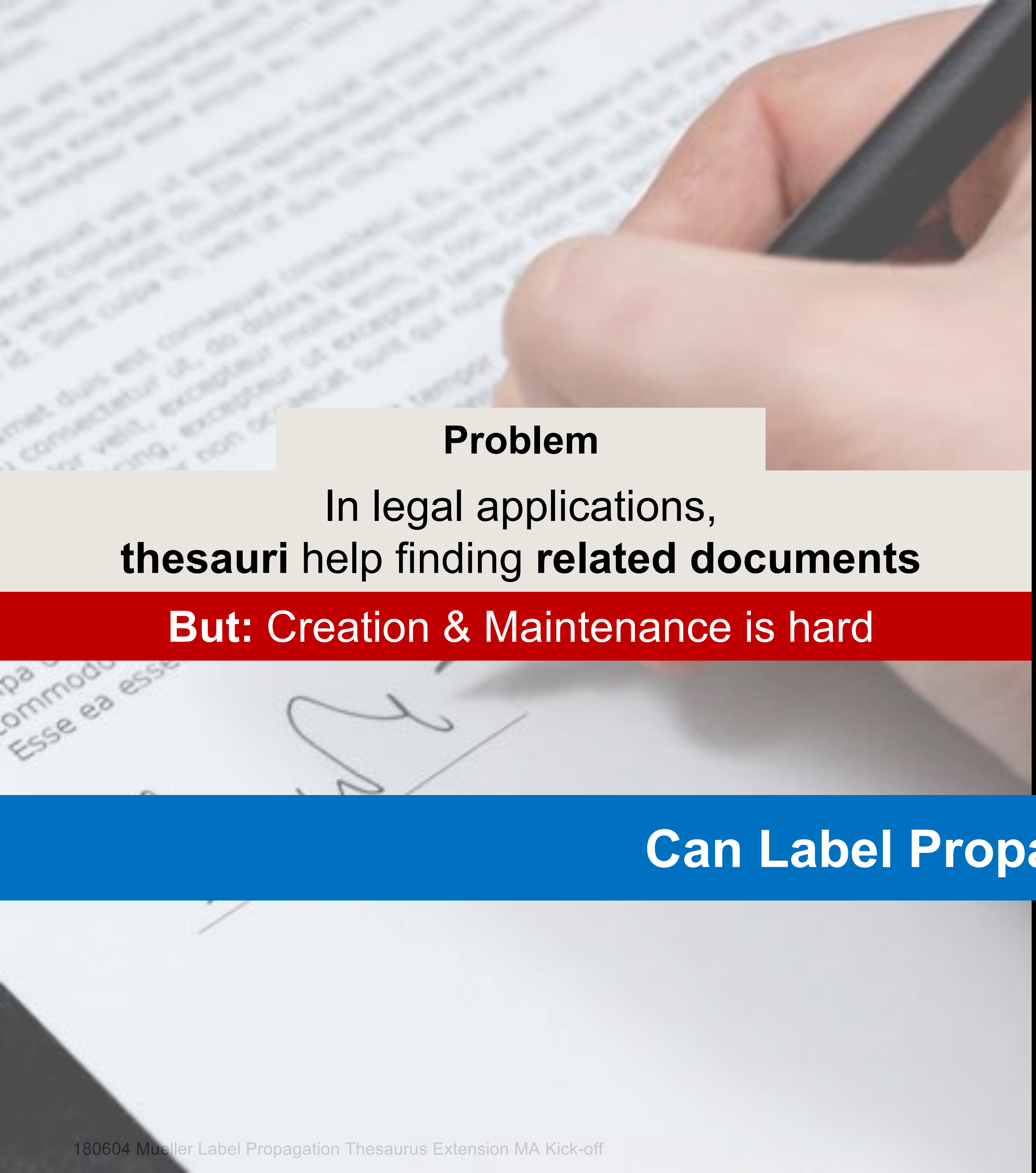
Label Propagation for Tax Law Thesaurus Extension

Markus Müller, 04.06.2018, Master’s Thesis Kick-Off Presentation

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

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



Problem

In legal applications,
thesauri help finding **related documents**

But: Creation & Maintenance is hard

Technology

Label Propagation can find communities in graphs
semi-supervised learning

Can Label Propagation help us?

