Word2Vec zur automatisierten Erstellung und Erweiterung von Thesauri

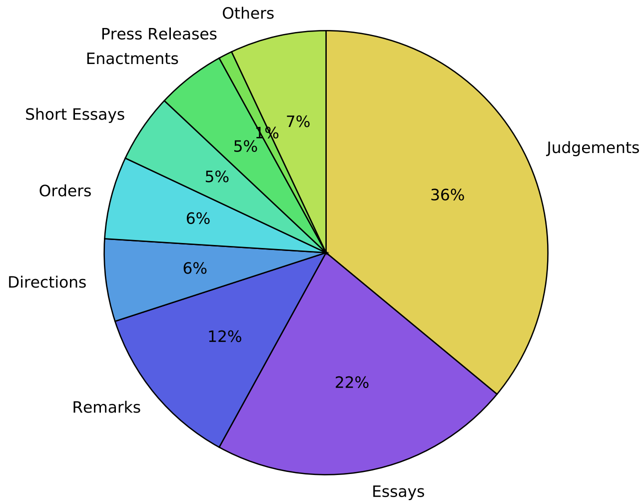
Jörg Landthaler, 9.3.2017, Munich

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

Dataset DATEV eG

Corpus Related to German Tax Law:

- ~130.000 documents
- Different document types
- ~150 million Tokens



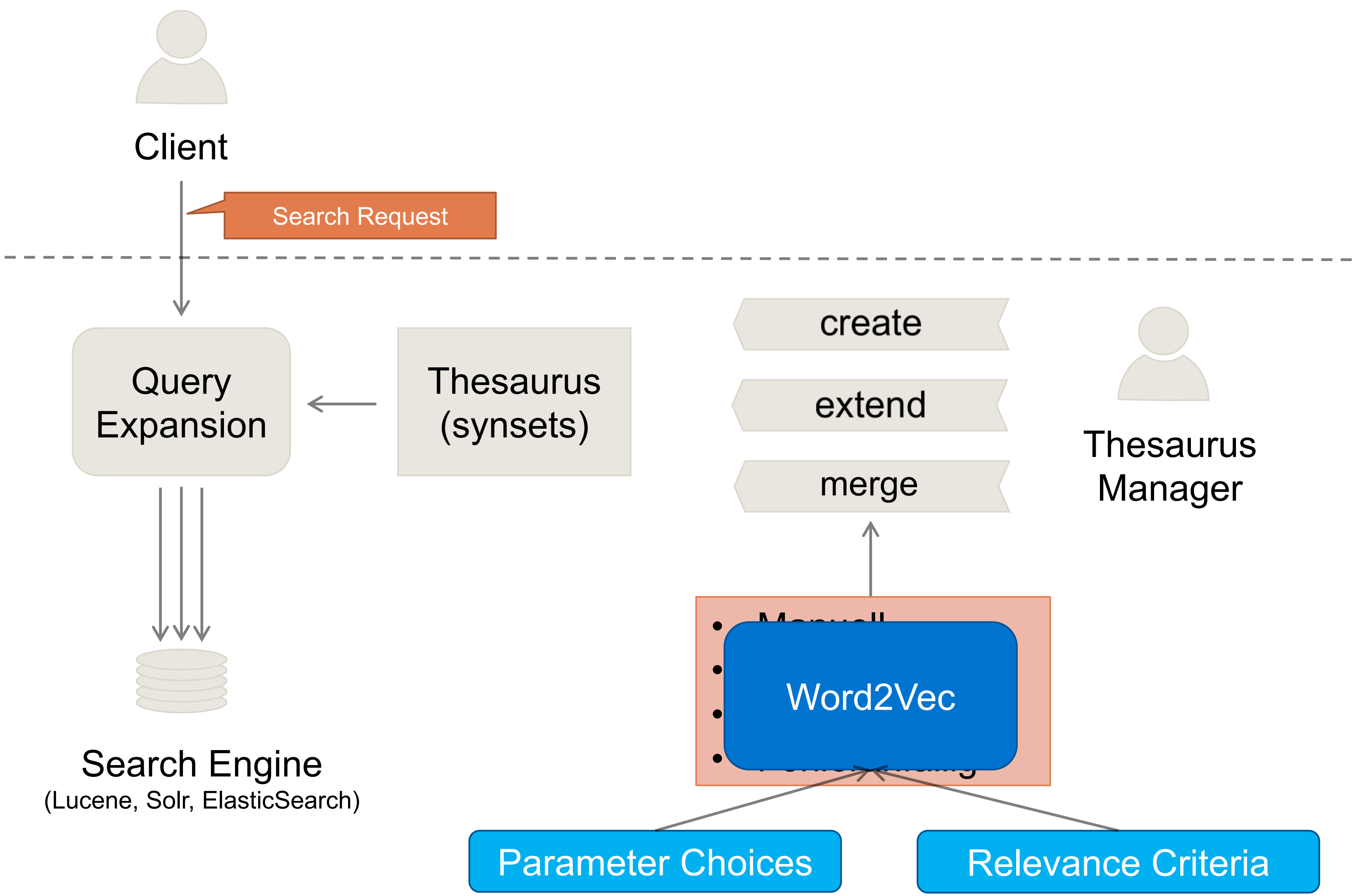
Thesaurus:

- ~16.000 concepts (12.000 synsets)
- ~ 36.000 synonyms
- We use subsets of synsets with words occurring at least N times (each word in synset)

N	Synsets	Terms	Terms/Group	Relations
250	275	622	2.26	932
500	158	358	2.26	542
750	112	260	2.32	420
1000	88	203	2.30	320

Example Synset: { 'fahrzeuge', 'gebrauchtfahrzeug', 'dienstwagen', 'pkw', 'firmenfahrzeug', 'fahrzeug‘ }

Thesauri for Information Retrieval



Word2Vec: A brief historical summary

The Distributional Hypothesis was introduced in 1954.

Harris, **1954**: Distributional Structure

Neural Probabilistic (Natural) Language Models are an old idea...

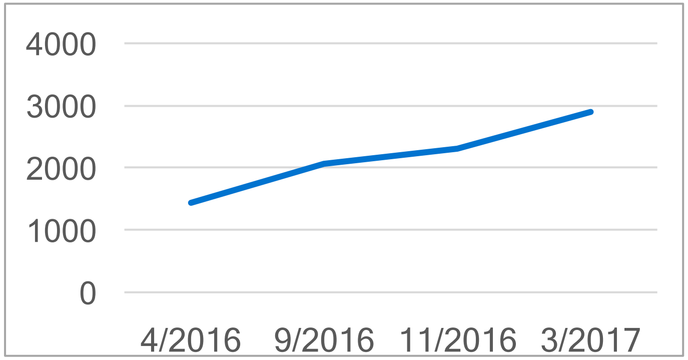
Hinton et al., **1986**: Learning distributed representations of concepts,

Bengio et al., **2003**: A Neural Probabilistic Language Model

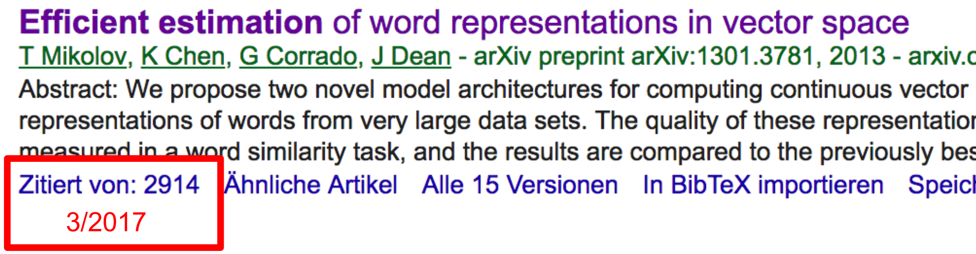
..., but now gain a lot of traction due to new and efficient algorithms!

Mikolov et al. **2013**: Efficient Estimation of word representations in vector space

And are a current trend in Natural Language Processing!



