TECHNISCHE UNIVERSITÄT DRESDEN

Dresden Database Systems Group

Word-Embedding-based Patent Retrieval

Großer Beleg

Hendrik Cziommer

Dresden, 13.05.2019

Background

- Interface:projects is Dresden based company providing search solutions

- *Deutsches Patent- und Markenamt* (**DPMA**) is a current client

Patents

Dresden Database Systems Group

Patent Application

Die Erfindung betrifft ein **elektronisch** gesteuertes **Federungssystem** für ein **Fahrrad** (1), enthaltend zumindest einem **Federelement** (3, 4), welches zwischen einem ersten Teil (10) des **Fahrrades** und einem zweiten Teil (14, 15) des **Fahrrades** (1) angeordnet ist, welche beweglich miteinander **verbunden** sind, wobei zumindest eine Kenngröße des **Federelementes** veränderbar ist, und zumindest einen Aktor (431), welcher auf das **Federelement** (3, 4) einwirkt, um die zumindest eine Kenngröße zu verändern, und ein **Elektronikmodul** (6), mit welchem ein Ansteuersignal für den zumindest einen Aktor (431) erzeugbar ist, wobei weiterhin ein **Steuerelement** (2, 63, 66) vorhanden ist, mit welchem das vom **Elektronikmodul** (6) erzeugte Ansteuersignal beeinflussbar ist, wobei das **Steuerelement** (2, 63) mit dem **Elektronikmodul** (6) über ein **Funksignal** (64) verbindbar ist und/oder der Aktor (431) mit dem **Elektronikmodul** (6) über ein Funksignal (64) verbindbar ist. Weiterhin betrifft die **Erfindung** ein entsprechendes Verfahren zur **Steuerung** eines **Federungssystems** für ein **Fahrrad** und ein Computerprogramm zu dessen Durchführung. [...]

Patents

Dresden Database Systems Group

Transformation into a Query

(**elektronisch** OR **elektronik**) AND (**fahrrad** OR **elektrorad** OR **e-bike** OR **pedelec** OR **zweirad**) AND(**feder*** OR **dämpf***) AND **steuer*** AND (***modul** OR ***element**) AND **pub < 2015** …

Patents

Dresden Database Systems Group

Transformation into a Query

Incomplete list of synonyms
(e.g. „elektrisch" is missing)

(**elektronisch** OR **elektronik**) AND (**fahrrad** OR **elektrorad** OR **e-bike** OR **pedelec** OR **zweirad**) AND(**feder*** OR **dämpf***) AND **steuer*** AND (***modul** OR ***element**) AND **pub < 2015** …

- Federbett
- Federhalter
- Federkleid
- …

- Steuererklärung
- Steuerberater
- Steuermann
- …

Patents

Problems

- Long processing time (up to hours for one query)
- Keyword extraction needs a lot of experience
- Query formulation is tedious
- Incomplete (manual created) Synonym list
- Wildcards introduce a lot of irrelevant documents

→ Low precision and low recall

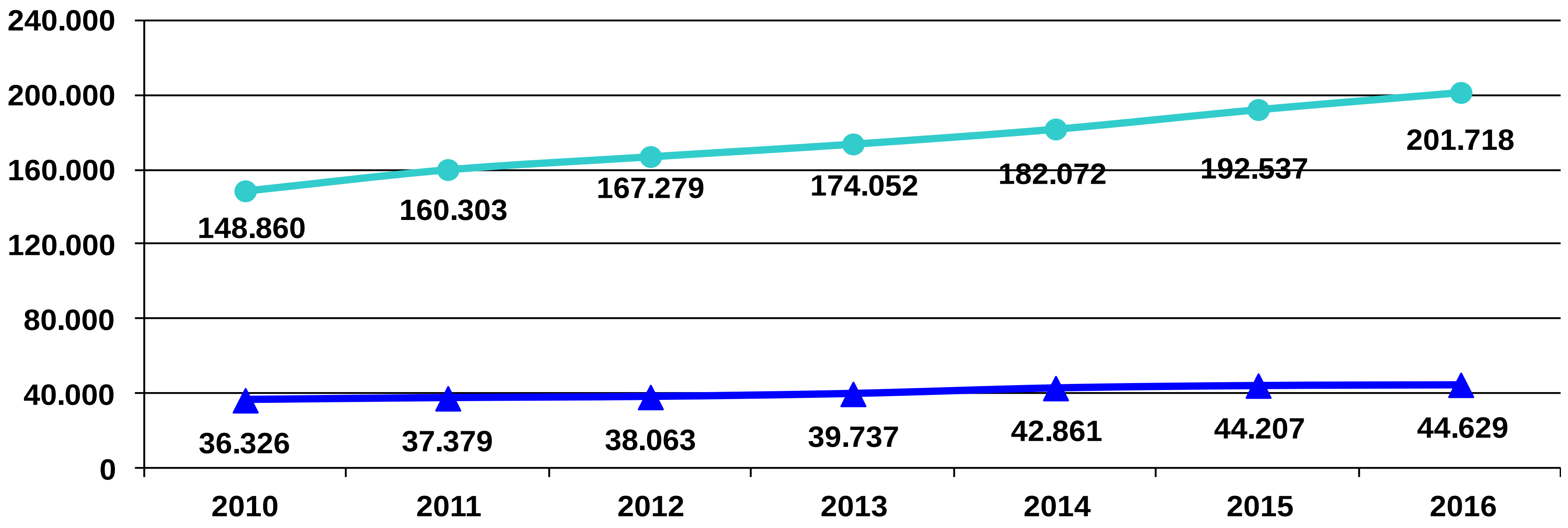
→Patent examiners have to look at a lot of documents

Patents


```
(elektronisch OR elektrisch OR elektronik) AND

(fahrrad OR elektrorad OR e-bike OR pedelec OR zweirad) AND

(feder* OR dämpf*) AND steuer* AND (*modul OR *element) AND

 ipc:B62M AND pub < 2015 …
```

Patents

Dresden Database Systems Group



pending Applications
(am Jahresende in den Patentabteilungen noch nicht abgeschlossene Prüfungsverfahren)

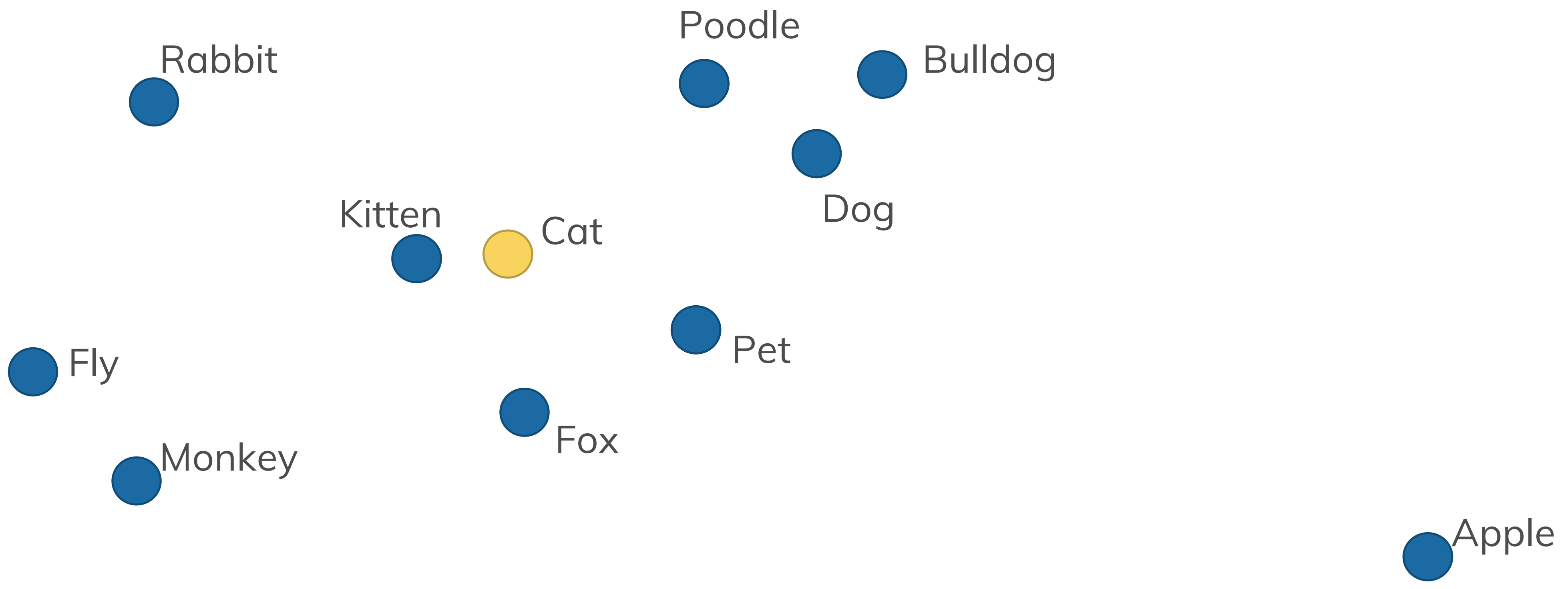
wirksame Prüfungsanträge und sonstige Zugänge
(insbesondere Zurückverweisungen vom BPatG, Abhilfen von Beschwerden, Wiedereinsetzungen)