Outline

Motivation

- Problem: Thesauri in the Legal Context
- Opportunity: Label Propagation on Graphs
- Related Work

Research Questions

Research Approach

- Technology Flow
- Concept
- Challenges

Timeline

Motivation

Problem: Thesauri in the Legal Context

What is a Thesaurus?

A Collection of Synonym Sets (Synsets)

{ example, instance, model, case, illustration, lesson, object, part, pattern, precedent, symbol,… }

Can contain other relations between words,
e.g. broader terms, narrower terms, top term, antonyms

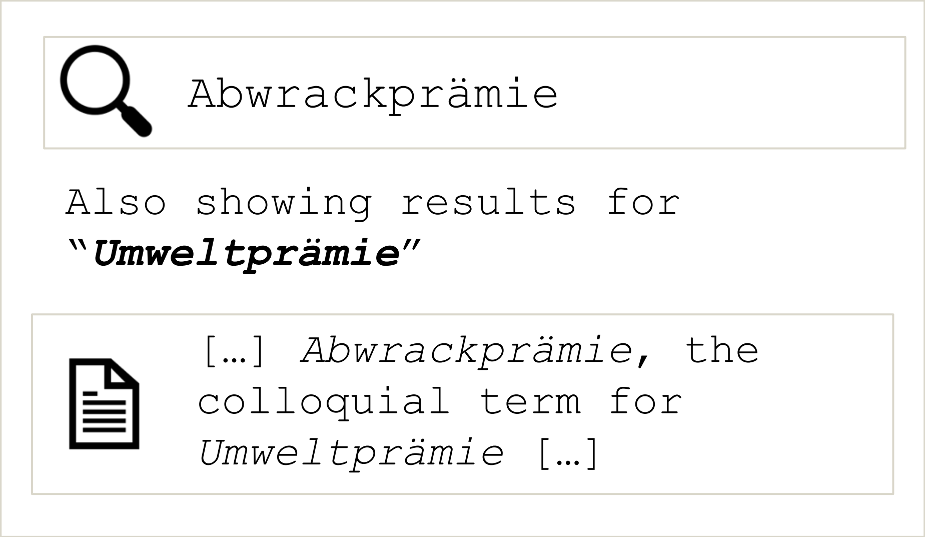
Example from Thesaurus.com

Motivation

Problem: Thesauri in the Legal Context

Why are Thesauri useful, especially in the Legal Domain?

Thesauri enhance Search
Search Query Expansion



Abwrackprämie

Also showing results for "***Umweltprämie***"

[…] *Abwrackprämie*, the colloquial term for *Umweltprämie* […]

Legal work deals with lots of texts

Laws, past cases, comments on laws…

Legal Content Providers
need to create & maintain thesauri

Wolters Kluwer 2016 [1]: "***Legal Thesauri*** *are the* ***backbone*** *of many application features in JURION*"

[1] C. Dirschl, "Thesaurus Generation and Usage at Wolters Kluwer Deutschland GmbH," *Jusletter IT 25. Februar 2016*, Feb. 2016.

