On the Use of Word Embeddings for Identifying Domain Specific Ambiguities in Requirements

Siba Mishra and Arpit Sharma

Department of Electrical Engineering and Computer Science
Indian Institute of Science Education and Research, Bhopal, India

September 24, 2019
Jeju Island, South Korea

Outline

Software Requirements

Requirements

- specify what a software is supposed to do
- serve as a legal agreement between the client and software development organization
- influence subsequent steps in software development
- provide a basis for testing
- are usually written in common natural language (NL)

Ambiguous Software Requirements

Ambiguity

- means that a single reader can interpret the requirement in more than one way
- means multiple readers come to different interpretations
- is one of the major cause of poor quality requirements
- may lead to time and cost overrun (worst case - project failure)

Domain Specific Ambiguity

- stakeholders with different technical backgrounds and domain expertise
- typical computer science (CS) terms may be interpreted differently by stakeholders (with no CS background)

Examples

- Platform (CS $\neq$ Petroleum)
- Tree (CS $\neq$ Environment)
- Cell (CS $\neq$ Biomedical)
- Operation (CS $\neq$ Military)
- State (CS $\neq$ Civil)

Domain Specific Ambiguity

- stakeholders with different technical backgrounds and domain expertise
- typical computer science (CS) terms may be interpreted differently by stakeholders (with no CS background)

Examples

- Platform (CS $\neq$ Petroleum)
- Tree (CS $\neq$ Environment)
- Cell (CS $\neq$ Biomedical)
- Operation (CS $\neq$ Military)
- State (CS $\neq$ Civil)

Goal : Detect domain specific ambiguous CS words

Table of Contents

Word Embeddings

- a powerful approach for analyzing language
- widely used in information retrieval and text mining
- dense representation of words (numeric vectors)
- capable of capturing the context of a word
- identifying semantically similar words, i.e., *cosine similarity*
- examples - GloVe (Stanford), Word2vec (Google), fastText (Facebook)

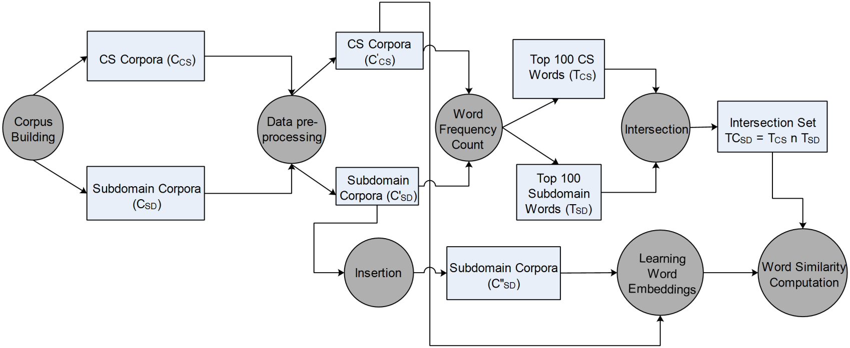
Word Embeddings

- a powerful approach for analyzing language
- widely used in information retrieval and text mining
- dense representation of words (numeric vectors)
- capable of capturing the context of a word
- identifying semantically similar words, i.e., *cosine similarity*
- examples - GloVe (Stanford), Word2vec (Google), fastText (Facebook)

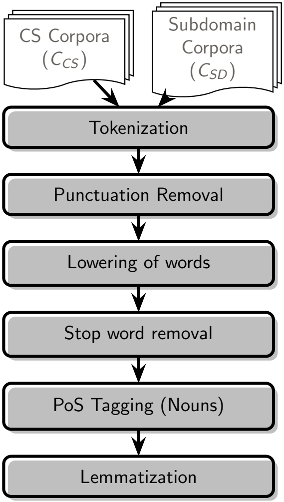
Word2Vec

Table of Contents

Our Approach



NLP Pipeline



Descriptive Statistics

Category Name	Pages	Total Sentences	Total Words	Vocabulary
Computer Science (CS)	9021	2,46,359	18,37,492	18,192
Building Engineering (BUE)	9002	3,52,005	25,77,515	23,538
Mechanical Engineering (MCEE)	7587	3,31,746	24,78,977	20,463
Electronic Engineering (ELCE)	7147	2,47,649	18,78,728	18,451
Civil Engineering (CIVE)	7071	2,83,337	21,42,500	20,427
Aerospace Engineering (AE)	4661	1,61,867	13,13,054	14,524
Chemical Engineering (CHEE)	4442	2,03,637	15,37,857	15,339
Environmental Engineering (ENVE)	2626	1,16,685	8,72,305	10,924
Marine Engineering (MAEE)	1369	31,712	2,23,956	4,880
Industrial Engineering (INEE)	1060	42,751	3,41,308	5,845
Military Engineering (MLEE)	932	32,068	2,42,944	5,027
Biomedical Engineering (BIEE)	924	52,599	3,87,492	8,214
Petroleum Engineering (PTEE)	419	15,148	1,21,614	2,965
Ceramic Engineering (CERE)	318	12,465	83,705	2,581