Stand: Februar 2017

Word and Document Embeddings

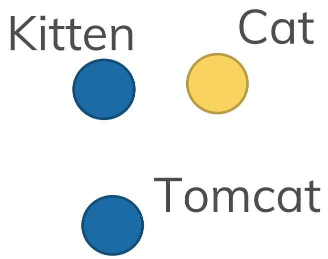
Word Embeddings

- Vector representation of words
- Capture word semantics

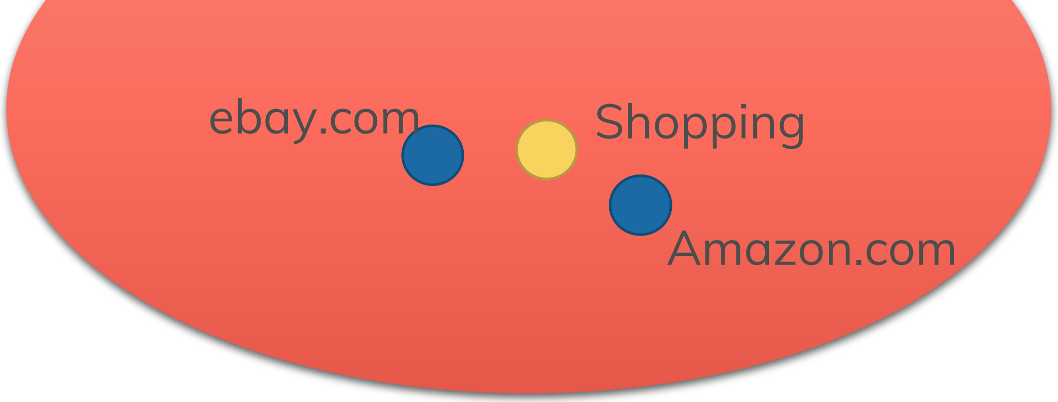


Use Cases

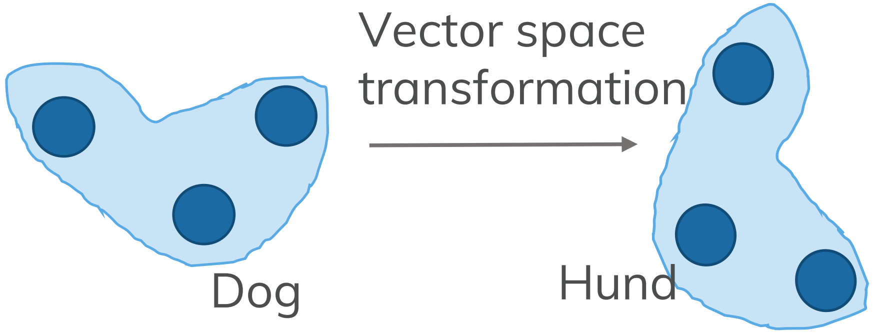
Synonyms



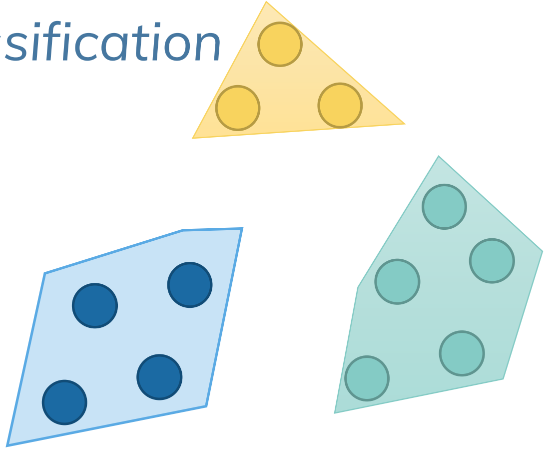
Information Retrieval



Translation



Classification



Training

- Idea: Similar words appear in similar contexts

  - "The **cat** sleeps.", "My pet is a **cat**.", "The **cat** meows."
  - "The **dog** sleeps.", "My pet is a **dog**.", "The **dog** barks."
    → **cat** and **dog** are similar, but slightly different

  - "I eat an **apple**.", "The **apple** is red.", "An **apple** is a fruit."
    → cat and apple are different

Training

Dresden Database Systems Group

- Word2Vec (CBOW) (Mikolov et al. 2013):
  - Considers the context of each word (context window)
  - If context is similar → word is similar

- fastText (Bojanowski et al. 2017) extends Word2Vec:
  - Considers subwords of words by using n-grams (character level)
  - One word vector is the sum of the vectors of the subwords
  - Better for words with same stem (plurals, inflection) compared to Word2Vec

Training

Dresden Database Systems Group

Die Erfindung betrifft (1) ein elektronisches Fahrrad-Federungssystem.

↓ Pre-processing

erfindung	betrifft	elektronisch	fahrrad	federungssystem

↓

Training

↓

Model

$$V_{Erfind.} = \begin{pmatrix} V_0 \\ \vdots \\ V_{299} \end{pmatrix} \quad V_{betrif.} = \begin{pmatrix} V_0 \\ \vdots \\ V_{299} \end{pmatrix} \quad V_{elektr.} = \begin{pmatrix} V_0 \\ \vdots \\ V_{299} \end{pmatrix} \quad V_{fahrr.} = \begin{pmatrix} V_0 \\ \vdots \\ V_{299} \end{pmatrix} \quad V_{federungssy.} = \begin{pmatrix} V_0 \\ \vdots \\ V_{299} \end{pmatrix} \quad \ldots$$

Document Embeddings

Dresden Database Systems Group

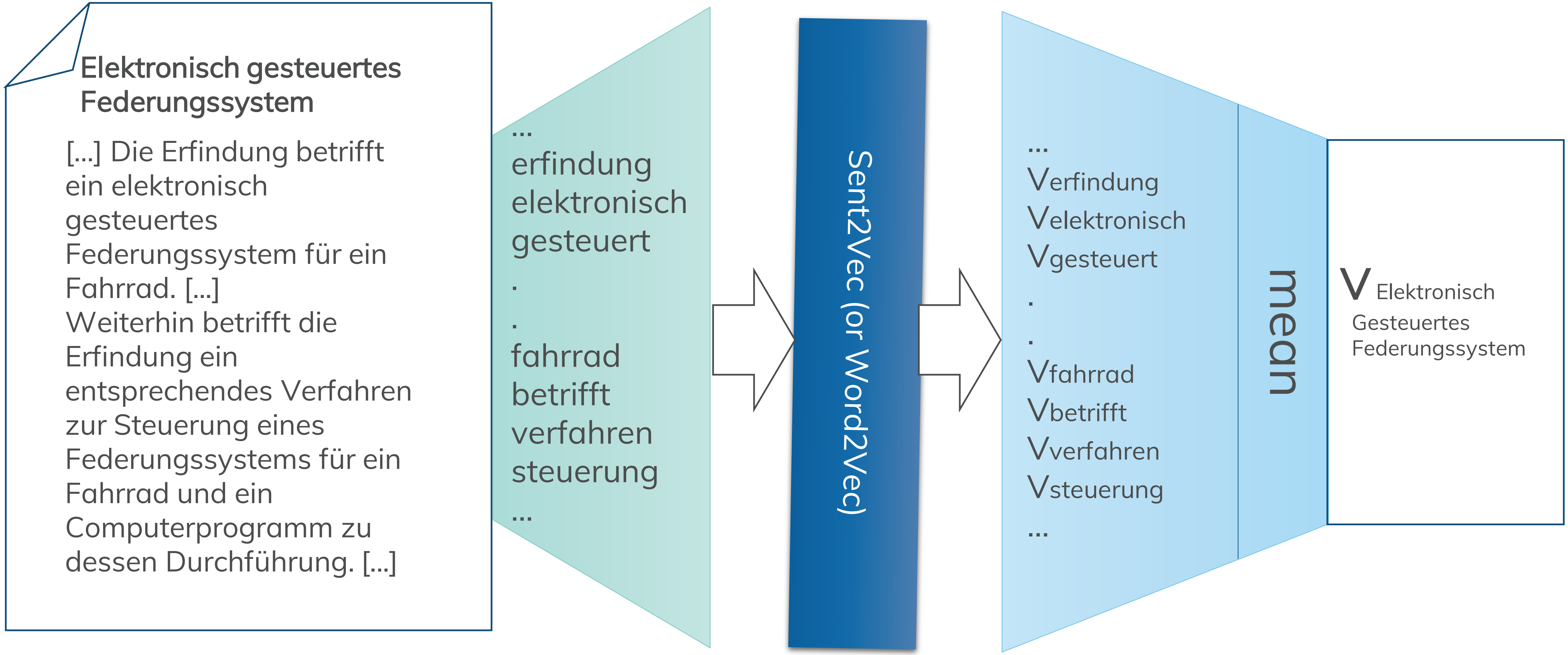
- Idea: Average of all Word Vectors in a Document
  - Outperforms other more complex approaches like Recurrent Neuronal Networks LSTM (Wieting et al.(2016b))
  - Sent2Vec(Pagliardini et al. 2017): Word Vectors are specifically optimized towards additive combination

Sent2Vec

Dresden Database Systems Group

- Word Vectors are specifically optimized towards additive combination
- Includes learning of n-grams (word level) embeddings by averaging
- Considers whole sentence as context opposed to word window

Document Embeddings

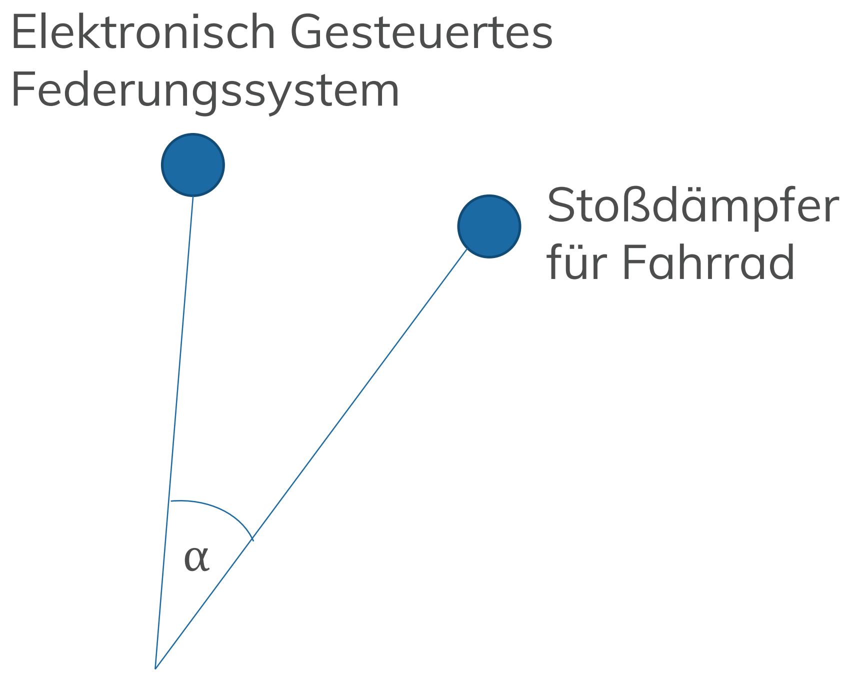


How to apply to Information Retrieval?

