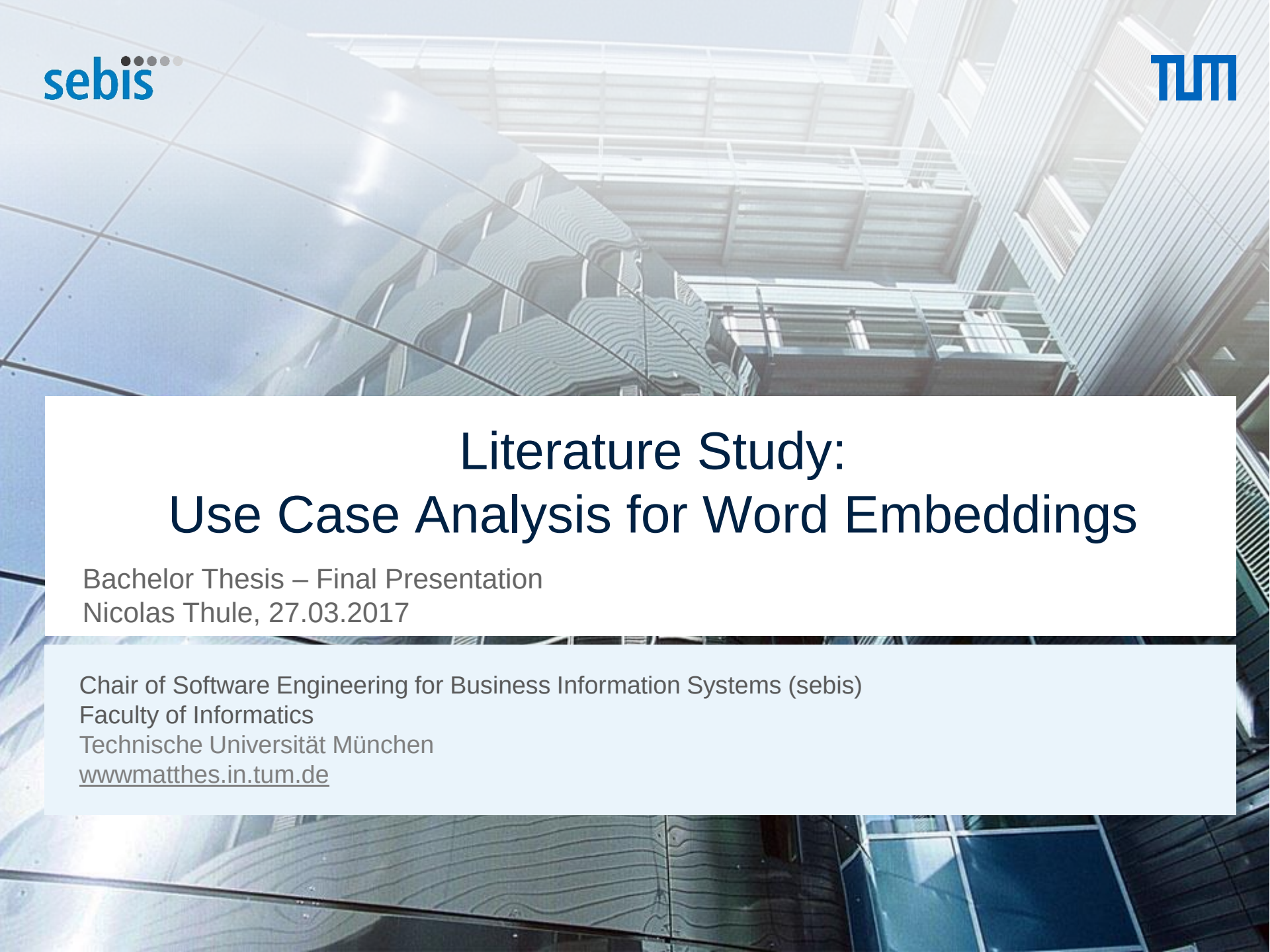


Literature Study: Use Case Analysis for Word Embeddings

Bachelor Thesis – Final Presentation

Nicolas Thule, 27.03.2017

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

Outline

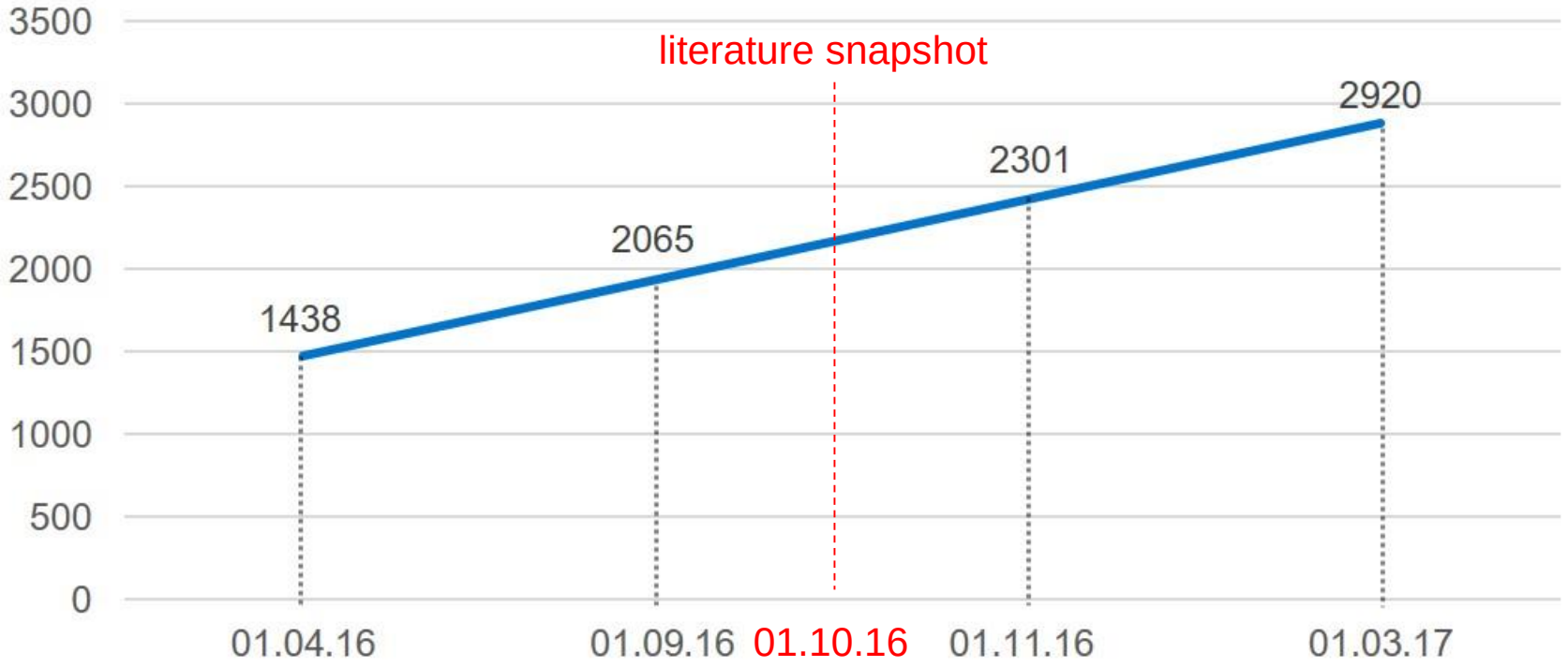
Motivation

Efficient estimation of word representations in vector space

T **Mikolov**, K Chen, G Corrado, J Dean - arXiv preprint arXiv:1301.3781, 2013 - arxiv.org

Abstract: We propose two novel model architectures for computing continuous vector representations of words from very large data sets. The quality of these representations is measured in a word similarity task, and the results are compared to the previously best ...

Zitiert von: 1438 Ähnliche Artikel Alle 8 Versionen Zitieren Speichern



=> no literature survey (esp. Use Cases) available

Outline

Data Acquisition

Crawl cited by publication list

01.10.2016
about 2100 publications in the Google Scholar Database referencing Mikolov et al.‘s work

Publish or Perish tool

