Determining Domain-Specific Differences of Polysemic Words Using Context Information

Daniel Töws and Leif Van Holland

Fraunhofer FKIE

What is a *platform*?

Google Cloud Platform

SAP Cloud Platform

YouTube

computing platform	car platform	train platform
hosting platform	weapons platform	scaffolding

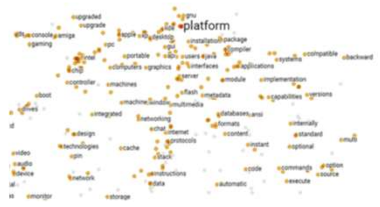
domain-specific

Fraunhofer FKIE

Goals

- Question: How could a **computer infer different meanings**?
- How would **humans** do it?
  - Knowledge about the world
  - **Context** of the sentence / text / …
  - ➡ **Surrounding words** are hints

➡

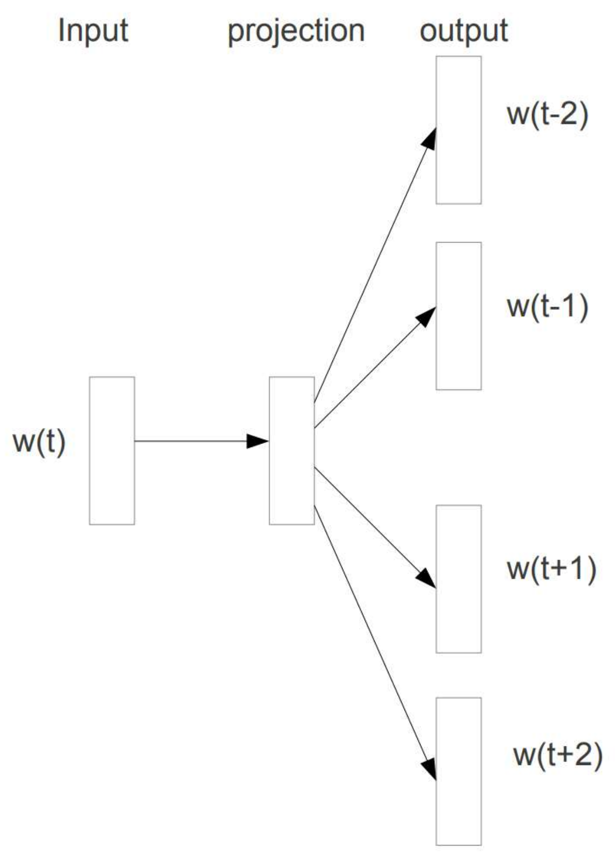


word embeddings

Fraunhofer FKIE

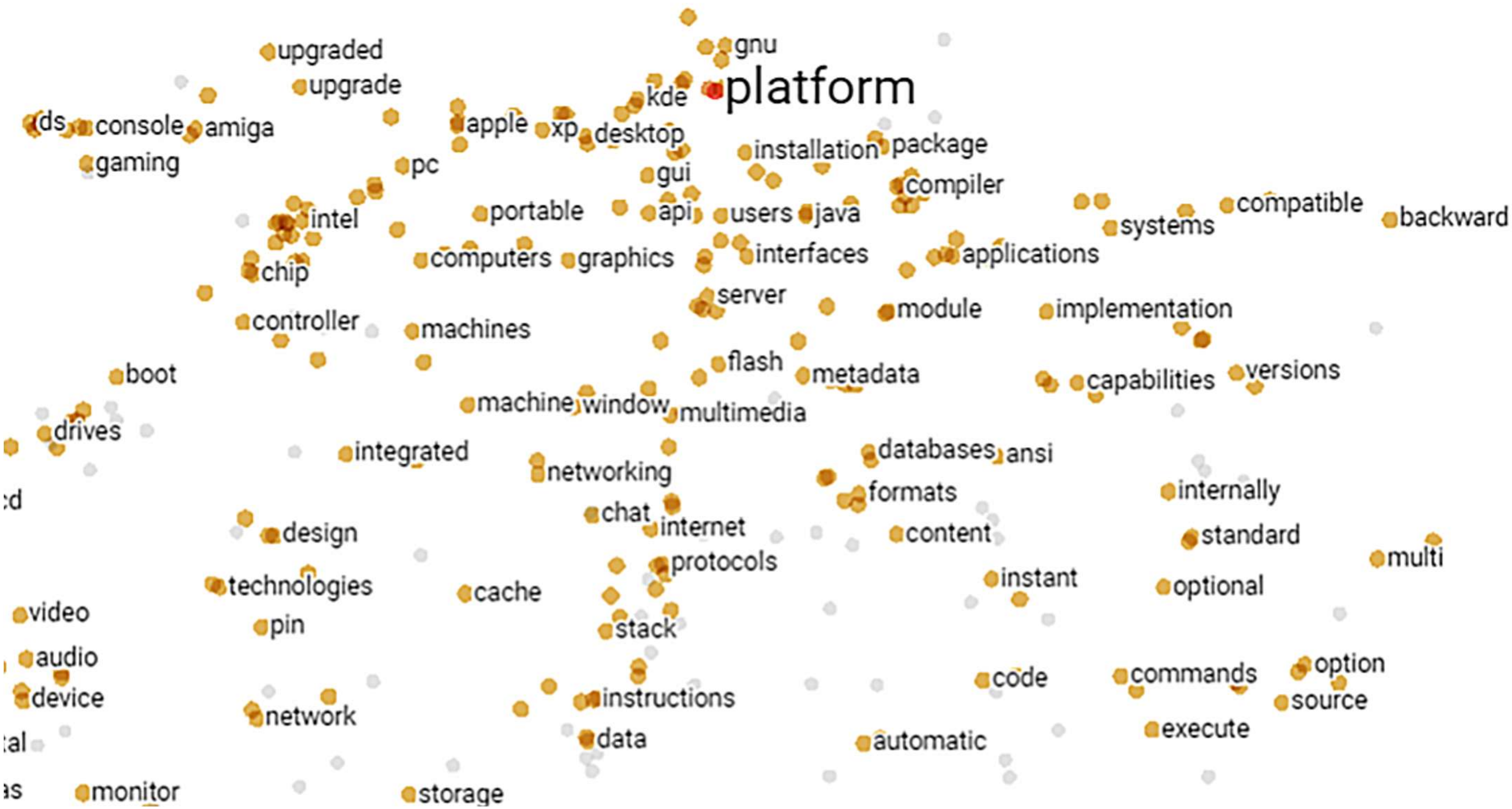
Word Embeddings

- "You shall know a word by the company it keeps" (Firth, J. R. 1957:11)
- Word2Vec (Mikolov et al., 2013) [1]
  - Efficient generation of a **vector space model** for words
  - represents **semantic relations** of words
  - *king - man + woman = queen*
  - Works if **training corpus** is **big**



[1] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119).

Example



Fraunhofer
FKIE

Goals

- Question: How could a computer infer different meanings?
- How would humans do it?
  - Knowledge about the world
  - Context of the sentence / text / …