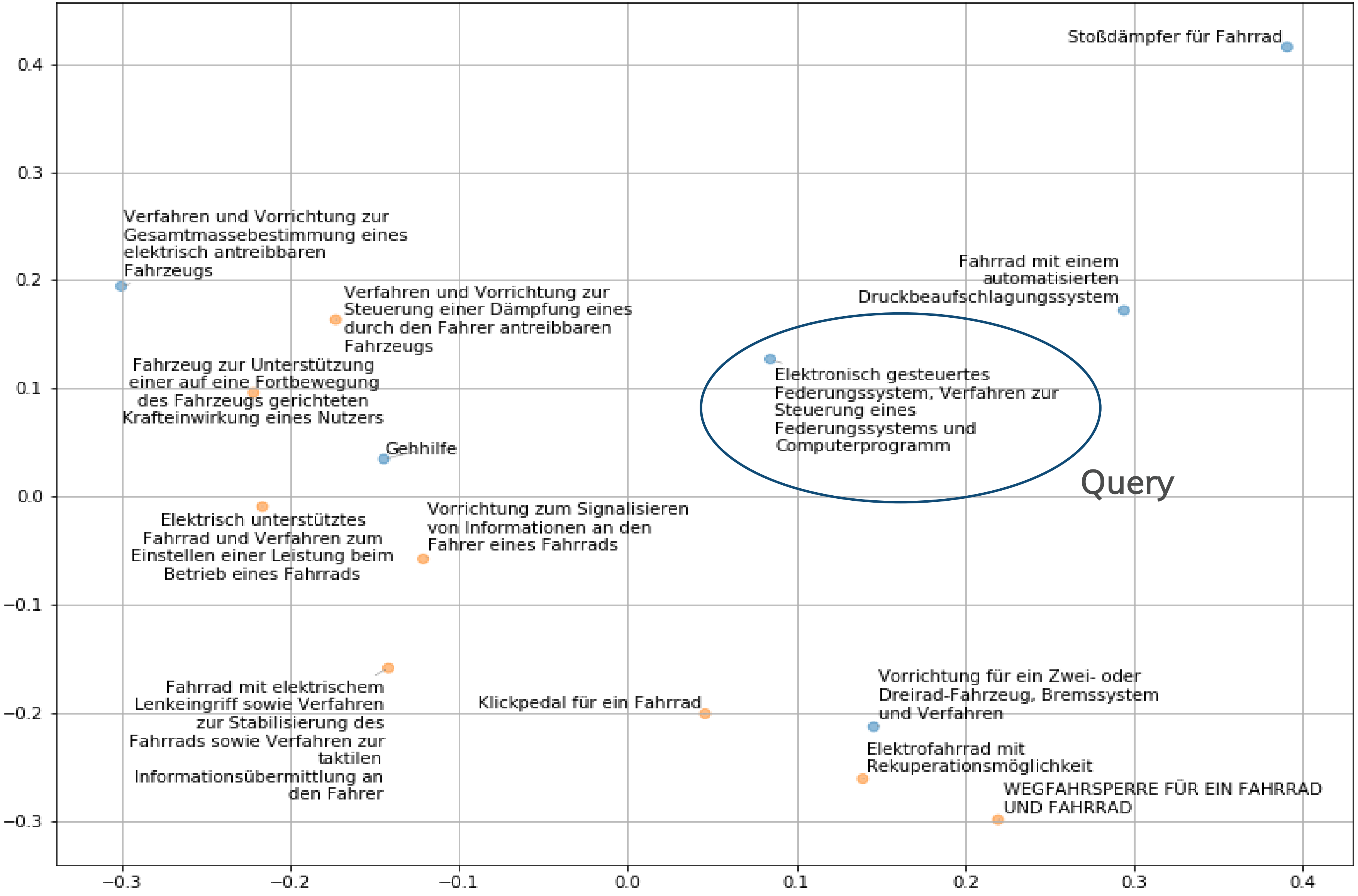
Evaluation – Similarity Score

Dresden Database Systems Group

- Determine cosine similarity: $\cos(\alpha)$
- $\cos(\alpha) = \frac{a \cdot b}{\|a\| \cdot \|b\|}$
- $\|a\| = \sqrt{{a_1}^2 + \cdots + {a_n}^2}$



Example



Search in (300-dimensional) original space

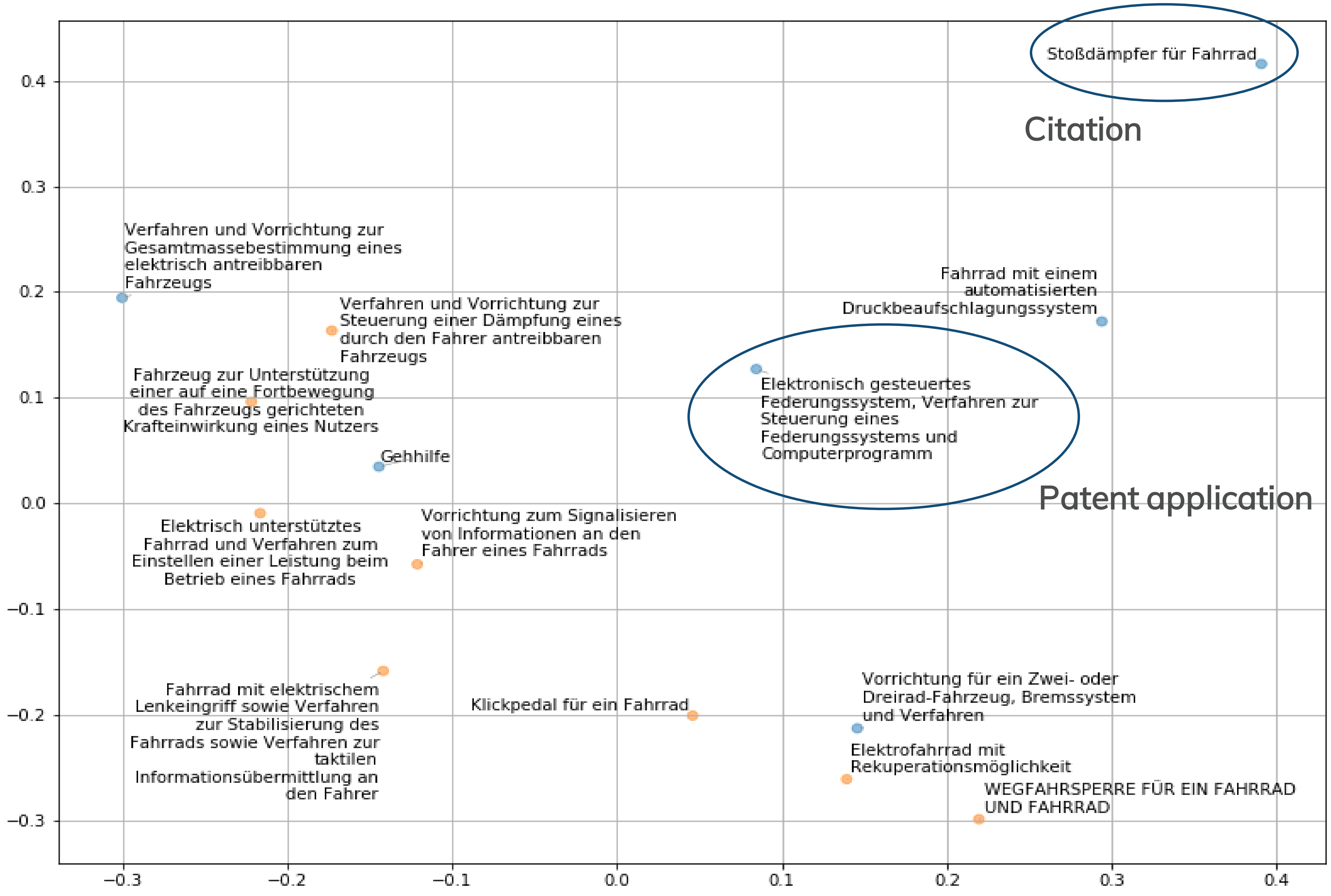
Results

Setup

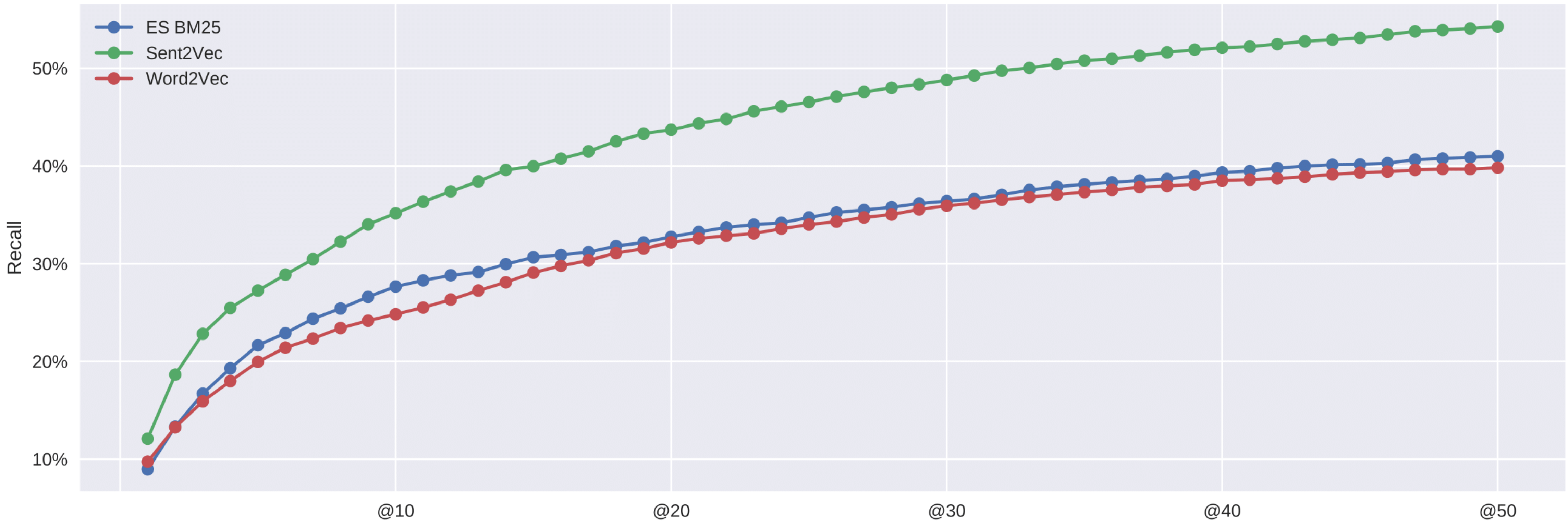
- Test to improved Recall
- Using "Citations" which are similar documents marked by patent experts
  → relevant document