crawls first 1000 publications by default
Publish or Perish

export to Excel and clean data

resulting in **989** publications as research basis

Data Acquisition

	A	B	C	D	E	F	G	H	
1	Cites	Authors	Title	Year	Source	Publisher	ArticleURL	CitesURL	GS
2	2271	T Mikolov, I S	Distributed r	2013	Advances in	papers.nips.	http://paper	https://scho	
3	1268	O Russakovs	Imagenet lar	2015	Internationa	Springer	http://link.s	https://scho	
4	899	J Penningtor	Glove: Globa	2014	EMNLP	pdfs.semant	https://pdfs.	https://scho	
5	631	QV Le, T Mik	Distributed F	2014	ICML	jmlr.org	http://www.	https://scho	
6	500	O Vinyals, A	Show and te	2015	Proceedings	cv-foundatic	http://www.	https://scho	
7	303	M Abadi, A A	Tensorflow:	2016	arXiv preprir	arxiv.org	https://arxiv	https://scho	
8	270	L Deng, D Yu	Deep learnir	2014	Foundations	nowpublishe	http://ftp.nc	https://scho	
9	264	A Frome, GS	Devise: A de	2013	Advances in	papers.nips.	http://paper	https://scho	
10	236	X Dong, E Ga	Knowledge v	2014	Proceedings	dl.acm.org	http://dl.acn	https://scho	