Motivation

Problem: Thesauri in the Legal Context

Creating and Maintaining a Thesaurus is hard

Wolters Kluwer 2016 [1]:

“*Thesaurus creation is a very challenging task*”

“*do not aim for having one single thesaurus in place […], but to have smaller, domain specific thesauri*” (e.g. tax law, tenancy law,…)

For each thesaurus:

- “*1 to 2 person months internal effort”*
- “*10 to 20k€ external costs*”
- “*Normally there are no processes for [maintenance] in place*“

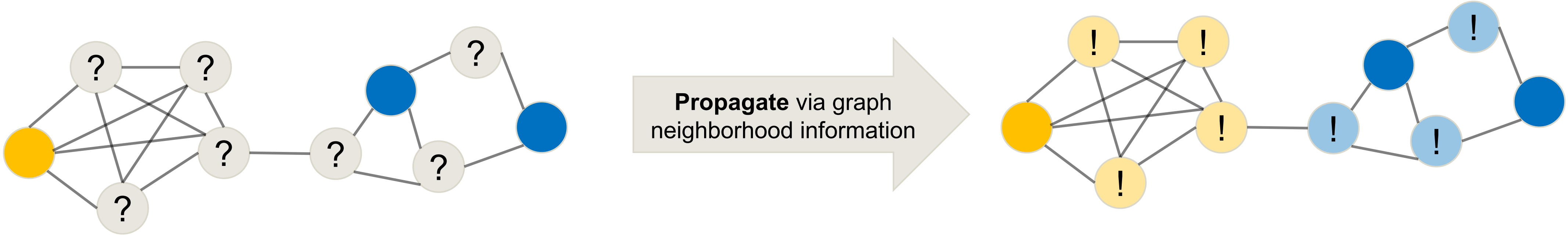
[1] C. Dirschl, “Thesaurus Generation and Usage at Wolters Kluwer Deutschland GmbH,” *Jusletter IT 25. Februar 2016*, Feb. 2016.

Motivation

Opportunity: Label Propagation on Graphs

Label Propagation

Family of **semi-supervised** machine learning methods
Use **few labeled** records & **graph structure**
to label a **large unlabeled** dataset



Very good **performance**, even with large datasets and lots of labels

Can we apply Label Propagation to
find new Synonyms?

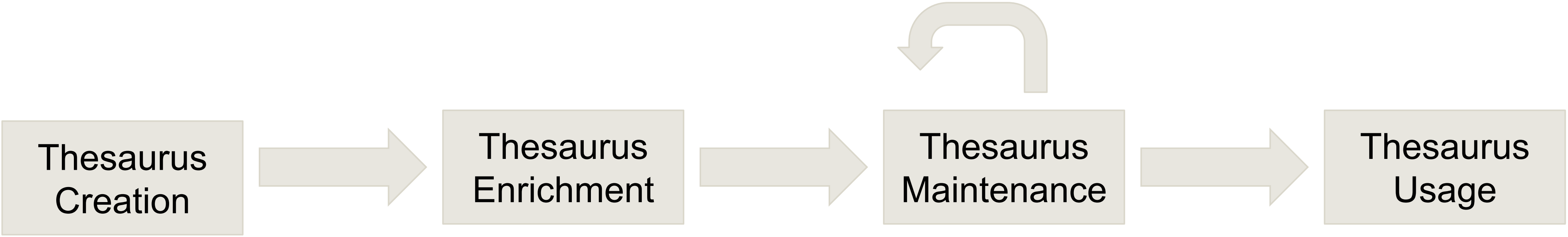
where Label ≙ Synset

Motivation

Opportunity: Label Propagation on Graphs

Process around Thesauri [1]

Process [1]



LP Opportunities

semi-automation

- *Suggest additions to manually created synsets*
- *Suggest new synsets*

thesauri linkage

Suggest relations between synsets of different thesauri

semi-automation, quality assurance

Suggest synset adjustments

[1] C. Dirschl, “Thesaurus Generation and Usage at Wolters Kluwer Deutschland GmbH,” *Jusletter IT 25. Februar 2016*, Feb. 2016.

Motivation
Related Work

Previous Research on **Thesaurus Extension** at sebis

[2] J. Landthaler, B. Waltl, D. Huth, D. Braun, and F. Matthes, “Extending Thesauri Using Word Embeddings and the Intersection Method,” 2018.