- Methods
  - Elastic Search’s BM25 implementation shorthanded by ES BM25
  - Document Embeddings with Word2Vec
  - Document Embeddings with Sent2Vec

Example


Stoßdämpfer für Fahrrad
Citation
Verfahren und Vorrichtung zur Gesamtmassebestimmung eines elektrisch antreibbaren Fahrzeugs
Fahrrad mit einem automatisierten Druckbeaufschlagungssystem
Verfahren und Vorrichtung zur Steuerung einer Dämpfung eines durch den Fahrer antreibbaren Fahrzeugs
Fahrzeug zur Unterstützung einer auf eine Fortbewegung des Fahrzeugs gerichteten Krafteinwirkung eines Nutzers
Elektronisch gesteuertes Federungssystem, Verfahren zur Steuerung eines Federungssystems und Computerprogramm
Gehhilfe
Patent application
Elektrisch unterstütztes Fahrrad und Verfahren zum Einstellen einer Leistung beim Betrieb eines Fahrrads
Vorrichtung zum Signalisieren von Informationen an den Fahrer eines Fahrrads
Fahrrad mit elektrischem Lenkeingriff sowie Verfahren zur Stabilisierung des Fahrrads sowie Verfahren zur taktilen Informationsübermittlung an den Fahrer
Klickpedal für ein Fahrrad
Vorrichtung für ein Zwei- oder Dreirad-Fahrzeug, Bremssystem und Verfahren
Elektrofahrrad mit Rekuperationsmöglichkeit
WEGFAHRSPERRE FÜR EIN FAHRRAD UND FAHRRAD
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
0.4

Result

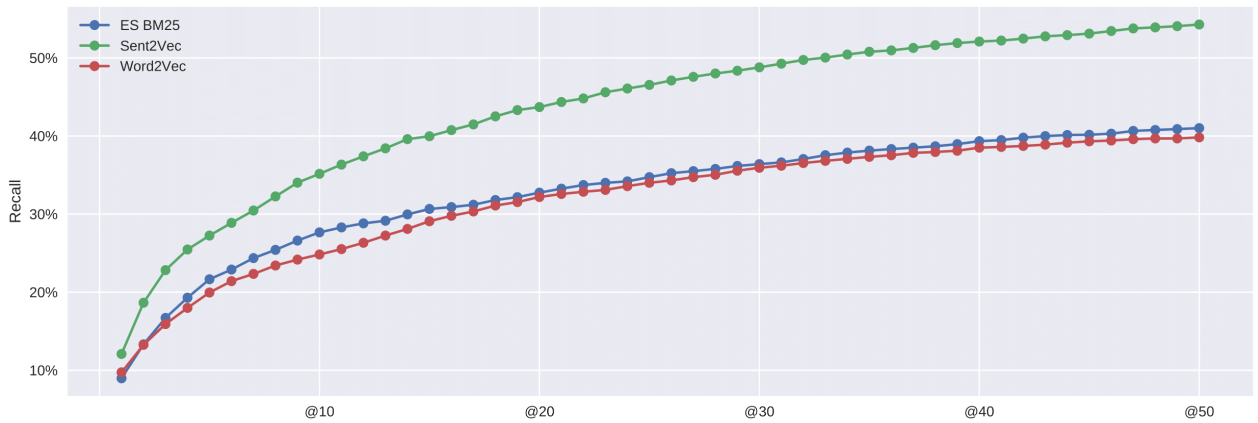


Result

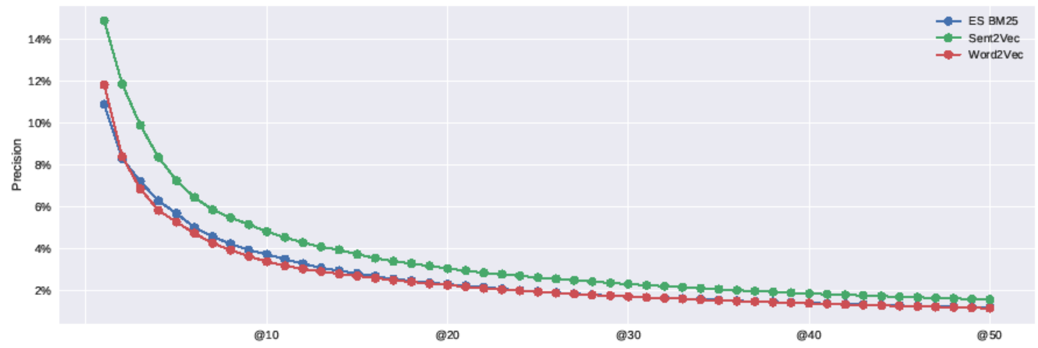
- Sent2Vec beats Elastic Search's BM25 implementation and Word2Vec

- ES BM25 is on par with Word2Vec

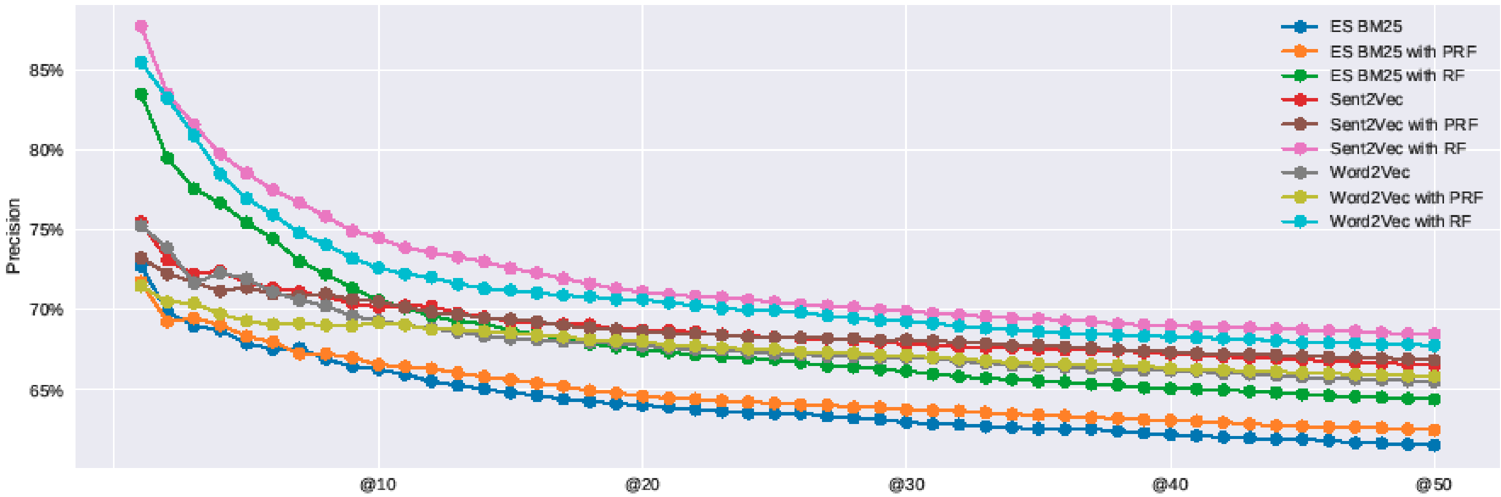
- Overall Recall is lower than excepted since most patents have only **one** citation



Result



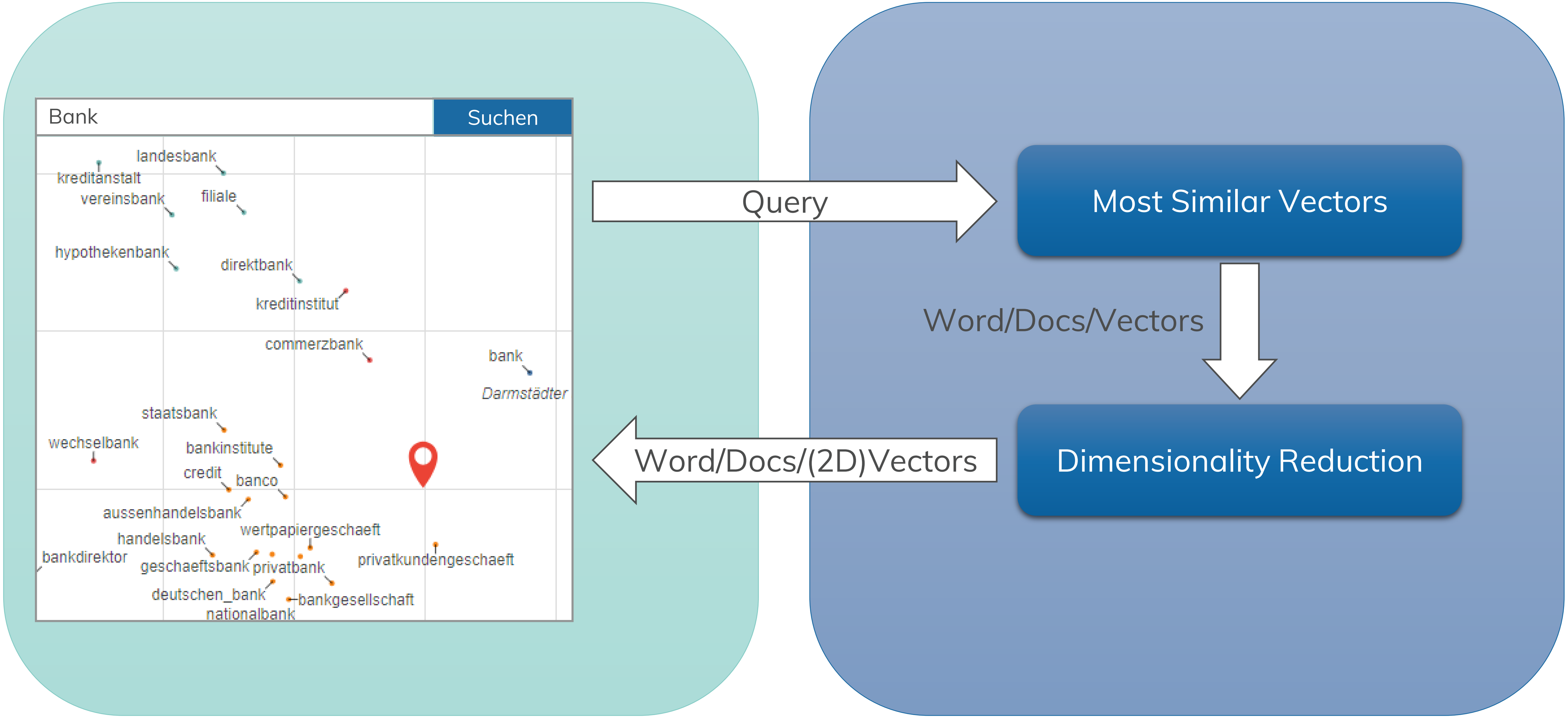
ES BM25
Sent2Vec
Word2Vec
Precision
14%
12%
10%
8%
6%
4%
2%
@10
@20
@30
@40
@50