Cites on Mikolov, 2013 on Google Scholar



Harris, Zellig S. "Distributional structure." *Papers in structural and transformational linguistics*. Springer Netherlands, 1970. 775-794.
Hinton, Geoffrey E. "Learning distributed representations of concepts." *Proceedings of the eighth annual conference of the cognitive science society*. Vol. 1. 1986.
Bengio, Yoshua, et al. "A neural probabilistic language model." *Journal of machine learning research* 3.Feb (2003): 1137-1155.
Mikolov et al 2013*: Efficient estimation of word representations in vector space*

NLP: Feature Transformationen

Traditional NLP

Vocabulary

John	learns	to	read	fast
1	0	0	0	0
0	1	0	0	0
0	0	1	0	0
0	0	0	1	0
0	0	0	0	1

Sparse „one-hot“ representation

Advantages:
- Simple to calculate
- Often good results

Disadvantages:
- Bag-of-Words assumption (ignores spatial order)
- Sparse vectors

Word Embeddings

Manually chosen size

John	learns	to	read	fast
2,34	4,87	1.01	8,34	3,22
-1,30	3,22	0.01	0.23	5.67
0.23	-1.0	2.44	2.34	0.01
1.33	3.9	3.84	1.04	-1.2
1.0	3.8	-2.08	4,55	2,66

Dense representation

Advantages:
- Fast, unsupervised training
- Co-occurrence frequency encoded

Disadvantages:
- Lack of (direct) quality measures
- Difficult to understand for humans
(numbers without known semantics)

Characteristics of Word Embeddings

Simple mathematical operations, e.g. addition and substraction lead to interesting results:

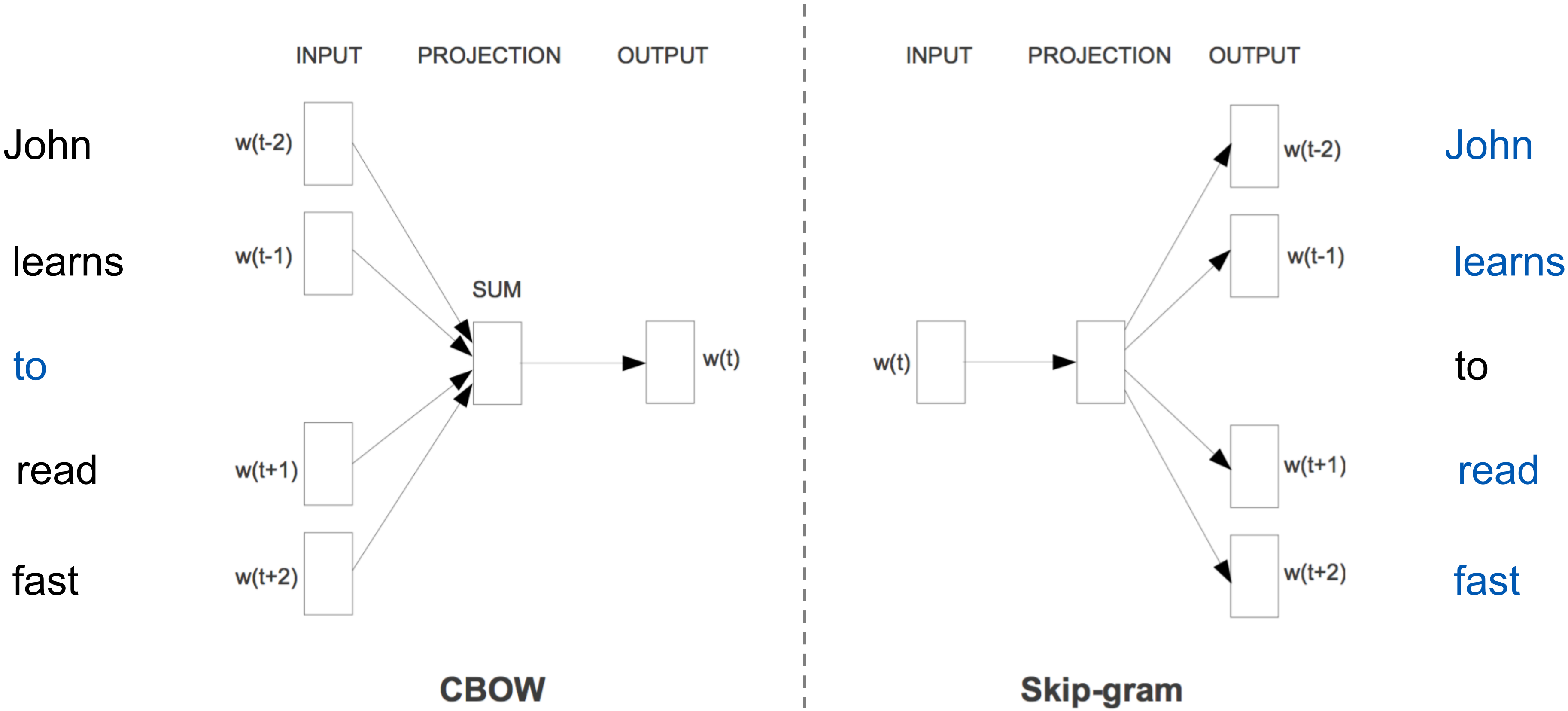
Clostest Vec w.r.t cosine similarity

Vec(„King“) – Vec („Man“) + Vec(„Woman“) -> Vec(„Queen“)

results in a vector close to the Vec (“Queen“) (w.r.t. cosine similarity e.g.)

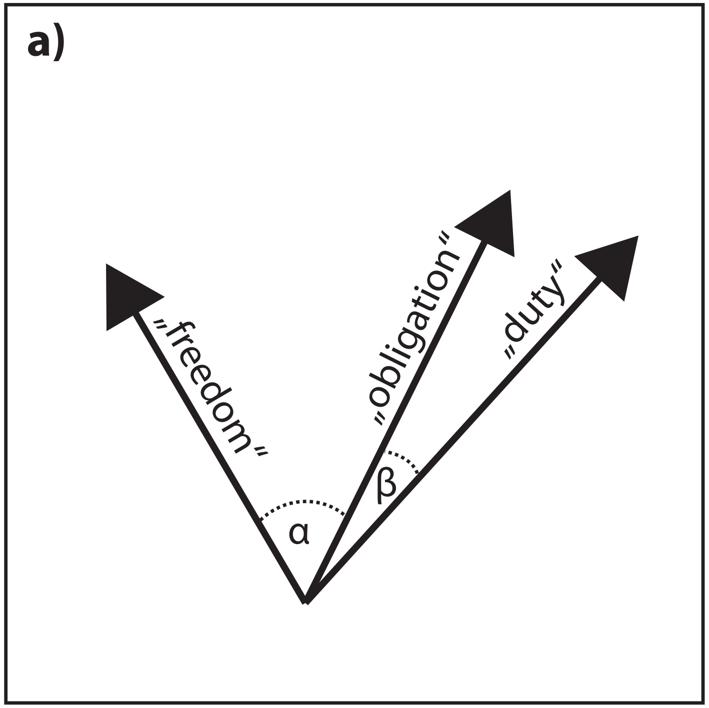
Mikolov et al. 2013, Distributed representations of words and phrases and their compositionality

A clever trick: “Unsupervised Classification”



Mikolov et al 2013: *Efficient estimation of word representations in vector space*

Characteristics of Word Embeddings



Semantically „similar“ words share a smaller angle w.r.t. a similarity measure than unrelated words.