Research on **Label Propagation** & its **Application**

[3] S. Ravi and Q. Diao, “Large Scale Distributed Semi-Supervised Learning Using Streaming Approximation,” *arXiv:1512.01752 [cs]*, Dec. 2015.
[4] A. Kannan *et al.*, “Smart Reply: Automated Response Suggestion for Email,” *arXiv:1606.04870 [cs]*, Jun. 2016.
[5] X. Zhu and Z. Ghahramani, “Learning from labeled and unlabeled data with label propagation,” 2002.
[6] Y. Bengio, O. Delalleau, and N. Le Roux, “Label Propagation and Quadratic Criterion,” *Semi-Supervised Learning*, Sep. 2006.

Cooperation with Prof. Günnemann
Data Mining and Analytics Group

[7] W. Gatterbauer, S. Günnemann, D. Koutra, and C. Faloutsos, “Linearized and Single-pass Belief Propagation,” *Proc. VLDB Endow.*, vol. 8, no. 5, pp. 581–592, Jan. 2015.
[8] D. Eswaran, S. Günnemann, C. Faloutsos, D. Makhija, and M. Kumar, “ZooBP: Belief Propagation for Heterogeneous Networks,” *Proc. VLDB Endow.*, vol. 10, no. 5, pp. 625–636, Jan. 2017.

Research Questions

Is LP a **suitable technology** for thesaurus extension in the legal domain?

Can we **model the thesaurus extension problem** on the LP technology?

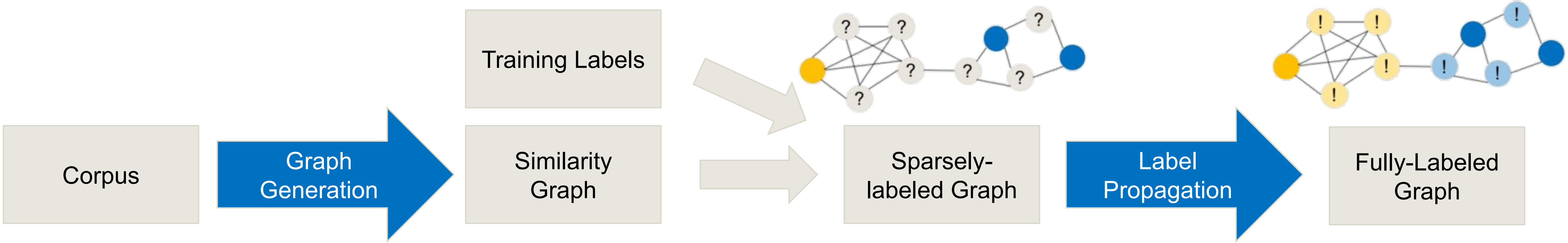
How can we get **semantic & context information into a graph** for LP?

How much **automation** for **thesaurus creation** is achievable with LP?

What LP **algorithms work best**?

Research Approach

Technology Flow



(from thesaurus by DATEV)

Training Labels

Corpus

Graph Generation

Similarity Graph

Sparsely-labeled Graph

Label Propagation

Fully-Labeled Graph

Tax Law Corpus provided by DATEV

Many options to try:

- Various word embeddings technologies
- Various hyper-parameters
- Various graph augmentation possibilities

Many options to try:

- Standard LP
- Label Spreading
- Belief Propagation
- Modifications by reference papers