  - ➔ Surrounding words are hints ➔

word embeddings

- Goal of our work
  - Determine **if a word is used differently** in two different corpora
  - Method should work on **corpora of arbitrary size**
  - Method should deliver **results *instantly***

Our Approach

- Given two **corpora** D_1, D_2, a *general-purpose* **word embedding** with vectors v_w, and a **word** t:
  - Determine **contexts** c_i of t in D_i $(i \in \{1,2\})$
  - Calculate **context centers**:

$$\boxed{center(c_i) = \frac{1}{|c_i|} \sum_{w \in c_i} \mathrm{IDF}_{D_i}(w) \cdot v_w}$$

  - Calculate **cosine similarity** of centers

$$\boxed{simc(t) = \frac{center(c_1) \cdot center(c_2)}{\|center(c_1)\| \cdot \|center(c_2)\|}}$$

Experimental setup

- Comparison with experiment from Ferrari et al. (2017) [2]:
  - Generated corpora by crawling **Wikipedia articles** of specific categories
  - Compared words from **Computer Science** with **five other categories**
- Analyzed **rank correlation** between our results and those of Ferrari et al.

[2] Ferrari, A., Donati, B., & Gnesi, S. (2017). Detecting domain-specific ambiguities: an NLP approach based on wikipedia crawling and word embeddings. In 2017 IEEE 25th International Requirements Engineering Conference Workshops (REW) (pp. 393-399). IEEE.