	A	B	C	D	E	F
1	Title	Authors	ArticleURL	Cites		
2	Distributed r	T Mikolov, I S	http://paper	2271		
3	Imagenet lar	O Russakovs	http://link.s	1268		
4	Glove: Globa	J Penningtor	https://pdfs.	899		
5	Distributed F	QV Le, T Mik	http://www.	631		
6	Show and te	O Vinyals, A	http://www.	500		
7	Tensorflow:	M Abadi, A A	https://arxiv	303		
8	Deep learnir	L Deng, D Yu	http://ftp.nc	270		
9	Devise: A de	A Frome, GS	http://paper	264		
10	Knowledge v	X Dong, E Ga	http://dl.acn	236		

research basis after cleaning

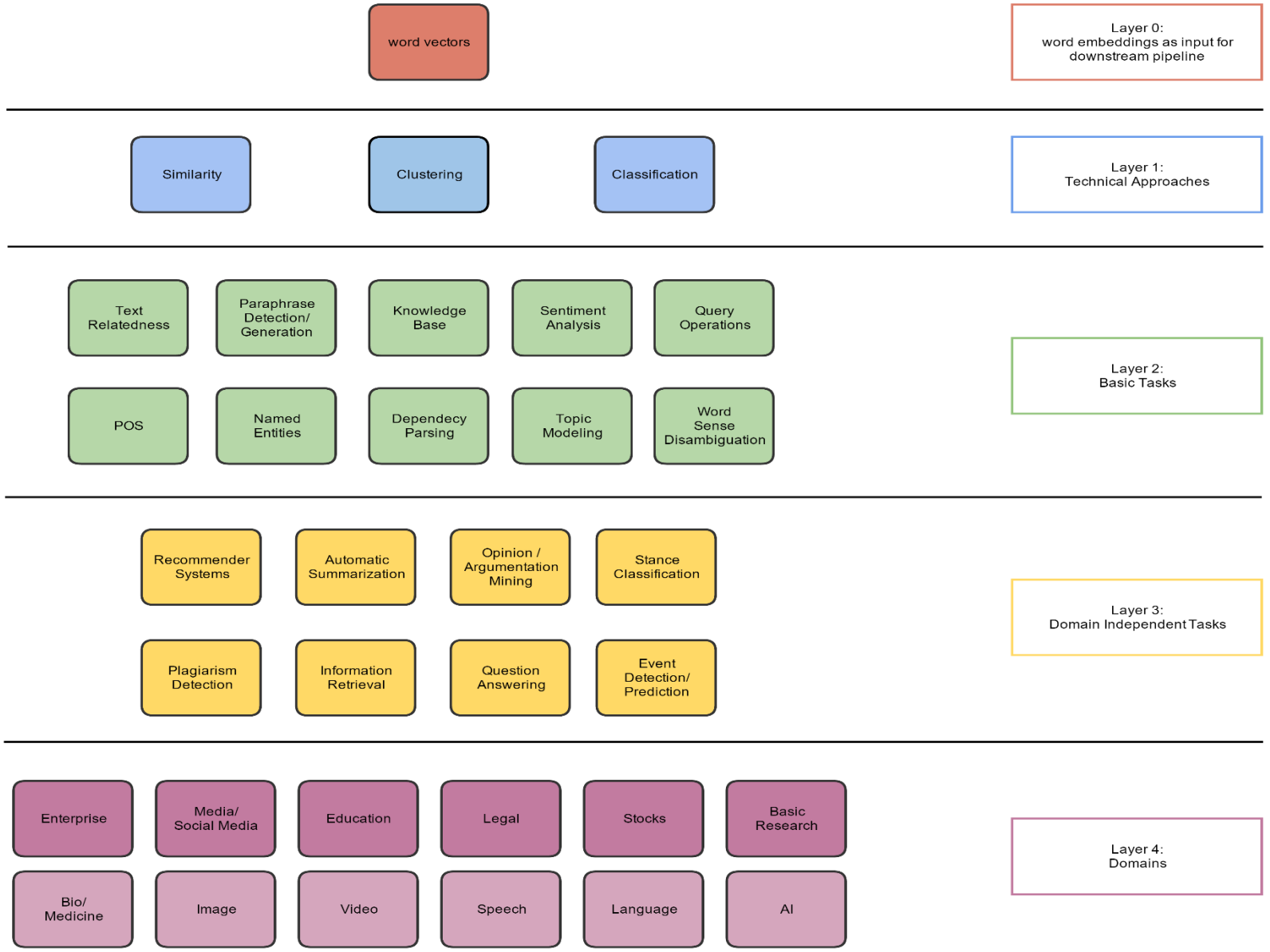
Outline

Research Questions and Contributions

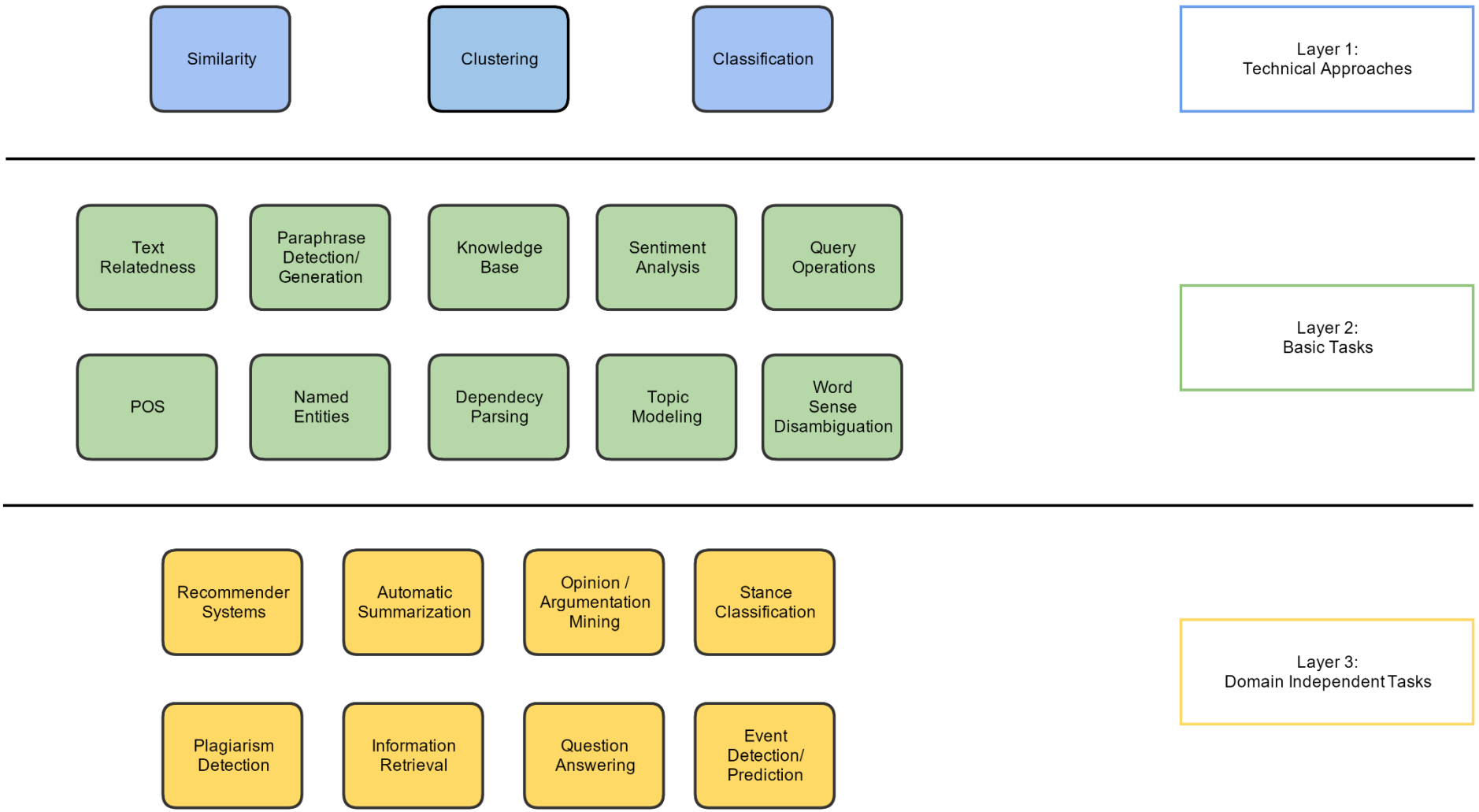
How can Use Cases for Word Embeddings be categorized?	Layer Model
What are the most frequent Use Cases?	Statistical Evaluation
Which Use Cases that apply to organizational contexts can be identified in the existing body of literature?	Statistical Evaluation
For the most interesting Use Cases: How do they work technically?	Utilization of the Layer Model
What further technological developments of Word2Vec can be identified in the body of literature?	Overview Further Developments

Outline

Layer Model


word vectors
Layer 0:
word embeddings as input for downstream pipeline
Similarity
Clustering
Classification
Layer 1:
Technical Approaches
Text Relatedness
Paraphrase Detection/ Generation
Knowledge Base
Sentiment Analysis
Query Operations
POS
Named Entities
Dependecy Parsing
Topic Modeling
Word Sense Disambiguation
Layer 2:
Basic Tasks
Recommender Systems
Automatic Summarization
Opinion / Argumentation Mining
Stance Classification
Plagiarism Detection
Information Retrieval
Question Answering
Event Detection/ Prediction
Layer 3:
Domain Independent Tasks
Enterprise
Media/ Social Media
Education
Legal
Stocks
Basic Research
Bio/ Medicine
Image
Video
Speech
Language
AI
Layer 4:
Domains