Table of Contents

Large-sized Subdomains

Words	Similarity Score	Most Similar Words (CS)	Most Similar Words (CIVE)
state	0.49546	class, algorithm, model, automaton, domain	land, link, highway, survey, government
source	0.47845	library, tool, application, specification	groundwater, recovery, growth, cycle, consumption, storage

Words	Similarity Score	Most Similar Words (CS)	Most Similar Words (AE)
space	0.59611	domain, set, regression, solution, element, number, property	mission, launch, spacecraft, shuttle, traffic, safety, satellite
system	0.16622	software, data, process, application, program	rocket, radio, vehicle, navigation, radar, power

Large-sized Subdomains

Words	Similarity Score	Most Similar Words (CS)	Most Similar Words (CHEE)
product	0.57344	source, code, cache, mode, requirement	steam, soil, ammonia, combustion, methane, compound
process	0.52690	command, code, layer, requirement, specification, memory, storage	hydrogen, carbon, combustion, water, emission, oxygen
environment	0.50088	driver, share, encryption, resource, database	biodiesel, coal, pollution, impact, waste, treatment

Example Sentences

- state (CS) : The state at which the automaton stops is called the final state.
- state (CIVE) : In 1872, Alexey Von Schmidt undertook the survey of the state line.

Medium-sized Subdomains

Words	Similarity Score	Most Similar Words (CS)	Most Similar Words (MLEE)
machine	0.65724	process, analysis, code, computation, data	defense, casualty, explosive, explosion, ammunition
operation	0.37621	object, block, integration, procedure, query	combat, hill, infantry, battle, attack

Words	Similarity Score	Most Similar Words (CS)	Most Similar Words (MAEE)
structure	0.56361	class, object, method, recursion, regression, procedure	tank, port, dock, yacht, plant, coast

Medium-sized Subdomains

Words	Similarity Score	Most Similar Words (CS)	Most Similar Words (ENVE)
tree	0.21638	heap, queue, insertion, sort, hash, merge, algorithm	hydrogen, reaction, biogas, reserve

Example Sentences

- tree (CS) : A left-leaning red–black (LLRB) tree is a type of self-balancing binary search tree
- tree (ENVE) : Van Mahotsav is an annual pan-Indian tree planting festival

Small-sized Subdomains

Words	Similarity Score	Most Similar Words (CS)	Most Similar Words (PTEE)
platform	0.60037	editor, email, desktop, apple, interface, sun, gui, firewall	equipment, sea, site, lift, construction, level
tool	0.55951	application, database, protocol, source, web, cloud, library	injection, hole, drill, perforation, valve, pump

Words	Similarity Score	Most Similar Words (CS)	Most Similar Words (CERE)
application	0.47303	tool, user, suite, platform, microsoft	water, cement, bone, steel, insulator, chemical, powder

Example Sentences

- platform (CS) : HoneyC is a platform independent open source framework written in Ruby
- platform (PTEE) : The first tower emerged in the early 1980s with the installation of Exxon's Lena oil platform

Table of Contents

1. A. Ferrari and S. Gnesi, "Using collective intelligence to detect pragmatic ambiguities," in 20th IEEE International Requirements Engineering Conference (RE), September 2012, pp. 191–200
2. A. Ferrari, B. Donati, and S. Gnesi, "Detecting domain-specific ambiguities: An NLP approach based on wikipedia crawling and word embeddings," in 25th IEEE International Requirements Engineering Conference Workshops (REW), September 2017, pp. 393–399
3. A. Ferrari, A. Esuli, and S. Gnesi, "Identification of cross-domain ambiguity with language models," in 5th International Workshop on Artificial Intelligence for Requirements Engineering (AIRE), August 2018, pp. 31–38

Table of Contents

Conclusions

- demonstrated the applicability of word2vec algorithm for detecting domain specific ambiguity
- demonstrated its applicability in both small and large software projects
- similarity threshold to detect ambiguous words

Future Work

- investigate applicability for large scale requirements specification
- detect similarity between natural language requirements in software product lines
- compare word2vec with other word embedding techniques, e.g., GloVe, fastText etc

Thank You