⇒ Larger synsets than before

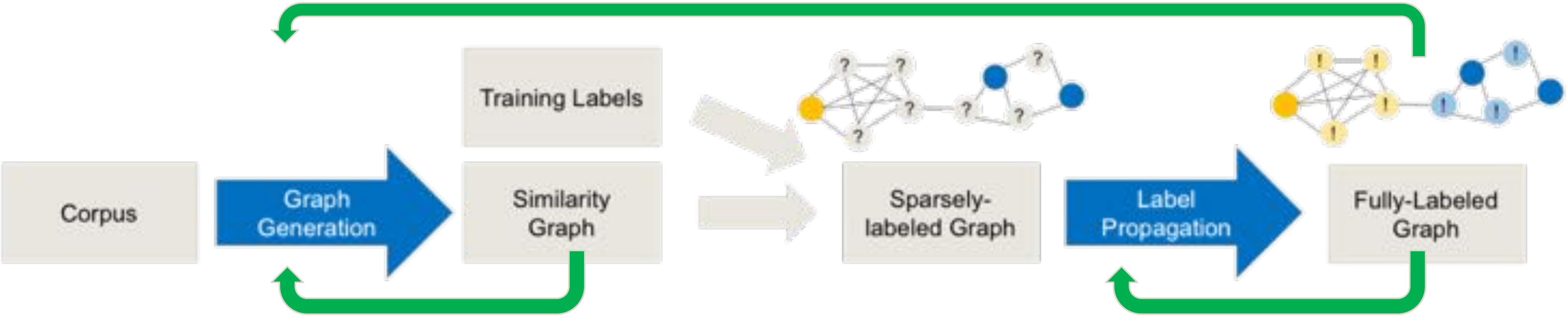
Research Approach

Concept

Build a framework for **iteratively** trying out many approaches

Evaluate

Continuous evaluation on corpus labels	**Continuous comparison** to baseline algorithms	**Manual review** by experts
Quantitative	*Quantitative*	*Qualitative*