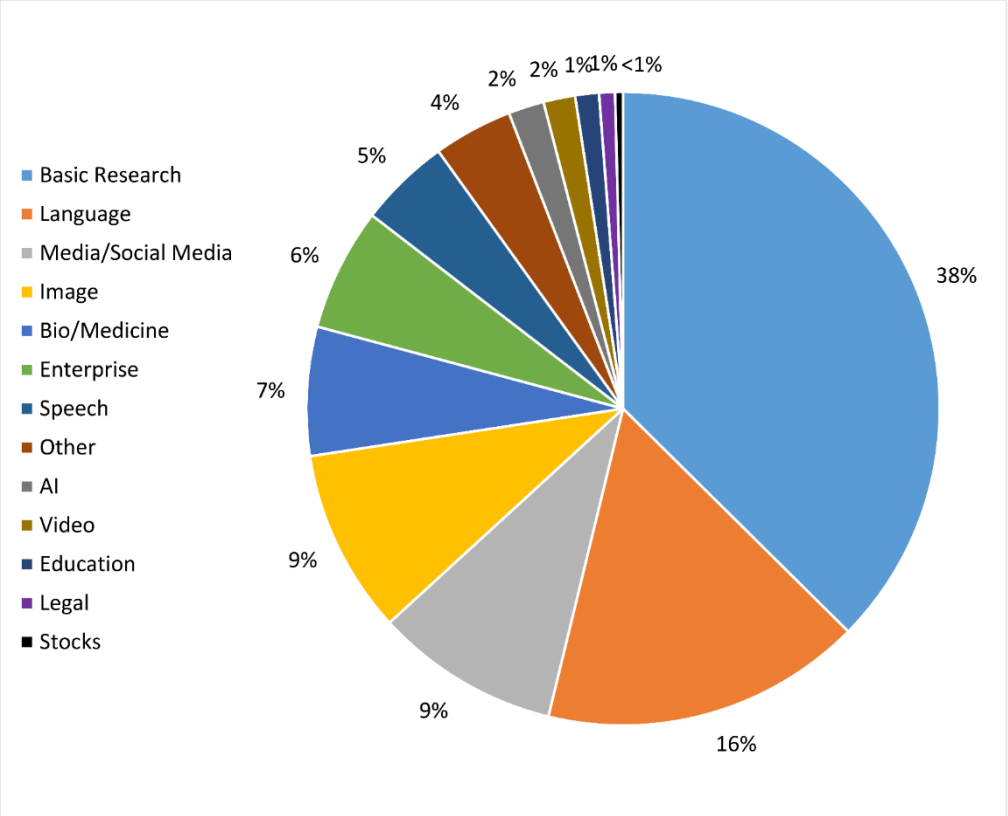
Layer Model


Similarity
Clustering
Classification
Layer 1:
Technical Approaches
Text Relatedness
Paraphrase Detection/ Generation
Knowledge Base
Sentiment Analysis
Query Operations
POS
Named Entities
Dependecy Parsing
Topic Modeling
Word Sense Disambiguation
Layer 2:
Basic Tasks
Recommender Systems
Automatic Summarization
Opinion / Argumentation Mining
Stance Classification
Plagiarism Detection
Information Retrieval
Question Answering
Event Detection/ Prediction
Layer 3:
Domain Independent Tasks

Layer Model

Enterprise	Media/ Social Media	Education	Legal	Stocks	Basic Research
Bio/ Medicine	Image	Video	Speech	Language	AI

Layer 4: Domains



Outline

1. Motivation
2. Data Acquisition
3. Research Questions
4. Layer Model
5. Utilization of the Layer Model
7. Further Developments
8. Lessons Learned
9. Future Work

Utilization of the Layer Model

Joshi, M., Hart, E., Vogel, M., and Ruvini, J.-D. (2015). **Distributed word representations improve ner for e-commerce.** In Proceedings of NAACL-HLT, pages 160–167.