Some results

- Compared 100 **most common nouns** of Computer Science to other categories
- Table contains **10 most similar** and **10 most dissimilar** words

Computer Science vs

Electr. Engineering		Sports	
science	0.9954	science	0.9789
code	0.9899	computer	0.9699
security	0.9898	software	0.9651
memory	0.9874	**research**	0.9647
file	0.9873	human	0.9645
language	0.9870	data	0.9608
algorithm	0.9867	input	0.9591
database*	0.9864	work	0.9586
software	0.9859	device	0.9578
user	0.9858	web	0.9572
⋮		⋮	
see	0.9619	**window**	0.8995
structure	0.9617	solution	0.8980
type	0.9605	**non**	0.8931
input	0.9581	network	0.8916
reference	0.9559	security	0.8902
source	0.9529	**field**	0.8895
technology	0.9465	**programming**	0.8854
game	0.9458	server	0.8837
field	0.9447	microsoft	0.8632
non	0.9238	file	0.8498

Fraunhofer FKIE

Correlation

Results of Ferrari et al.

Our Reimplementation of Ferrari et al.

Our Approach

corpus	(a) Ferrari et al.		(c) Wiki		(d) Wiki + IDF	
	ρ	p	ρ	p	ρ	p
Electronic Engineering	0.5882	$<$0.0001	0.4703	$<$0.0001	0.6079	$<$0.0001
Literature	0.4717	$<$0.0001	0.5844	$<$0.0001	0.5921	$<$0.0001
Mechanical Engineering	0.4385	$<$0.0001	0.5442	$<$0.0001	0.5736	$<$0.0001
Medicine	0.4877	$<$0.0001	0.5429	$<$0.0001	0.6164	$<$0.0001
Sports	0.4684	$<$0.0001	0.3839	0.0001	0.4710	$<$0.0001

Correlation

Results of Ferrari et al.

Our Reimplementation of Ferrari et al.

Our Approach

corpus	(a) Ferrari et al.		(c) Wiki		(d) Wiki + IDF	
	ρ	p	ρ	p	ρ	p
Electronic Engineering	0.5882	<0.0001	0.4703	<0.0001	0.6079	<0.0001
Literature	0.4717	<0.0001	0.5844	<0.0001	0.5921	<0.0001
Mechanical Engineering	0.4385	<0.0001	0.5442	<0.0001	0.5736	<0.0001
Medicine	0.4877	<0.0001	0.5429	<0.0001	0.6164	<0.0001
Sports	0.4684	<0.0001	0.3839	0.0001	0.4710	<0.0001

moderate correlation across all categories

Correlation

Results of Ferrari et al. ⟷ Our Reimplementation of Ferrari et al. ⟷ Our Approach

corpus	(a) Ferrari et al.		(c) Wiki		(d) Wiki + IDF	
	ρ	p	ρ	p	ρ	p
Electronic Engineering	0.5882	<0.0001	0.4703	<0.0001	0.6079	<0.0001
Literature	0.4717	<0.0001	0.5844	<0.0001	0.5921	<0.0001
Mechanical Engineering	0.4385	<0.0001	0.5442	<0.0001	0.5736	<0.0001
Medicine	0.4877	<0.0001	0.5429	<0.0001	0.6164	<0.0001
Sports	0.4684	<0.0001	0.3839	0.0001	0.4710	<0.0001

again, moderate correlation

Correlation

Results of Ferrari et al.

Our Reimplementation of Ferrari et al.

Our Approach

corpus	(a) Ferrari et al.		(c) Wiki		(d) Wiki + IDF	
	ρ	p	ρ	p	ρ	p
Electronic Engineering	0.5882	<0.0001	0.4703	<0.0001	0.6079	<0.0001
Literature	0.4717	<0.0001	0.5844	<0.0001	0.5921	<0.0001
Mechanical Engineering	0.4385	<0.0001	0.5442	<0.0001	0.5736	<0.0001
Medicine	0.4877	<0.0001	0.5429	<0.0001	0.6164	<0.0001
Sports	0.4684	<0.0001	0.3839	0.0001	0.4710	<0.0001

all correlations are significant

Fraunhofer FKIE

Applying this to RE

- Used **four different** RE datasets describing parts of **one big project**
- Compared usage of common words for **pairs of datasets**