ES BM25
ES BM25 with PRF
ES BM25 with RF
Sent2Vec
Sent2Vec with PRF
Sent2Vec with RF
Word2Vec
Word2Vec with PRF
Word2Vec with RF
Precision
85%
80%
75%
70%
65%
@10
@20
@30
@40
@50

Prototype

Prototype


Bank
Suchen
landesbank
kreditanstalt
vereinsbank
filiale
hypothekenbank
direktbank
kreditinstitut
commerzbank
bank
Darmstädter
staatsbank
wechselbank
bankinstitute
credit
banco
aussenhandelsbank
handelsbank
wertpapiergeschaeft
bankdirektor
geschaeftsbank
privatbank
privatkundengeschaeft
deutschen_bank
bankgesellschaft
nationalbank
Query
Most Similar Vectors
Word/Docs/Vectors
Dimensionality Reduction
Word/Docs/(2D)Vectors

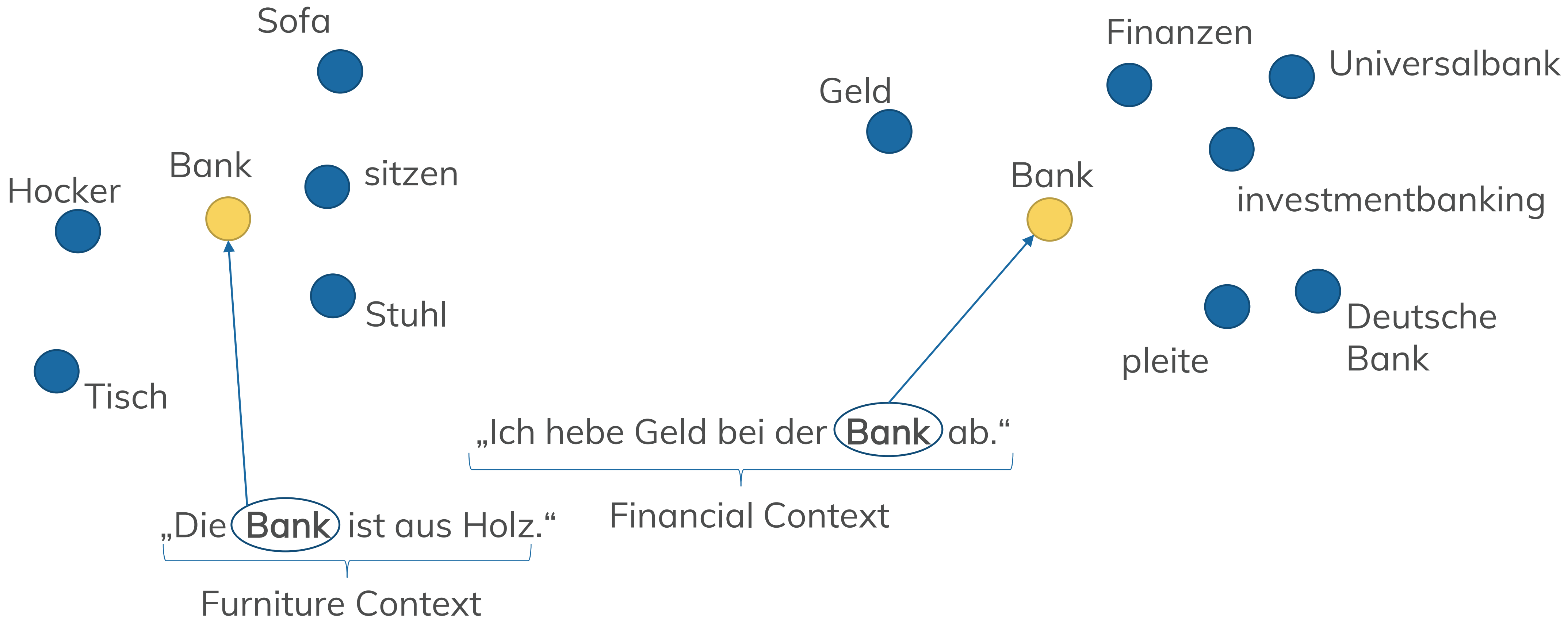
Conclusion

- For the domain a patents we find the following
  - Word2Vec based Document Embedding is on par with ES BM25
  - Sent2Vec performs better than ES BM25 and Word2Vec

- We presented a prototype

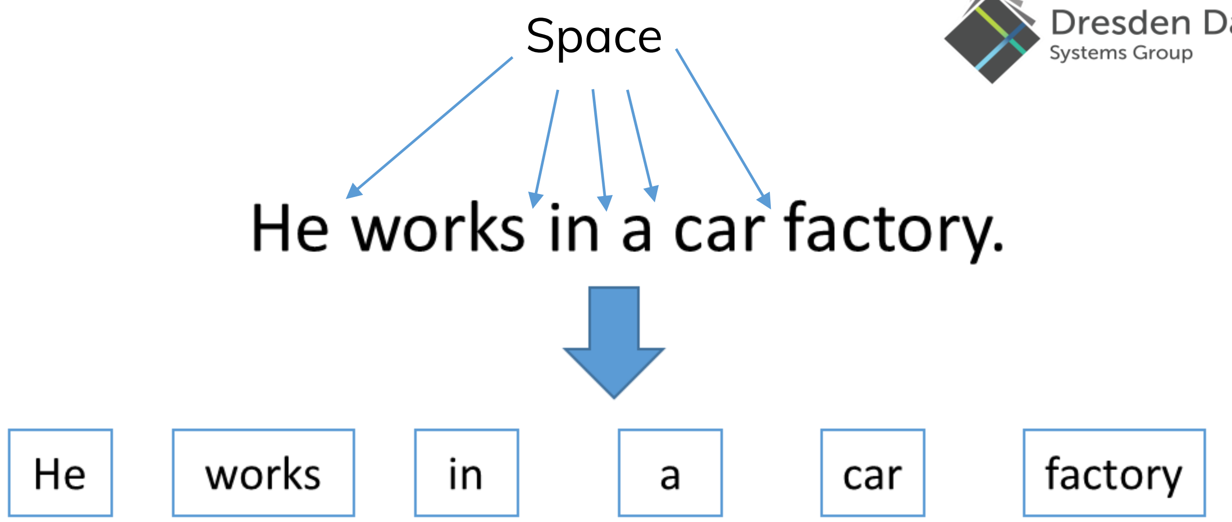
Outlook

- LASER and BERT are new approaches for embedding words based on their context (takes sentence as input)



Tokenization

- English:



Tokenization

- English:

He works in a car factory.

↓

He	works	in	a	car	factory

- Japanese / Chinese:

(called: Segmentation)

他在汽车工厂工作。

↓

他在汽车工厂工作

?

Topic

- Corpus pre-processing methods for Japanese and Chinese
- Training of Word Embeddings
- Evaluation of Word Embeddings Models

Ideas

- Segmenting Tools
- Improvement of Segmenting using Named Entity Recognition
- Transliteration from Japanese/Chinese to Latin script

Characteristics of Chinese and Japanese

Chinese

- Sentences consist of characters

$$\text{Tom}\underbrace{\text{想要}}_{\text{want}}\overbrace{\text{看}}^{\text{watch}}\underbrace{\text{电视。}}_{\text{TV}}$$

- Word ≠ character:

包	bāo	bag	bag
包 括	bāo kuò	bag + to enclose	to include
面 包	miàn bāo	face + bag	bread
手 提 包	shǒu tí bāo	hand + to hold + bag	handbag
打 包	dǎ bāo	to beat + bag	to pack

Chinese

200 words: yi

勚	yì		toil
翼	yì		wing
邑	yì		city
泄	yì	(adj.)	idle
饴	yí		syrup
椅	yǐ		chair
咿	yī		(onomat.)
漪	yī		ripple
猗	yī		(interj.)
洟	yí		snivel
埸	yì		border
帟	yì		canopy
熤	yì		(person)
饐	yì		rancid
圯	yí	(n.)	bridge
衣	yī		clothes gown / to dress / to wear
铱	yī		iridium
鷖	yī		widgeon
荑	yí		to weed

https://www.chinese-dictionary.org/

Chinese

200 words: yi

20 words: yī

勩	yì		toil
翼	yì		wing
邑	yì		city
泄	yì	(adj.)	idle
饴	yí		syrup
椅	yǐ		chair
咿	yī		(onomat.)
漪	yī		ripple
猗	yī		(interj.)
洟	yí		snivel
埸	yì		border
帟	yì		canopy
熠	yì		(person)
饐	yì		rancid
圯	yí	(n.)	bridge
衣	yī		clothes gown / to dress / to wear
铱	yī		iridium
鹥	yī		widgeon
荑	yí		to weed

https://www.chinese-dictionary.org/

Chinese

200 words: yi

20 words: yī

1 word: 衣(yī, clothes)