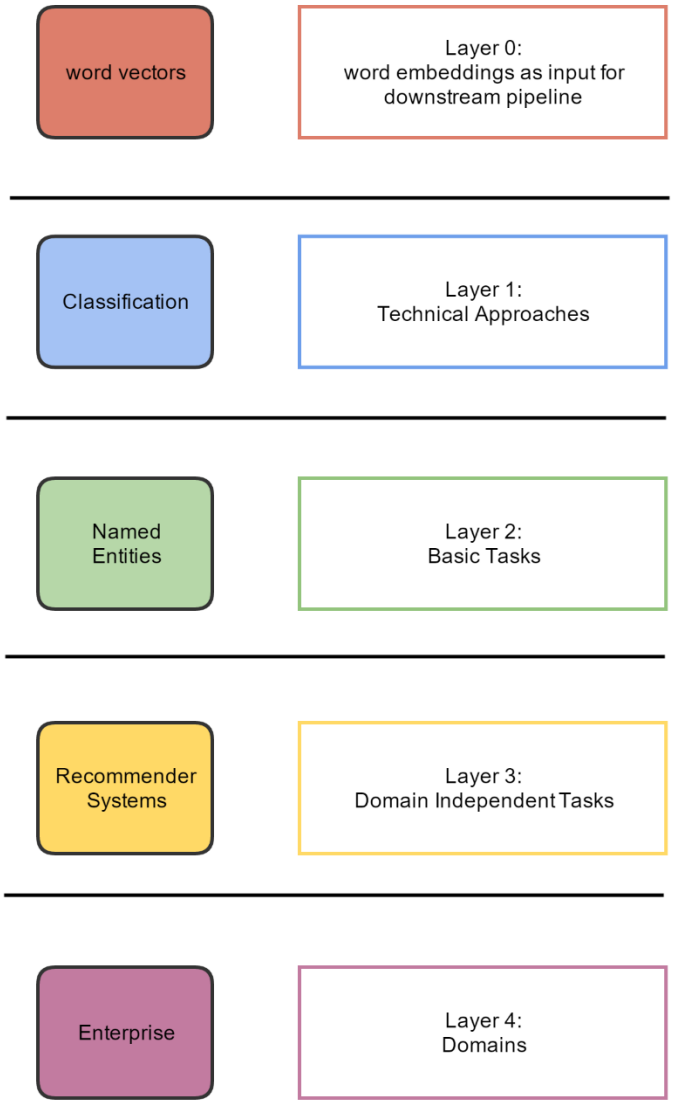
L0: Use of Word2Vec

L1: Conditional Random Field model -> classificiation

L2: Named Entity Recognition (NER)

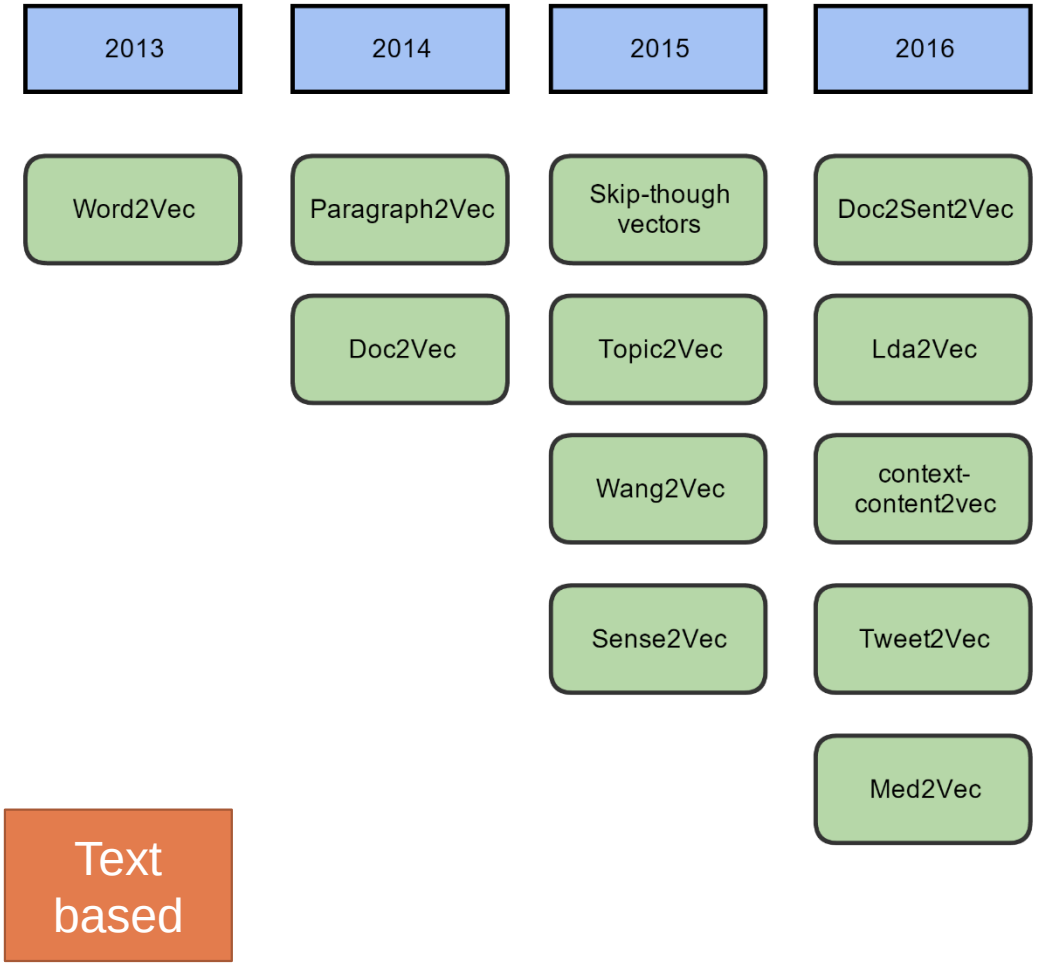
L3: NER as improvements for a Recommender System proposed

