Master's thesis

Master's Programme in Computer Science

An analysis of the semantic shifts of citations

Jiayue Xue

June 8, 2021

Faculty of Science

University of Helsinki

Contact information

P. O. Box 68 (Pietari Kalmin katu 5)
00014 University of Helsinki,Finland

Email address: info@cs.helsinki.fi
URL: http://www.cs.helsinki.fi/

HELSINGIN YLIOPISTO – HELSINGFORS UNIVERSITET – UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty	Koulutusohjelma — Utbildningsprogram — Study programme
Faculty of Science	Master's Programme in Computer Science

Tekijä — Författare — Author

Jiayue Xue

Työn nimi — Arbetets titel — Title

An analysis of the semantic shifts of citations

Ohjaajat — Handledare — Supervisors

Dr. Alan Medlar, Prof. Dorota Glowacka

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
Master's thesis	June 8, 2021	26 pages

Tiivistelmä — Referat — Abstract

The semantic shifts in natural language is a well established phenomenon and have been studied for many years. Similarly, the meanings of scientific publications may also change as time goes by. In other words, the same publication may be cited in distinct contexts. To investigate whether the meanings of citations have changed in different scenarios, which is also called in the semantic shifts in citations, we followed the same ideas of how researchers studied semantic shifts in language. To be more specific, we combined the temporal referencing model and the Word2Vec model to explore the semantic shifts of scientific citations in two aspects: their usages over time and their usages across different domains. By observing how citations themselves changed over time and comparing the closest neighbors of citations, we concluded that the semantics of scientific publications did shift in terms of cosine distances.

ACM Computing Classification System (CCS)
Computing methodologies → Machine learning

Avainsanat — Nyckelord — Keywords

semantics of citations, semantic shift

Säilytyspaikka — Förvaringsställe — Where deposited

Helsinki University Library

Muita tietoja — övriga uppgifter — Additional information

Networking study track

Contents

1 Introduction

Once an English word was created, its content will not be changed. However, as people may use natural language for different purposes, it is completely possible that the meanings of English words will be changed or extended. In fact, this phenomenon is called the semantic shifts in language [18], [9] and was confirmed for a long time. For instance, in 1900s, the word "gay" meant cheerful. As time went by, its meaning shifted towards homosexuality. In the research area, many early works resorted to word embedding models to investigate on how the semantics of natural language shifted over time. Although different researchers preferred different word embedding models, the basic idea can be split into three steps. First, the whole corpus was split into several sub-corpora by different time periods. Then, the researchers encoded the meanings of the words in sub-corpora into their word vectors representations. Finally, semantic shifts can be observed by comparing the word vectors in different time period of a certain word.

The content of a publication will not change after it is publish, it might be cited by other publications by different purposes and thus its meaning in different contexts might be different as well. For instance, one paper utilized the neural network proposed by one publication to implement its system. however, an other paper also cited the same publication but this time it made some modifications to the original model and achieved better performance. The difference of the meanings is minor in the first place. However, it could accumulate as time goes by and finally become large enough to be observed. One extreme example is the neural network. It was first proposed in 1960s. At that time, it did not attract much attention. However, the neural networks have been developed a lot in recent years and become one of the most popular topic in the world.

One property shared by English words and scientific citations is that their content will not be changed once created but their meanings might become different from their original ones. In that, we can adopt the same idea of how researchers studied the semantic shifts in language to investigate the semantic shifts in citations. In this thesis, we adopted a strategy called temporal referencing to attach our concerned information to the original publications and utilized a neural network called Word2Vec to encode the semantics of the

citations. By this practice, we could understand the meaning of a particular publication by other English words whose word vectors were similar to the one of that publication. In addition, we were able to observe how one publication evolved as time went on by comparing its own word vectors in different time period. Note that our definitions of semantic shifts or changes are evaluated by cosine distance. In addition, when we discuss the semantic shift of a citation over time, we mean the cosine similarity between the first year it was cited and each successive year.

On one hand, the semantic shifts over time could happen in every domain. On the other hand, for publications that were cited by multiple domains, they are more likely to be cited in distinct contexts and thus their meanings might change more dramatically than other kinds of publications. In that, we proposed the following research questions and tried to answer them in this thesis:

1. Do the semantics of publications tend to change over time or be stable?
2. Whether semantic shifts are more likely to happen in cross-domain publications or not?

2 Background

2.1 Word Embedding

Feature engineering techniques are frequently applied in the machine learning field. Especially, when dealing with tasks related to natural language, transforming the human-readable text into a machine-understandable representation is a necessary step before conducting any other processing. After many years' development, there are multiple ways to achieve this goal, such as Bag of Words [19], TF-IDF, and Word Embedding. Given a paragraph, the basic idea of Word Embedding is to map all the words in the text to vectors of the same dimension.

2.1.1 Word2Vec

Word2Vec is the collective name of two neutral network models: Continues Bag of Words (CBOW) and Skip-gram [12]. The aim of CBOW is to predict the probability that the word appear in the current position according to the context of the current position. In contrast, Skip-gram try to use the word in the current position to predict the probabilities of the context of the current position.

Although the design philosophy of these two networks are different, they still share a lot in common. For instance, their networks both consist of three layers: input layer, projection layer and output layer:

1. The input of the input layer is English words encoded by one-hot coding
2. The projection layer has K hidden units and their functionality is to map the N-dimension input vector to a K-dimension feature vector which is calculated by the $N \times K$ weight matrix
3. The output layer also has output units, where every unit corresponds to the one original word. Finally, Softmax activation function is applied to the output vector to computed the probability of every word

Figure 2.1: Semantic Shift Example from Reference [7]

2.2 Semantic Shift

Many works utilized word embedding methods to explore semantic shift and achieved promising results. One typical example in figure 2.1 came from Figure 1 of reference [7]. Although different researchers preferred different word embedding models, their ideas were similar:

1. Split the entire corpus into several sub-corpora by different time periods
2. In every sub-corpus, learn vector representations of the words in the vocabulary by a certain type of word embedding model
3. Apply an suitable metric to quantify semantic shift of a word by comparing the vectors of that word across different sub-corpora

One key step in this procedure is how to compare these word vectors.