Research Approach

Challenges

Analogy **Synset = Label**
might not work out of the box

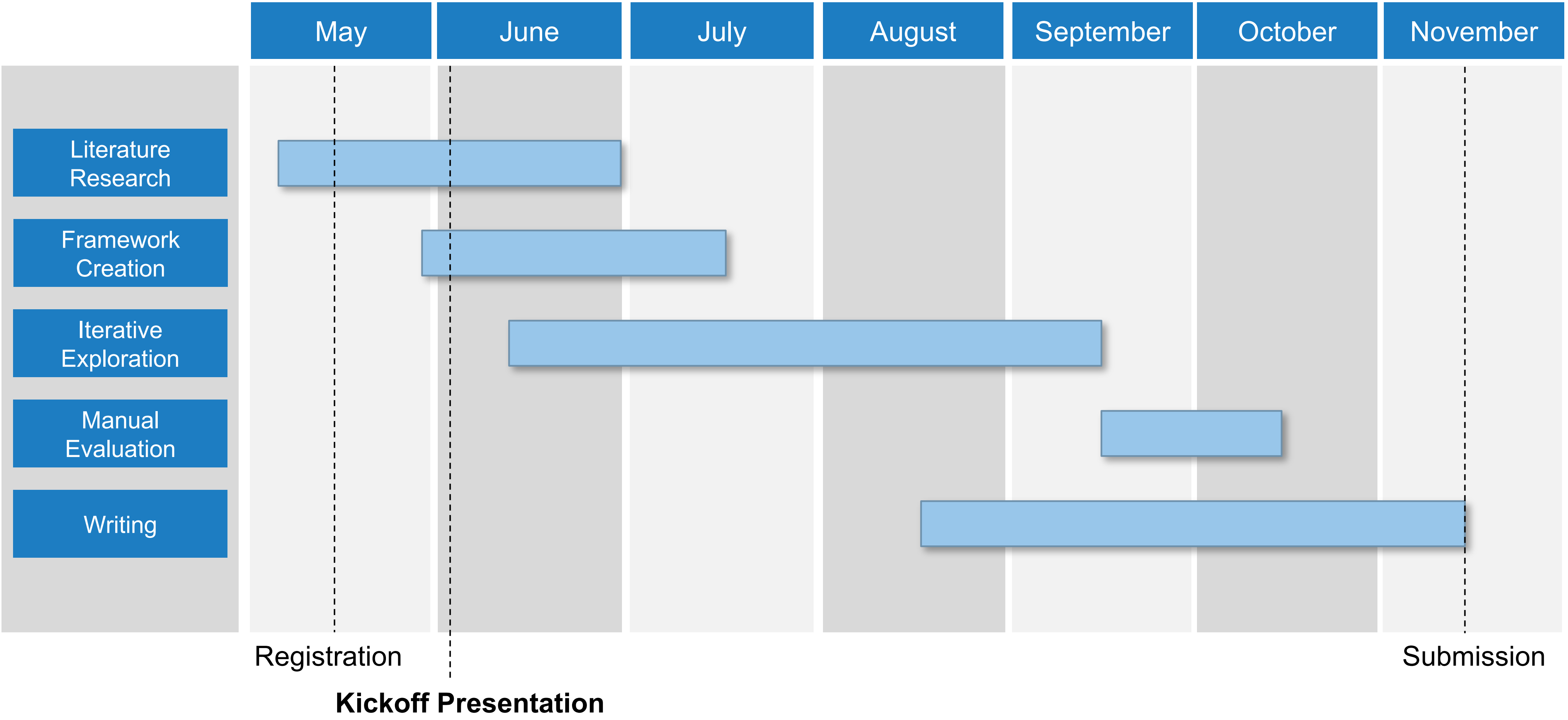
Many different options and approaches

Several aspects out of scope:

- compound words
- multi-label (multiple synsets/word)

Will it work on a **different corpus**?

Timeline


May
June
July
August
September
October
November
Literature Research
Framework Creation
Iterative Exploration
Manual Evaluation
Writing
Registration
Kickoff Presentation
Submission

References

[1] C. Dirschl, “Thesaurus Generation and Usage at Wolters Kluwer Deutschland GmbH,” *Jusletter IT 25. Februar 2016*, Feb. 2016.

[2] J. Landthaler, B. Waltl, D. Huth, D. Braun, and F. Matthes, “Extending Thesauri Using Word Embeddings and the Intersection Method,” 2018.

[3] S. Ravi and Q. Diao, “Large Scale Distributed Semi-Supervised Learning Using Streaming Approximation,” *arXiv:1512.01752 [cs]*, Dec. 2015.

[4] A. Kannan *et al.*, “Smart Reply: Automated Response Suggestion for Email,” *arXiv:1606.04870 [cs]*, Jun. 2016.

[5] X. Zhu and Z. Ghahramani, “Learning from labeled and unlabeled data with label propagation,” 2002.

[6] Y. Bengio, O. Delalleau, and N. Le Roux, “Label Propagation and Quadratic Criterion,” *Semi-Supervised Learning*, Sep. 2006.

[7] W. Gatterbauer, S. Günnemann, D. Koutra, and C. Faloutsos, “Linearized and Single-pass Belief Propagation,” *Proc. VLDB Endow.*, vol. 8, no. 5, pp. 581–592, Jan. 2015.

[8] D. Eswaran, S. Günnemann, C. Faloutsos, D. Makhija, and M. Kumar, “ZooBP: Belief Propagation for Heterogeneous Networks,” *Proc. VLDB Endow.*, vol. 10, no. 5, pp. 625–636, Jan. 2017.

TUM
sebis

Master’s Student Informatics

Markus Müller
www.muellermarkus.com

Technische Universität München
Faculty of Informatics
Chair of Software Engineering for Business
Information Systems

Boltzmannstraße 3
85748 Garching bei München

Tel   +49.89.289.17132
Fax   +49.89.289.17136

mail@muellermarkus.com
wwwmatthes.in.tum.de

Backup

Supervised, Semi-Supervised, Transductive

Supervised learning: Learn on labeled training instances, perform prediction on unknown test data.

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

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

Chapter 6: Network Data, Mining Massive Datasets, Stephan Günnemann, WS 2016/17

Comment

In literature, propagation is often referred to as semi-supervised learning, but actually it is transductive learning. A solution would be to place both the inductive and the transductive approaches as categories of semi-supervised learning.

Backup

DATEV Corpus Stats

~130.000 separate texts

~140 Mio. words

~180.000 distinct words

© sebis