L4: E-commerce -> Enterprise

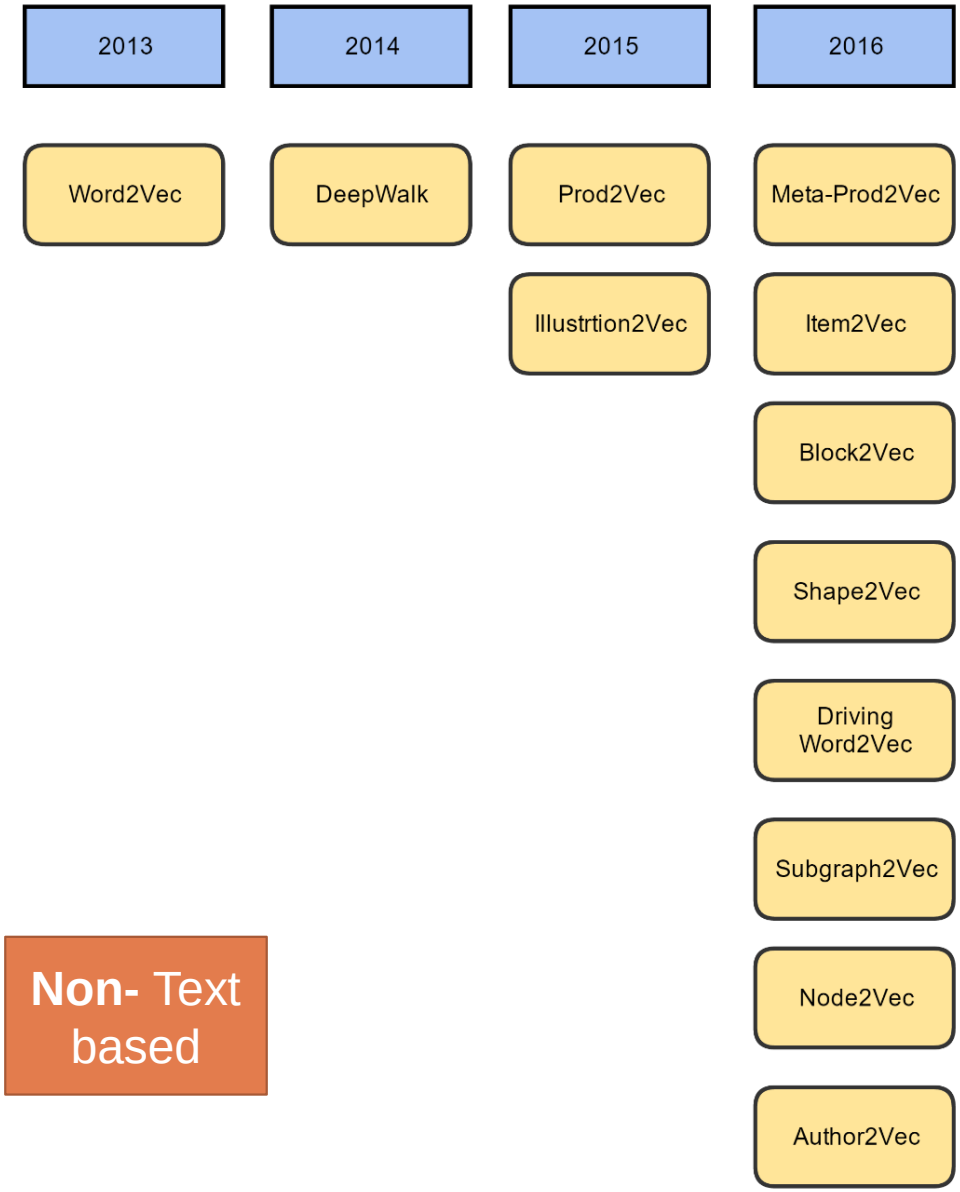


Outline

Evaluation - Further Developments of Word2Vec



Evaluation - Further Developments of Word2Vec



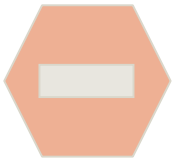
Outline

Lessons Learned

Research with Word Embeddings distributed over several Research Areas

Transferability of the methods and techniques

Elevated utilization of Word Embeddings for NLP and non-NLP tasks



Research mostly is too task or domain specific

Outline

1. Motivation
2. Data Acquisition
3. Research Questions
4. Layer Model
5. Utilization of the Layer Model
7. Further Developments
8. Lessons Learned
9. Future Work