勩	yì		toil
翼	yì		wing
邑	yì		city
泄	yì	(adj.)	idle
饴	yí		syrup
椅	yǐ		chair
咿	yī		(onomat.)
漪	yī		ripple
猗	yī		(interj.)
洟	yí		snivel
埸	yì		border
帟	yì		canopy
熤	yì		(person)
饐	yì		rancid
圯	yí	(n.)	bridge
衣	yī		clothes gown / to dress / to wear
铱	yī		iridium
鹥	yī		widgeon
荑	yí		to weed

https://www.chinese-dictionary.org/

Chinese

200 words: yi

20 words: yī

1 word: 衣(yī, clothes)

Spoken language: ambiguous
Written language: unambiguous

勩	yì		toil
翼	yì		wing
邑	yì		city
泄	yì	(adj.)	idle
饴	yí		syrup
椅	yǐ		chair
咿	yī		(onomat.)
漪	yī		ripple
猗	yī		(interj.)
洟	yí		snivel
埸	yì		border
帟	yì		canopy
熤	yì		(person)
饐	yì		rancid
圯	yí	(n.)	bridge
衣	yī		clothes gown / to dress / to wear
铱	yī		iridium
鷖	yī		widgeon
荑	yí		to weed

https://www.chinese-dictionary.org/

Japanese

- Sentences can consist of 4 systems:

Rōmaji Kanji

Tomはテレビが観たい。

Katakana

Hiragana

- Compound words

日本語情報処理

Japan Language Information Processing

Japanese

Japanese

- One Character – Multiple Pronunciations (Readings)
  - 日**hi** (sun, day)
  - 日曜日**nichi** yō **bi**
  - 二日futsu **ka** (two days)
  - 今日kyou (today)

- One word – Multiple characters/ representations
  - To color: sasu - 差す / 注す / さす

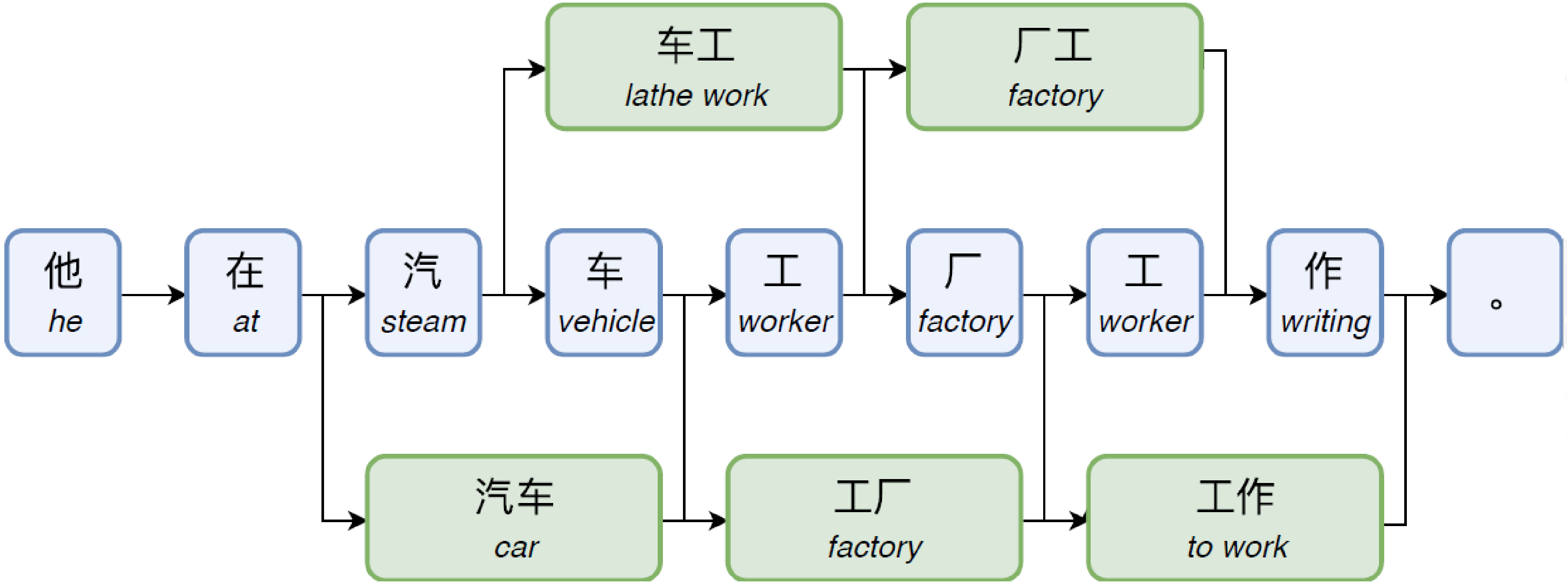
- One pronunciation – multiple words
  - Sasu: to offer, to hold up, to pour into, to color, to shine on, to aim at, …