Word2Vec: Mikolov et al 2013*: Efficient estimation of word representations in vector space*

Example result lists for individual words

pkw

1) kfz (0.778566)
2) pkws (0.754268)
3) fahrzeu
4) kraftfa
5) firmen
6) fahrze
7) wagen
8) firmenfa
9) fahrzeuge (0.6438
10) personenkraftwage
11) porsche (0.628974)
12) dienstwagen (0.627725)
13) autos (0.625358)
14) bmw (0.623232)
15) fahrzeug (0.618870)

Ranking-Position (RP): 12

umweltprämie

1) abwrackprämie (0.426274)
2) abwrackhilfe (0.415489)

71594)

testmiete (0.336994)
grenzentlastung (0.333919)
11) serienhauses (0.333742)
12) verbauchsbesteuerung (0.332857)
13) produktionsunabhangigen (0.331591)
14) konjunkturzulage (0.330321)
15) inhalteanbietern (0.327988)

Problems:

- How to measure quality of ranking lists?
- How to improve quality of ranking lists?

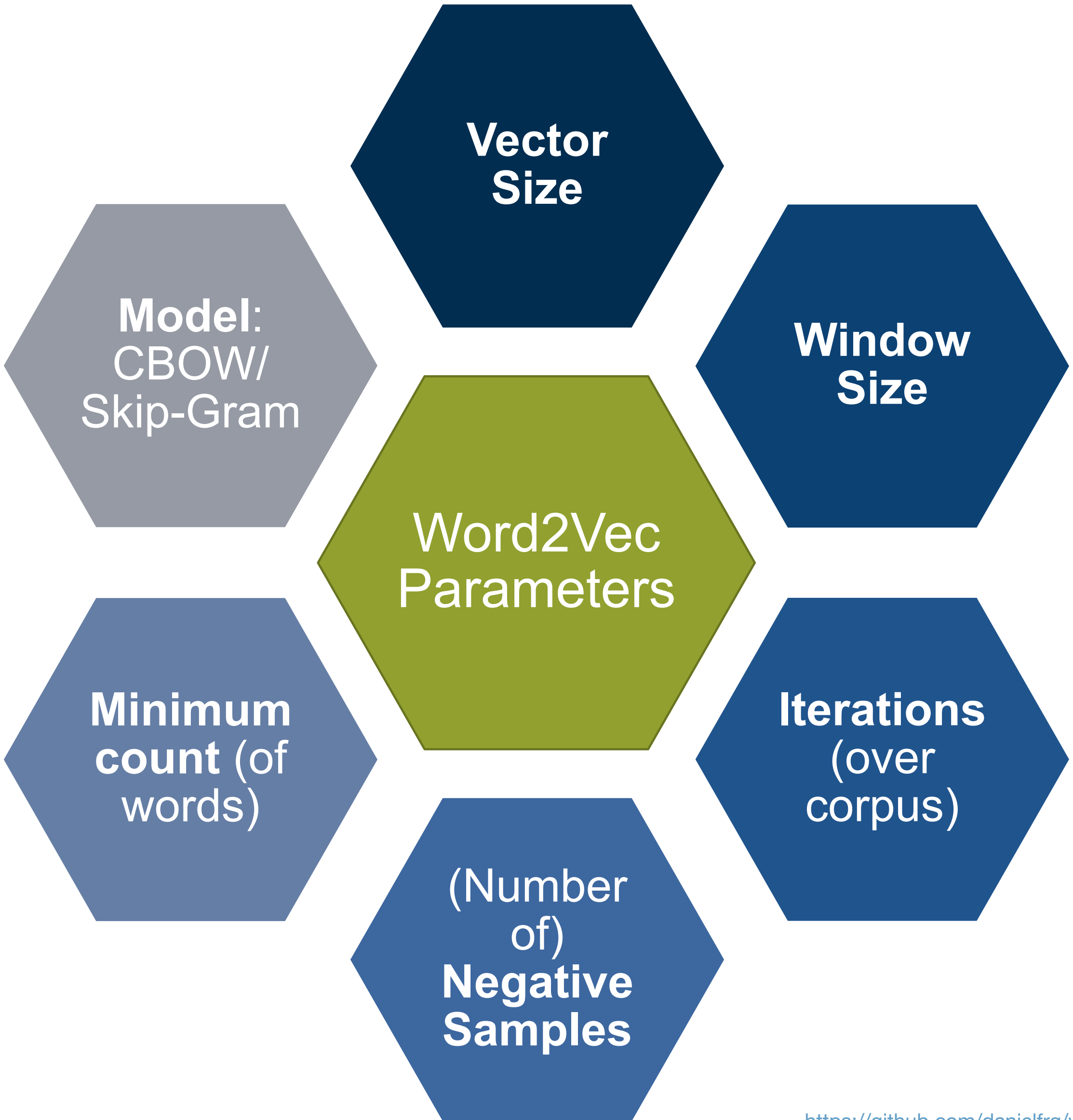
An implicit quality measure for Word2Vec