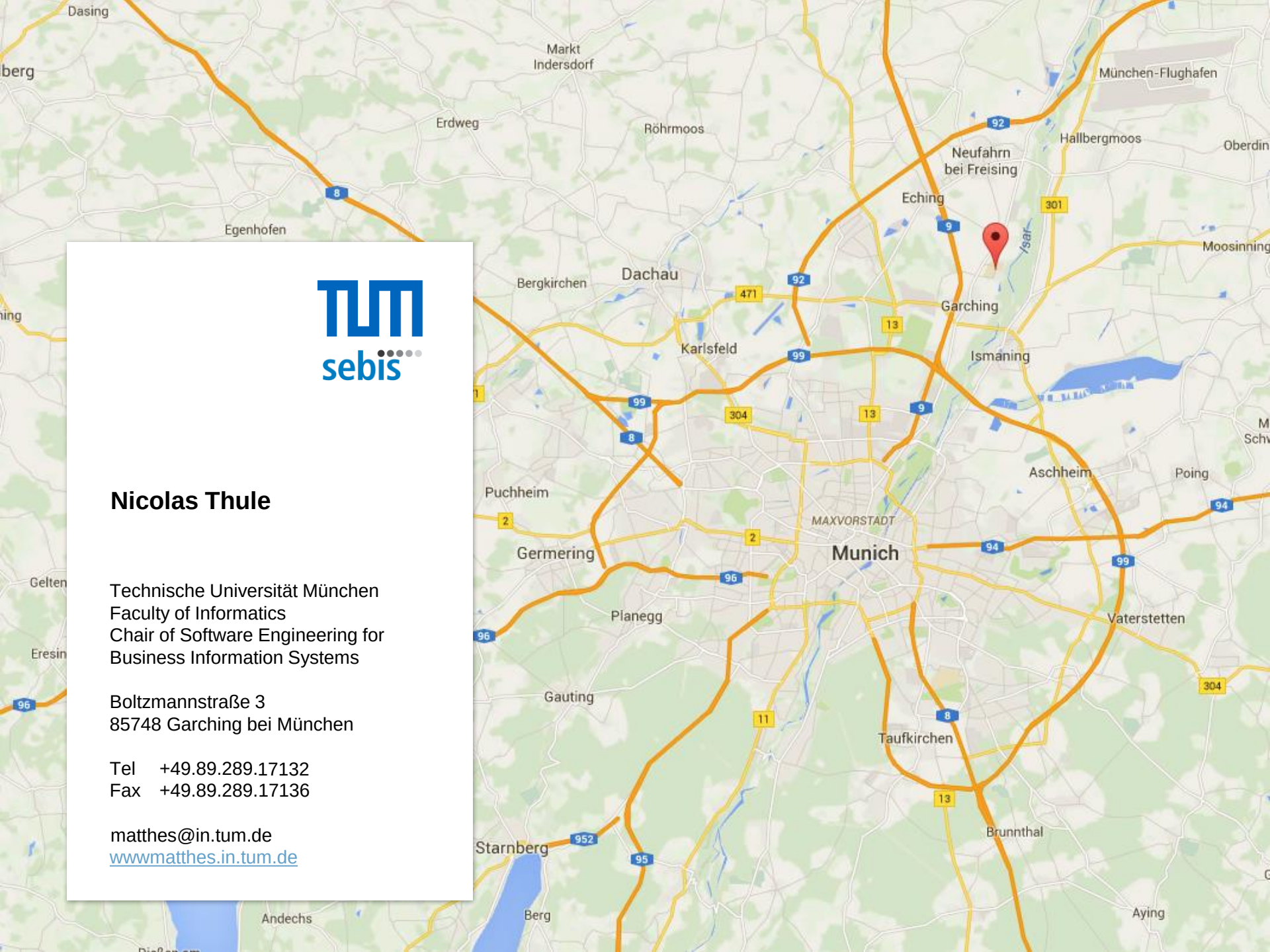
Future Work

Model related

- Extend Layers
- Refine Descriptions
- ….

Literature related

- Most cited lists
- Further statistics on the snapshot and general literature
- ….



Nicolas Thule

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

matthes@in.tum.de
wwwmatthes.in.tum.de

Sources

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

http://www.harzing.com/resources/publish-or-perish

Joshi, M., Hart, E., Vogel, M., and Ruvini, J.-D. (2015). Distributed word representations improve ner for e-commerce. In Proceedings of NAACL-HLT, pages 160–167.

[Thule2017] Thule, N. (2017). Literature Study: Use Case Analysis for Word Embeddings.