System Overview

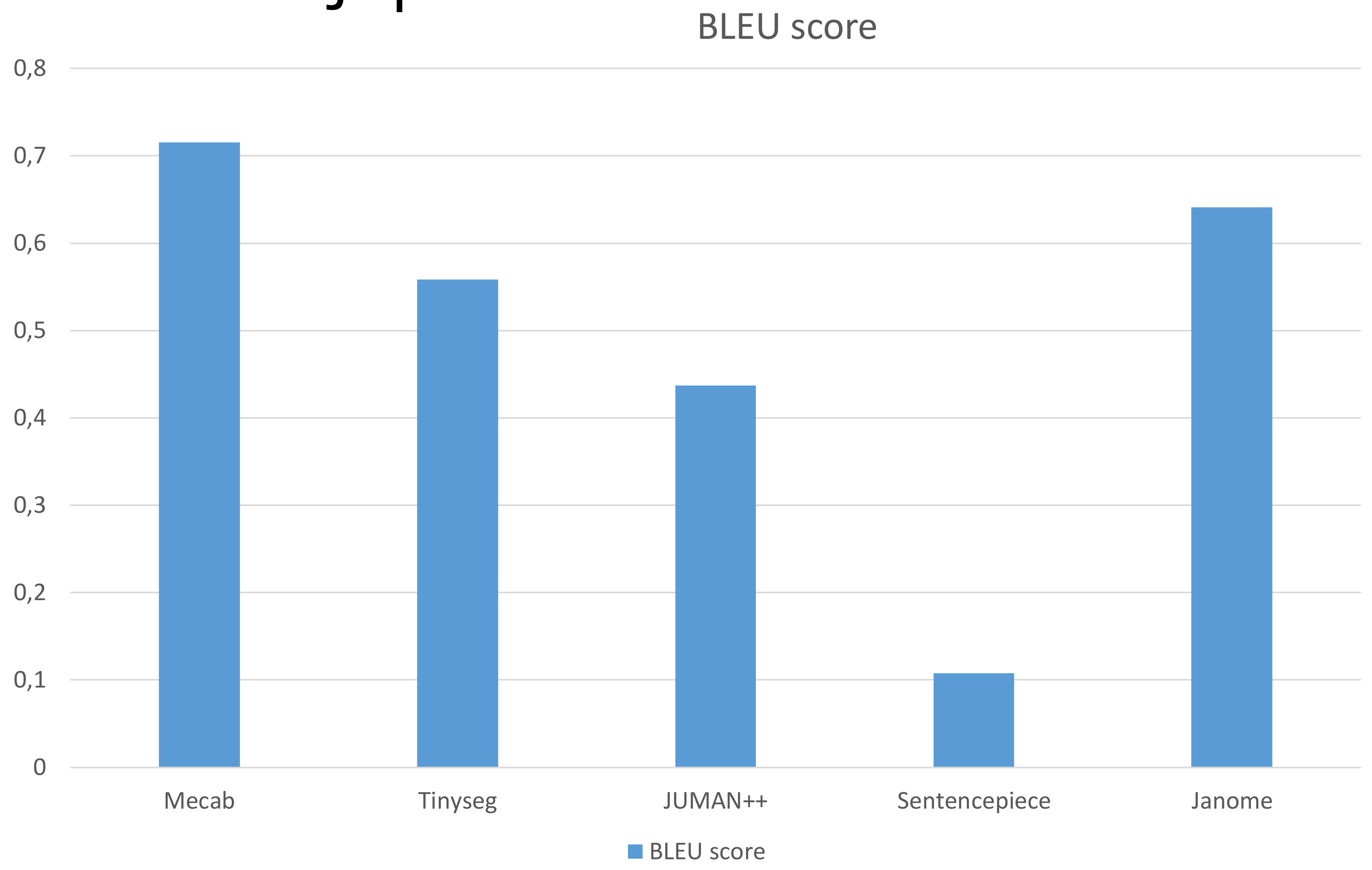
Segmentation

Segmentation

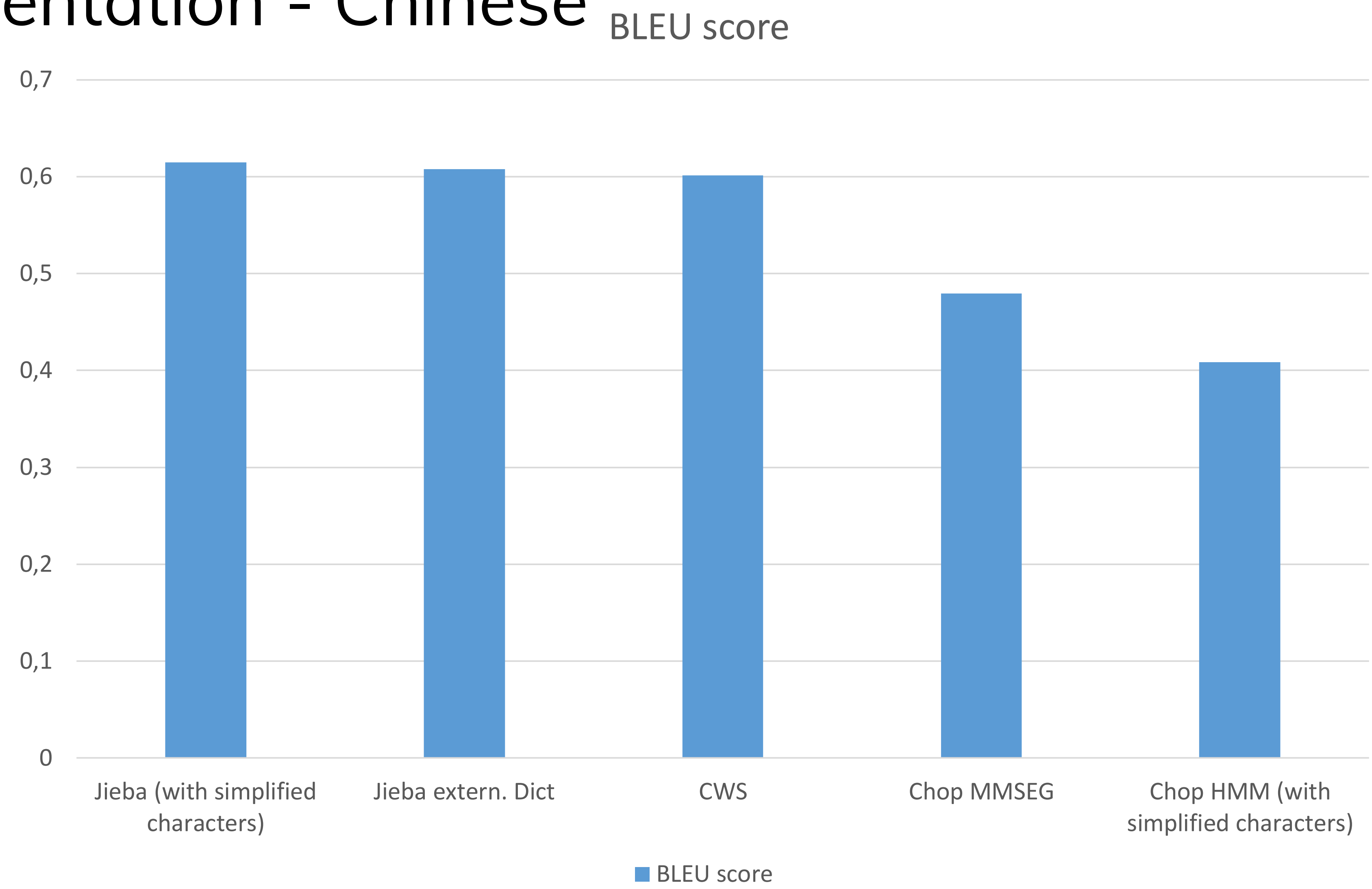
- Ambiguous segmentation in Chinese and Japanese



Segmentation - Japanese



Segmentation - Chinese



Named Entity Recognition

Named Entity Recognition (NER)

- Finds entities like location, organizations, person names in text

Google rebrands its business apps ✓

Let me google that for you. ✗

Named Entity Recognition (NER)

- Support for segmentation

→Split proper nouns better

他在阿里巴巴工作。

↓ NER

他在<u>阿里巴巴</u>工作。

↓ Replace entities

他在placeholder工作。

↓ Segmentation

他	在	placeholder	工作

↓ Replace placeholder

他	在	阿里巴巴	工作

Overview - Process

去北京首都国际机场的火车从3号站台出发。
The train for Beijing Capital International Airport leaves from platform 3.

Named Entity Recognition

去北京首都国际机场(3,9, LOC)的火车从3号站台出发。
The train for Beijing_Capital_International_Airport(15,51,LOC) leaves from platform 3.

Segmentation

去，北京首都国际机场，的，火车，从，3，号，站台 ,出发
The，train，for，Beijing_Capital_International_Airport，leaves，from，platform，3

Overview - Process

Dresden Database Systems Group

去北京首都国际机场的火车从3号站台出发。
The train for Beijing Capital International Airport leaves from platform 3.

Named Entity Recognition

去北京首都国际机场(3,9, LOC)的火车从3号站台出发。
The train for Beijing_Capital_International_Airport(15,51,LOC) leaves from platform 3.

Segmentation

去，北京首都国际机场，的，火车，从，3，号，站台 ，出发
The，train，for，Beijing_Capital_International_Airport，leaves，from，platform，3

Transliteration → **Transformation**

qù, běijīngshǒudūguójìjīchǎng, huǒchē, cóng, 3, háo, zhàntái, chūfā
train, Beijing_Capital_International_Airport, leaves, platform, 3

Training Word Embeddings

Training Word Embeddings

Facebooks[1] implemention of word embeddings

- Consindering subwords of words by using n-grams (3 to 6 by default)
- One word vector is the sum of the vectors of the subwords
- Better for words with same stem (plurals, inflection) compared to word2vec

[1]https://github.com/facebookresearch/fastText

Overview - Process

Segmentation

去，北京首都国际机场，的，火车，从，3，号，站台 ,出发
The, train, for, Beijing_Capital_International_Airport, leaves, from, platform, 3

Transliteration → Transformation

qù, běijīngshǒudūguójìjīchǎng, huǒchē, cóng, 3, háo, zhàntái, chūfā
train, Beijing_Capital_International_Airport, leaves, platform, 3

Train Word Embeddings

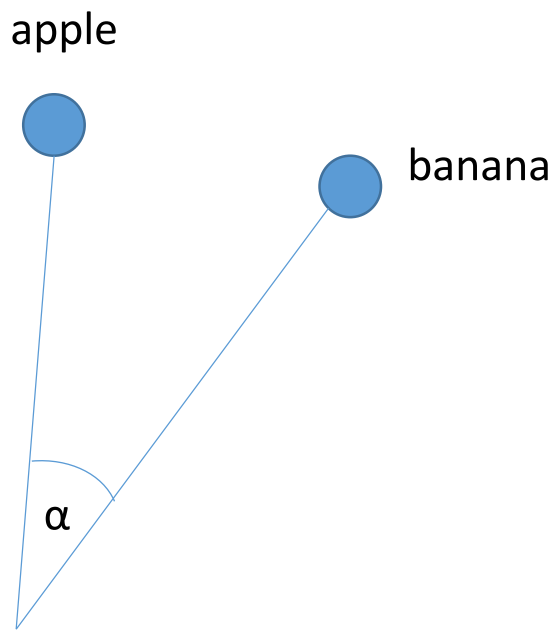
$V_{\text{huǒchē}} = \begin{bmatrix} \\ \\ \end{bmatrix}$ $V_{\text{train}} = \begin{bmatrix} \\ \\ \end{bmatrix}$

Evaluation

Evaluation – Similarity Score

- Determine cosine similarity: $\cos(\alpha)$

- $\cos(\alpha) = \frac{a \cdot b}{\|a\| \cdot \|b\|}$
- $\|a\| = \sqrt{{a_1}^2 + \cdots + {a_n}^2}$



Evaluation – Similarity Score

Dresden Database Systems Group

Test Set

Word 1	Word 2	y
apple	banana	7
professor	cucumber	0.3
money	cash	9.5

Evaluation – Similarity Score

Test Set

Word 1	Word 2	y
apple	banana	7
professor	cucumber	0.3
money	cash	9.5

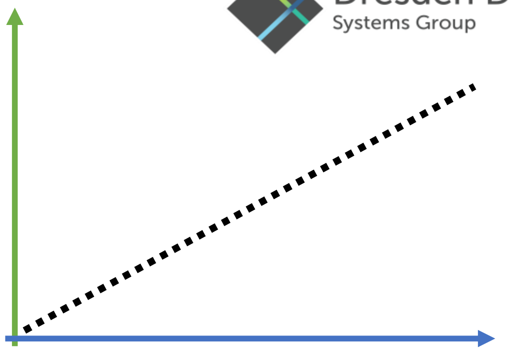
Predicted

y_{pred}
0.7
0.03
0.95

Correlation between y and y_{pred}
(Spearman‘s rank correlation coefficient)