RP-Score: Average ranking position of word pairs for given thesaurus.

		synset1		synset2		synset3		synset4			
Häufigkeit		**umweltpramie**	**abwrackpramie**	**bfh**	**bundesfinanzhof**	**ao**	**abgabenordnung**	**bank**	**geldinstitut**	**kreditinstitut**	**kreditanstalt**
19	**umweltpramie**	1	13851	405249	406939	399476	457489	346357	399708	399396	195716
15	**abwrackpramie**	33	1	364758	318489	431965	451863	128183	294940	206033	205234
729858	**bfh**	190688	171407	1	39	19017	56230	508397	440417	247016	534566
37981	**bundesfinanzhof**	241059	131777	8	1	201645	131140	409572	486169	247839	508755
193527	**ao**	153226	304444	28569	140415	1	26	451847	322382	178515	245277
61569	**abgabenordnung**	423108	422205	85853	123266	3	1	298294	182632	319463	90879
24588	**bank**	167398	40531	505011	443437	498257	362275	1	2775	86	1783
324	**geldinstitut**	223928	105345	449207	495341	399026	194346	326	1	2	11594
6189	**kreditinstitut**	229393	51140	289010	257827	255902	322890	28	7	1	2932
299	**kreditanstalt**	27920	132629	530303	516554	407219	254721	333	53194	3643	1

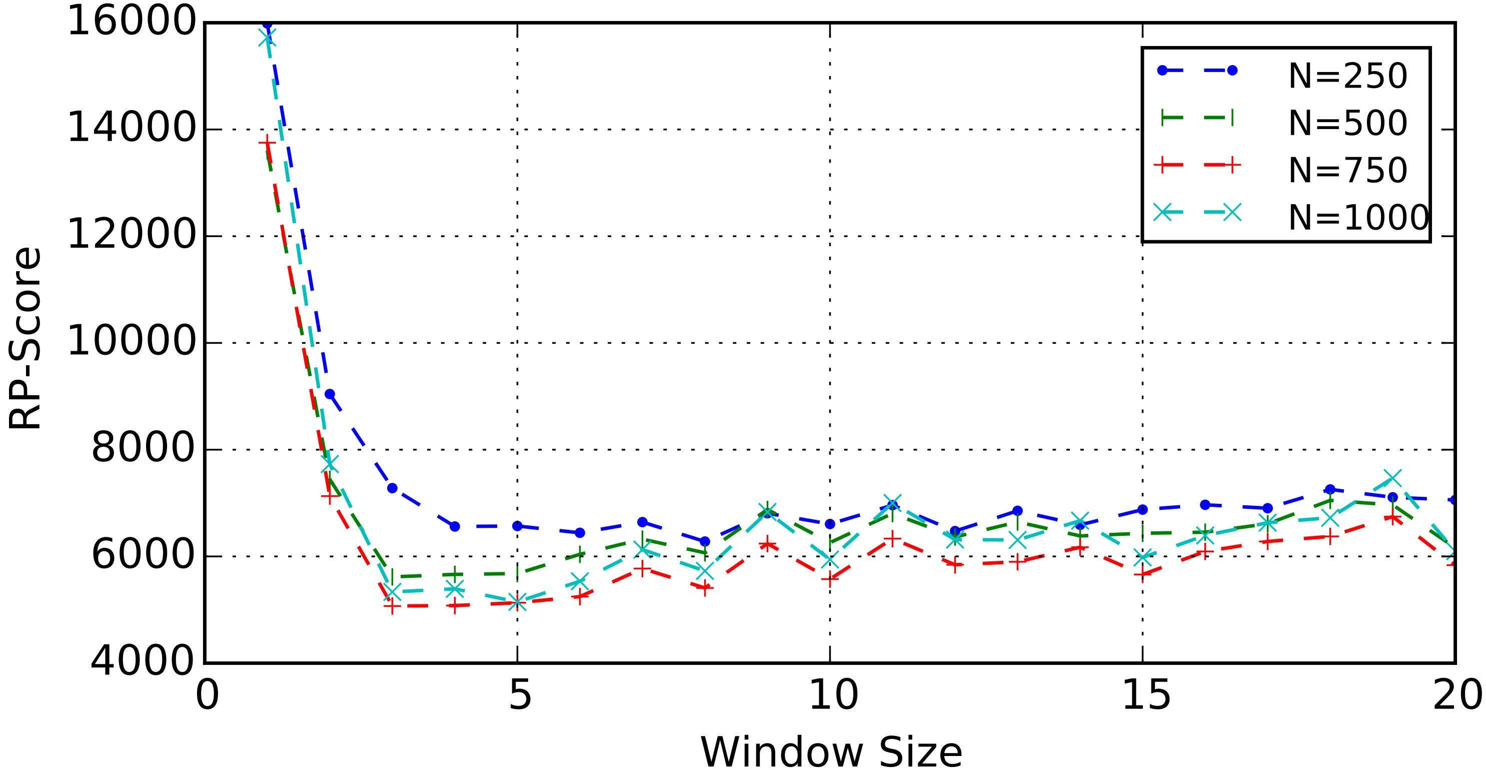
Landthaler et al., ICAIL 2017, submitted

