Junfang JIA, Valeriia Tumanian, Guoqiang LI, 2021. Discovering semantically related technical terms and web resources in Q&A discussions. *Frontiers of Information Technology & Electronic Engineering*, 22(7):969-985. https://doi.org/10.1631/FITEE.2000186

Discovering semantically related technical terms and web resources in Q&A discussions

Key words: Technical terms; Web resources; Word embedding; Q&A web site; Clustering task; Recommendation task

Corresponding author: Guoqiang LI
E-mail: jiajunfang816@163.com; valeria.tumanyan@yandex.ua; li.g@sjtu.edu.cn
ORCID: Junfang JIA, https://orcid.org/0000-0002-3451-8487; Valeriia TUMANIAN https://orcid.org/0000-0001-6651-191X; Guoqiang LI, https://orcid.org/0000-0001-9005-7112

Motivation

- Developers often use a set of related techniques and seek information from relevant web resources. Thus, discovering semantically related technical terms and web resources offers the opportunity to design appealing services to facilitate information retrieval and information discovery.
- We can develop some recommendation systems for serendipitous discovery of information.
- However, keyboard-based search methods and correlation-similarity-based methods cannot discover semantically related but heterogeneous technical terms and web resources that are not similar in attributes or content.

Main idea

The goal is to propose a neural-language-model-based framework for discovering semantically similar technical terms and web resources from community of Q&A discussions. The whole framework is composed of three modules:

1) pseudo-document generation from the community Q&A discussions (SO discussion threads) that contain both tags and URLs;
2) technical-term and web-resource embedding learning: an extended skip-gram model takes technical terms and web resources as an input and predicts the technical terms and web resources as an output;
3) similarity computation.

Main idea

To evaluate the vector representations of technical terms and web resources, we conduct three studies:

1) clustering task: *"can learned technical term and web-resource embeddings using this simple neural network capture semantic regularities of the related technical terms and web resources?"*
2) search task: *"can the learned technical term and web-resource embeddings support the recommendation of semantically related technical terms and web resources?"*
3) semantic reasoning task which covers *semantic addition* and *analogical reasoning*.

Methods for T1

For the clustering task, we:

- use *K*-means, cluster technical terms and web resources based on the cosine similarity of the learnt vector representations;
- manually inspect the semantic relatedness of technical terms and web resources in the resulting clusters, and identify three categories: concept-centric, technique-centric, and task-centric;
- quantitively analyze the semantic relatedness of technical terms and web resources using the intra- and inter-cluster content similarities;
- comparatively study the clusters obtained using *K*-means with the clusters obtained using the LDA methods.

Major results for T1

- Technical terms and web resources can be well clustered around concepts, techniques, and tasks.
- The learnt vector representations can capture the semantic relatedness of the corresponding technical terms and web resources.
- Our technical-term and web-resource embeddings provide an alternative to the traditional topic models to obtain high-quality clusters of semantically related technical terms and web resources.