Evaluation – Similarity Score

Spearman coefficient = 1



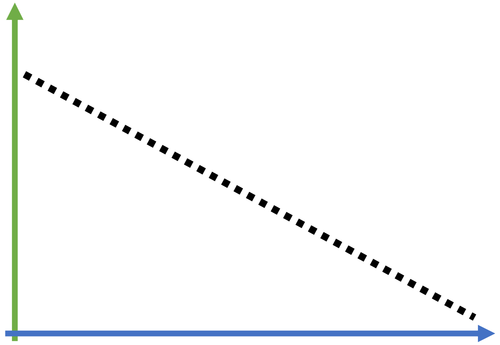
Test Set

Word 1	**Word 2**	**y**
apple	banana	7
professor	cucumber	0.3
money	cash	9.5

Predicted

y_{pred}
0.69
0.032
0.87

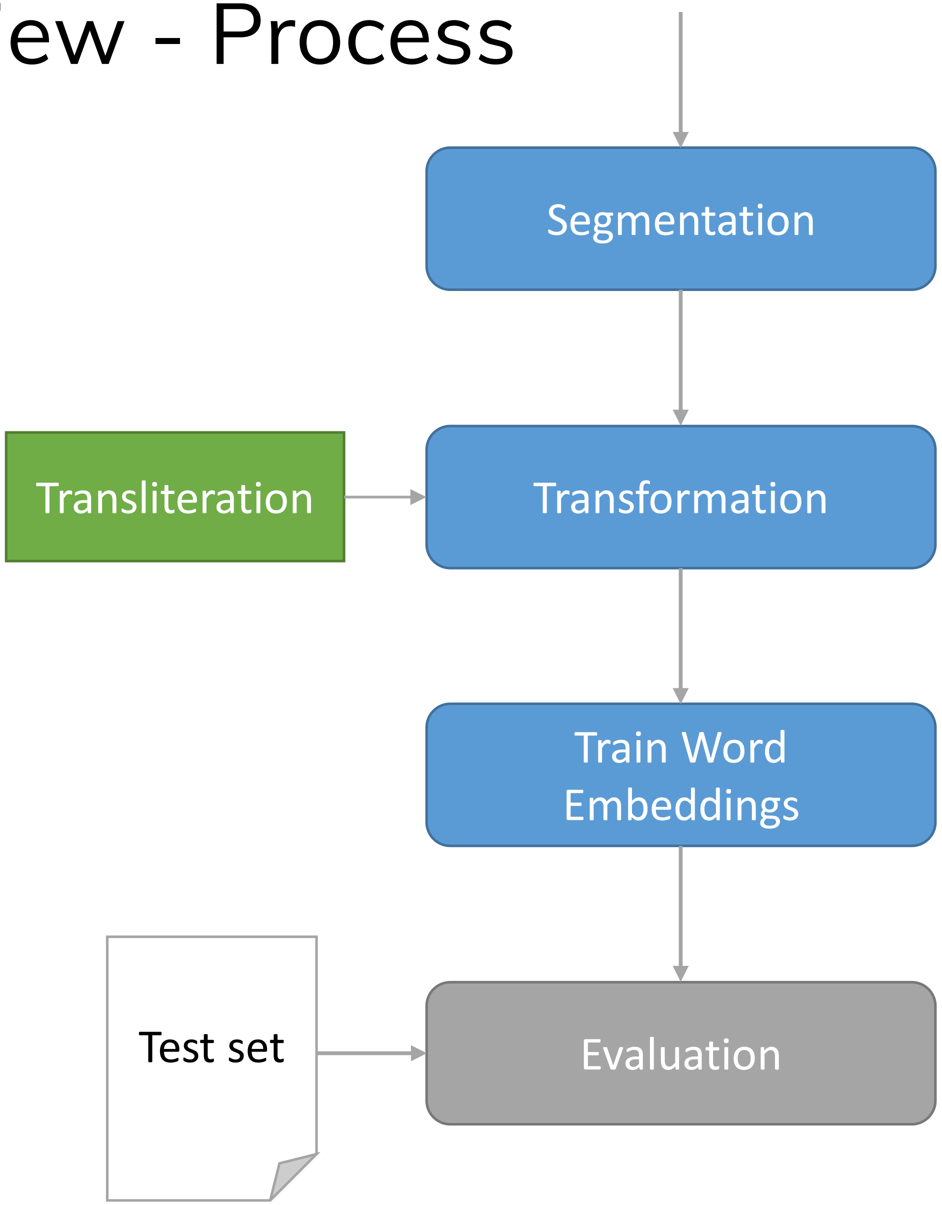
Spearman coefficient = -1



Spearman coefficient = 0

Correlation between y and y_{pred}
(Spearman‘s rank correlation coefficient)

Overview - Process



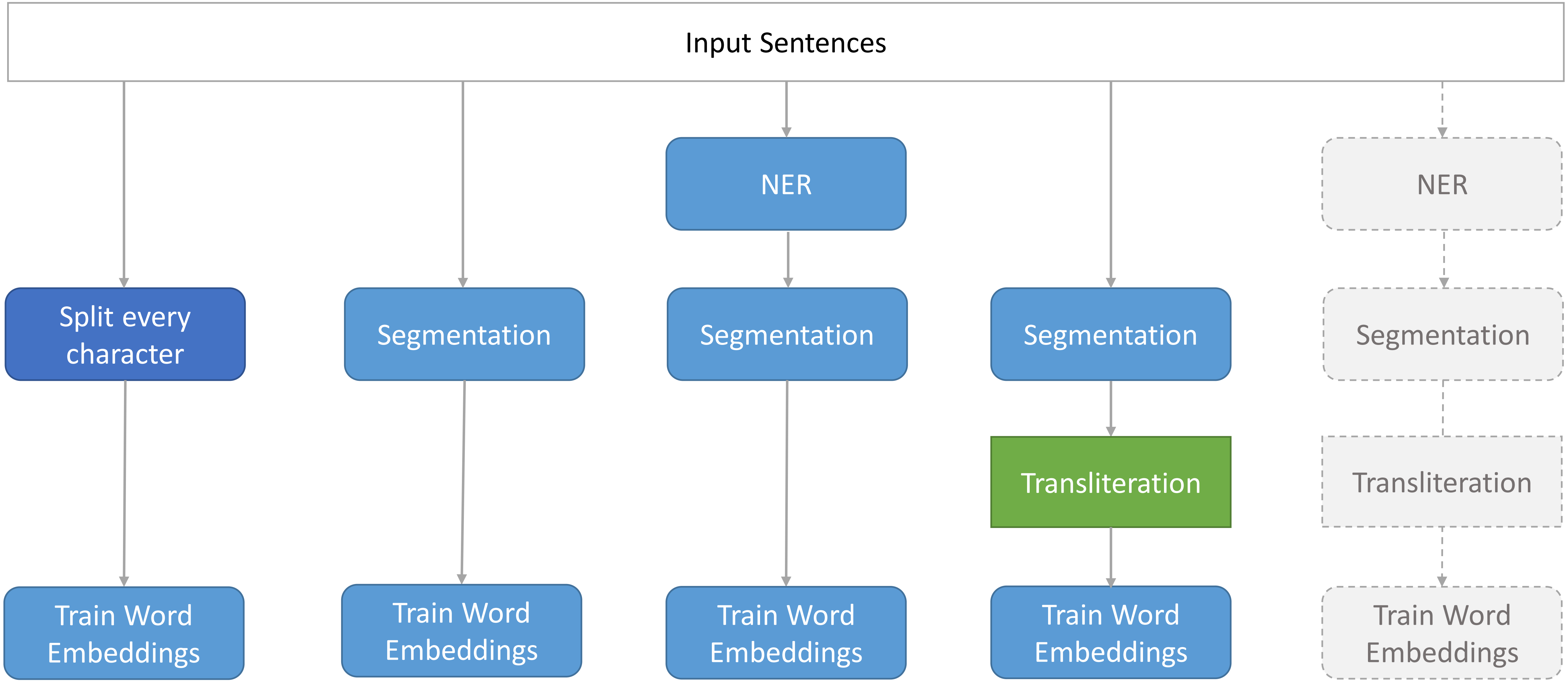
去，北京首都国际机场，的，火车，从，3，号，站台 ,出发
The, train, for, Beijing_Capital_International_Airport, leaves, from, platform, 3

qù, běijīngshǒudūguójìjīchǎng, huǒchē, cóng, 3, háo, zhàntái, chūfā
train, Beijing_Capital_International_Airport, leaves, platform, 3

$V_{\text{huǒchē}} = \begin{pmatrix} \ \end{pmatrix}$ $V_{\text{train}} = \begin{pmatrix} \ \end{pmatrix}$

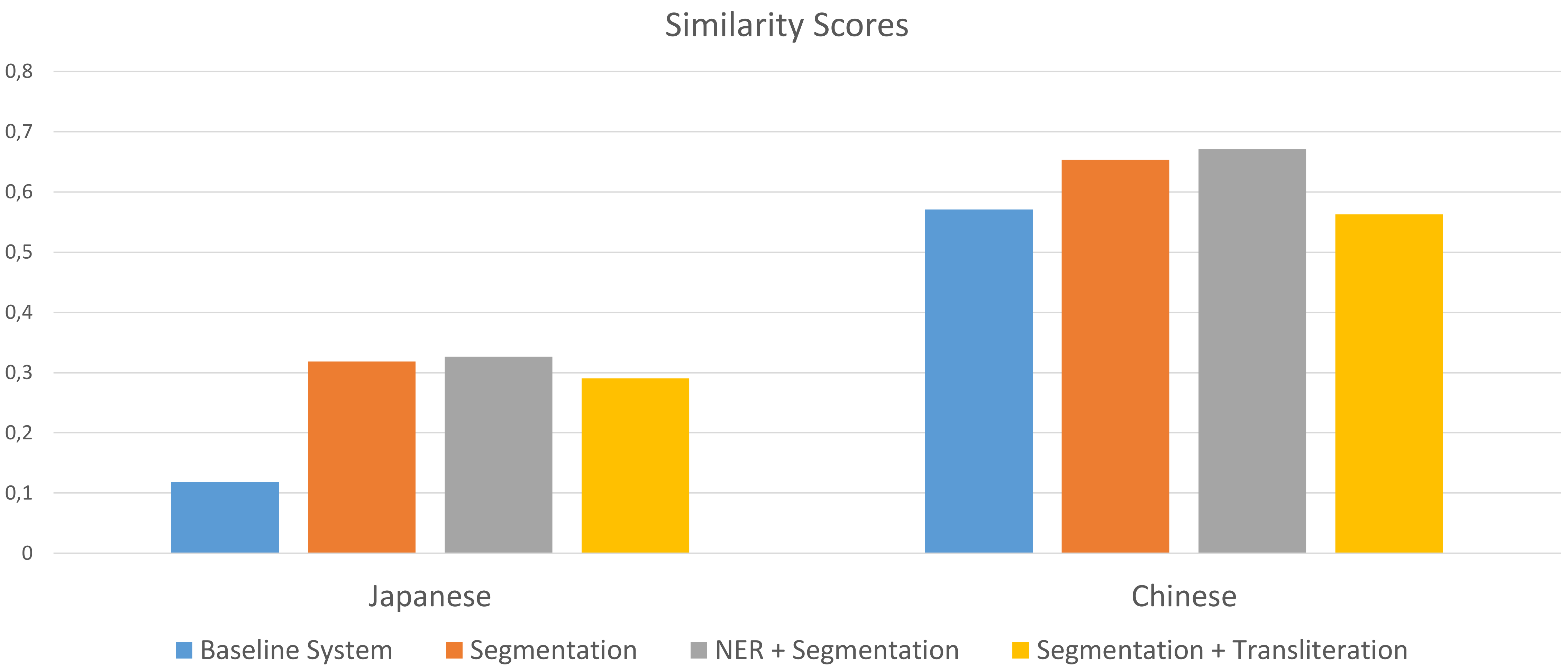
similarity score = 0.7

Configurations



Results

Results



Similarity Scores
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
Japanese
Chinese
Baseline System
Segmentation
NER + Segmentation
Segmentation + Transliteration

Conclusion

Trained Word Embeddings using Segmentation and NER

- Transliteration results in lower scores
- Suitable segmentation and NER improves results
- Can be used for synonyms, IR, translation

End

Questions?