Influencing Parameters on Word2Vec Outcome Quality



https://github.com/danielfrg/word2vec, accessed on 20.12.2016

Parameter Choice for Word2Vec: Example "Window Size"



Landthaler et al., ICAIL 2017, submitted

Example result lists for different parameters

Umweltpramie
Parameter Iterations = 19

1) **umweltpramie (0.426274)**
2) **abwrackhilfe (0.415489)**
3) abgabeordnung (0.377226)
4) **pramie (0.376614)**
5)
6)
7)
8)
9)
10)
11) atelier (0.333742)
12) archivraum (0.332857
13) stahlradiator (0.33159
14) wartefrist (0.330321)
15) anschrift (0.327988)

Umweltpramie
Parameter Iterations = 20

1) **abwrackhilfe (0.426274)**
2) **abwrackpramie (0.415489)**
3) architekturkopien (0.377226)
4) zuordnungsentgelts (0.376614)
5) 94)
11) serienhauses (0.333742)
12) erung (0.332857)
13) angigen (0.331591)
14) (0.330321)
15) .327988)

Problem:

- How to select only “relevant” terms?

(previously known solutions: fixed length lists, threshold)

Solution: Intersections

1) **umweltpramie**
2) **abwrackhilfe**
3) **pramie**

Landthaler et al., ICAIL 2017, submitted

Conclusion & Outlook

Conclusion

- Word2Vec can be used to calculate **useful suggestions** for Thesauri Managers
- The given Corpus & Thesaurus from Datev enabled us to **determine good parameters** for Word2Vec / to assess how well Word2Vec is suited to detect synsets automatically (on German texts)
- **Intersections** of syonym lists lead to stable synsets (irrelevant words are removed)