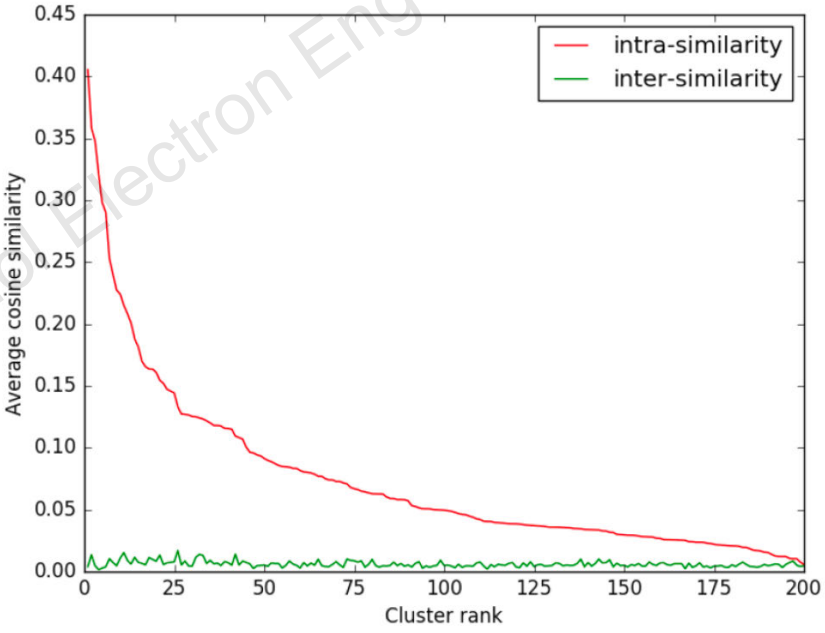

Fig. 2 Intra-cluster and inter-cluster similarities: embedding-based K-means

Methods for T2

For the search task, we:

- design four types of search tasks that can be useful for various online applications, denoted as *term –>? term, term –>? url, url –>? term,* and *url –>? url;*
- reduce these search tasks to a simple *K*-nearest neighbor search in the joint vector space between technical terms and web resources;
- quantitively analyze the co-occurrence frequency of a given technical term (or web resource) and recommended technical terms and/or web resources.

Major results for T2

Table 3 Embedding-based recommendation versus Stack Overflow related tags

query tag	cordova	http	algebra	gmail	hadoop
Our method	phonegap-plugins	httprequest	polynomial-math	imap	mapreduce
	appcelerator	http-status-codes	quadratic	hotmail	hdfs
	phonegap-build	http-headers	symbolic-math	gmail-imap	hive
	sencha-touch-2	content-length	calculus	email-integration	hbase
	jquery-mobile	http-request	discrete-mathematics	pop3	cloudera
Stackoverflow Related Tags	android	java	math	email	mapreduce
	javascript	php	algorithm	php	java
	ios	android	python	smtp	hive
	jquery	javascript	java	android	hdfs
	jquery	post	c++	imap	apache-pig

- The learnt technical-term and web-resource embeddings can complement existing keyword-based and co-occurrence-based search systems in numerous online applications (as our recommendations are more semantically related).

Methods for T3

- In the semantic reasoning task, we report an exploratory study using simple algebraic operations on the learned technical-term and web-resource embeddings in two types of semantic reasoning tasks:

1) semantic addition;
2) analogical reasoning: a question of the form *"a is to A as ? Is to B"* can be inferred from the words whose embedding is most similar to the vector *a – A + B.*

Major results for T3

- Our model can recommend highly related technical terms and web resources for the given input.
- We observe some unrelated web resources in our results which could be filtered out by incorporating domain knowledge into a word embedding or by further considering the discussion context in which they appear.

Table 4 Examples of semantic addition

Technical term	Web resource
ios + facebook	
facebook-ios-sdk	https://github.com/facebook/facebook-ios-sdk
sharekit	https://dev.twitter.com/docs/ios
fbconnect	http://getsharekit.com
mgtwitterengine	https://github.com/sharekit/sharekit
three20	http://www.getsharekit.com

Table 5 Examples of analogical reasoning

Technical term	Web resource
http://www.python.org/dev/peps/pep-0008 - python + java	
annotation-processing	http://www.oracle.com/technetwork/java /codeconventions-135099.html#367
checkstyle	http://www.oracle.com/technetwork/java/javase /documentation/ codeconvtoc-136057.html
apache-commons-lang	http://www.oracle.com/technetwork/java/javase /documentation/ codeconventions-135099.html#367
anonymous-class	http://code.google.com/p/javadude/wiki/annotations
xtend	http://docs.oracle.com/javase/specs/jls/se7/html/jls-4.html

Conclusions

- Different from existing approaches, the underlying assumption of our approach is that semantically similar or related technical terms and web resources would be present in similar technical-term and web-resource contexts.
- The learned technical-term and web-resource factor of representations works surprisingly well for clustering semantically related technical terms and web resources, even when they are not similar in content.
- The learned factor representations have a wide potential in reducing complex search and semantic reasoning tasks to simple *K*-nearest neighbor search and simple algebraic operations in the technical-term and web-resource and bedding space.