P_1 vs. P_3		P_1 vs. P_3'		P_2 vs. P_3		P_3 vs. P_3'	
bieten	0.9874	bieten	0.9884	bieten	0.9705	verbund	0.9940
möglichkeit	0.9874	möglichkeit	0.9881	möglichkeit	0.9705	service	0.9938
nutzer	0.9771	nutzer	0.9770	nutzer	0.9691	möglichkeit	0.9933
fähig	0.9422	fähig	0.9622	ermöglichen	0.9584	bieten	0.9932
chat	0.9397	konfigurieren	0.9618	bereitstellen	0.9510	nutzer	0.9931
⋮		⋮		⋮		⋮	
mission	0.9177	anzeigen	0.9052	$\boxed{\text{services}}$	$\boxed{0.9008}$	informationen	0.9251
automatisch	0.8941	durchzuführen	0.9012	gemäß	0.8966	endgerät	0.9148
informationen	0.8637	entsprechend	0.8961	mobilen	0.8911	clients	0.8703
$\boxed{\text{service}}$	$\boxed{0.8625}$	nutzung	0.8899	informationen	0.8730	planning	0.8701
durchzuführen	0.8540	$\boxed{\text{service}}$	$\boxed{0.8750}$	plattform	0.8621	durchzuführen	0.8436

P_1, P_2 and P_3 describe different aspects of a service

Fraunhofer FKIE

Applying this to RE

- Used **four different** RE datasets describing parts of **one big project**
- Compared usage of common words for **pairs of datasets**

P_1 vs. P_3		P_1 vs. P_3'		P_2 vs. P_3		P_3 vs. P_3'	
bieten	0.9874	bieten	0.9884	bieten	0.9705	verbund	0.9940
möglichkeit	0.9874	möglichkeit	0.9881	möglichkeit	0.9705	service	0.9938
nutzer	0.9771	nutzer	0.9770	nutzer	0.9691	möglichkeit	0.9933
fähig	0.9422	fähig	0.9622	ermöglichen	0.9584	bieten	0.9932
chat	0.9397	konfigurieren	0.9618	bereitstellen	0.9510	nutzer	0.9931
⋮		⋮		⋮		⋮	
mission	0.9177	anzeigen	0.9052	services	0.9008	informationen	0.9251
automatisch	0.8941	durchzuführen	0.9012	gemäß	0.8966	endgerät	0.9148
informationen	0.8637	entsprechend	0.8961	mobilen	0.8911	clients	0.8703
service	0.8625	nutzung	0.8899	informationen	0.8730	planning	0.8701
durchzuführen	0.8540	service	0.8750	plattform	0.8621	durchzuführen	0.8436

P_3 and P_3' are from the same subproject

Applying this to RE

- Used **four different** RE datasets describing parts of **one big project**
- Compared usage of common words for **pairs of datasets**

P_1 vs. P_3		P_1 vs. P_3'		P_2 vs. P_3		P_3 vs. P_3'	
bieten	0.9874	bieten	0.9884	bieten	0.9705	verbund	0.9940
möglichkeit	0.9874	möglichkeit	0.9881	möglichkeit	0.9705	service	0.9938
nutzer	0.9771	nutzer	0.9770	nutzer	0.9691	möglichkeit	0.9933
fähig	0.9422	fähig	0.9622	ermöglichen	0.9584	bieten	0.9932
chat	0.9397	konfigurieren	0.9618	bereitstellen	0.9510	nutzer	0.9931
⋮		⋮		⋮		⋮	
mission	0.9177	anzeigen	0.9052	services	0.9008	informationen	0.9251
automatisch	0.8941	durchzuführen	0.9012	gemäß	0.8966	endgerät	0.9148
informationen	0.8637	entsprechend	0.8961	mobilen	0.8911	clients	0.8703
service	0.8625	nutzung	0.8899	informationen	0.8730	planning	0.8701
durchzuführen	0.8540	service	0.8750	plattform	0.8621	durchzuführen	0.8436

Future Work

- **Thorough analysis** of the values produced
  - Tests on **other corpora**
  - What are **constraints** on corpus size / word count / …
  - How meaningful are values in **small corpora**?
- **How** can the value be used?
  - **Qualitative study**: How do users incorporate the value in a RE process?
  - What if **only one corpus** is given?

Questions