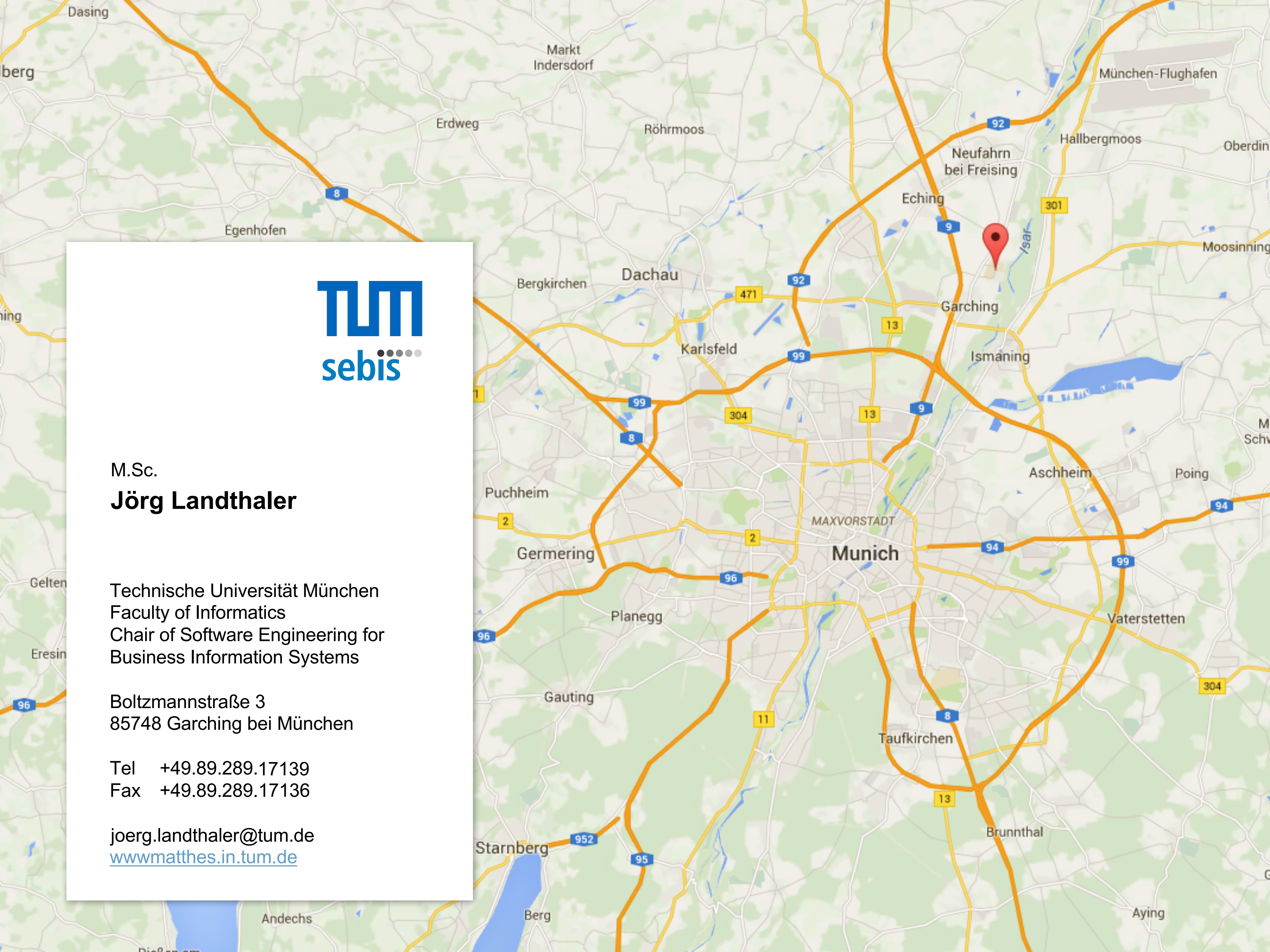
Outlook

Open questions w.r.t. Word2Vec for Thesauri are:

- How can **initial synsets** or **initial key terms** be identified? (building a thesaurus from scratch)
- Can Word2Vec models be merged?
- What is the effect of **multi-lingual** corpuses?

Another interesting topic is: „Can Word Embeddings be used for **Summarization** and **Tagging** tasks?“

If you are interested in these topics, ask me later at Stammtisch!



M.Sc.
Jörg Landthaler

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

Boltzmannstraße 3
85748 Garching bei München

Tel +49.89.289.17139
Fax +49.89.289.17136

joerg.landthaler@tum.de
wwwmatthes.in.tum.de