2.2.1 Alignment-based Strategies

The alignment-bases strategies mainly consist of two steps: encoding and alignment. For instance, [7] presented two methods to encode English words: PPMI [11] and SVD [3].

Since the word vectors were trained based different sub-corpora and thus laid in their own vector spaces. In order to compare them, a series of transformations are inevitable to align these word vectors into the same vector space. In terms of PPMI, alignment was not necessary because the word vectors lay in the same vector space in the first place. For SVD, many researchers proposed different ways to solve the problem. [10] adopted heuristic local

alignments for every word. [7] utilized orthogonal Procrustes with an application of SVD [17] to accomplish the alignment step. To be more specific, suppose W^t is the matrix of the word vectors, the optimization object is:

$$R^t = \arg\min \left\| QW^t - W^{t+1} \right\| \tag{2.1}$$

The main drawback of the alignment-based strategies is that the alignment introduces extra noise into the model, which is reason why we would like to introduce the temporal referencing strategy in the next section.

In terms of the problem, the alignment can be avoided by concentrating on word neighbors [6] and temporal referencing [4].

2.2.2 Temporal Referencing

The temporal referencing model [4] is designed to learn new representations of English words across different time periods. In this model, corpora from different time periods were merged into one large corpus. The trick to introduce time-specific information is to create a set of new words and inject them into the corpus. Suppose the target word is w and we would like to learn its meanings in three time periods t_1, t_2, t_3. Then three new words $w_{t_1}, w_{t_2}, w_{t_3}$ will be injected into the corpus by replacing the places of the original word w in the sentences. For instance, if the sentence "Tom ate w." appeared in t_2, the new sentence will be "Tom ate w_{t_2}".

In the temporal referencing, all the vectors are naturally located in the same vector space and can be compared without introducing additional alignment. Consequently, the outcome of the model is less noisy. What's more, temporal referencing is suitable for almost all vector space learning strategies because it is straightforward and has no extra requirements. The only assumption of the Temporal referencing is that the semantics of the context are stable over time, which is valid in most cases.

2.3 Semantic of Citations

Many works utilized embedding strategies to explore different aspects of publications. For instance, [5] and [20] mainly leveraged citation network to represent publications. In addition, [1] visualized publications by their citation context.

[14] studied the semantic of citations in the context of citation recommendation. In this paper, the semantic of citations were split into two genres: implicit semantics and explicit semantics. Implicit semantics referred to representing words by word embedding. The authors used the Word Mover's Distance between the word vectors to measure the similarity of that two words. Explicit semantics referred to grounded knowledge. The idea was to first identify the named entities in the documents, and then measure the similarity by some measures, such as Entity Overlap. In essence, the semantic of citation is a natural extension of the semantic of word, which treats citations as a special kind of words.

3 Dataset

In this section, we introduce the two datasets used in the thesis: unarXive [16] and GlUE. The unarXive contains all the information related to the publications; GlUE is introduced into the model to enhance the performance of the model.

3.1 unarXive

The unarXive dataset contains more than 1 million documents and 29.2 million citations. Every document represents the plain-text of one publication available on the arXiv.org. One citation is analogous to a link between two publications. Since one publication tends to cite many other publications, the number of citations in the dataset is much larger than the number of documents.

Besides the main information described before, the unarXive dataset also collected many kinds of auxiliary information (e.g. the publication date, discipline, ...) and saved them in a database, which was very helpful and convenient when we tried to select suitable publications for incipient analyses.

3.1.1 Statistics

Discipline: The distribution of the disciplines of the publications in the dataset was plotted in figure 3.1. Obviously, physics, math, statistics, and computer science were the major source of this dataset. In that, when exploring the semantic stability across different domains, we selected these four domains as our interested target.

Citation Frequency: The log-scale distribution of citation frequencies of all the publications was plotted in figure 3.2.

Obviously, most of the publications were cited less than 100 times. Since learning the

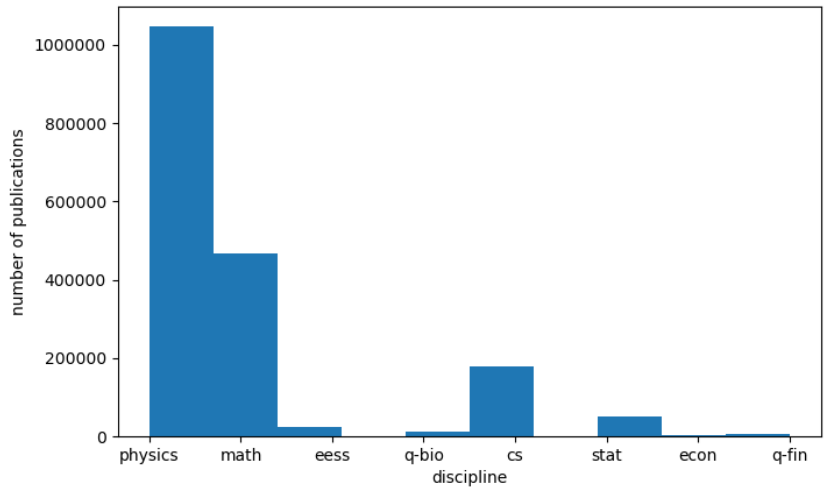

Figure 3.1: Discipline Distribution of all Publications

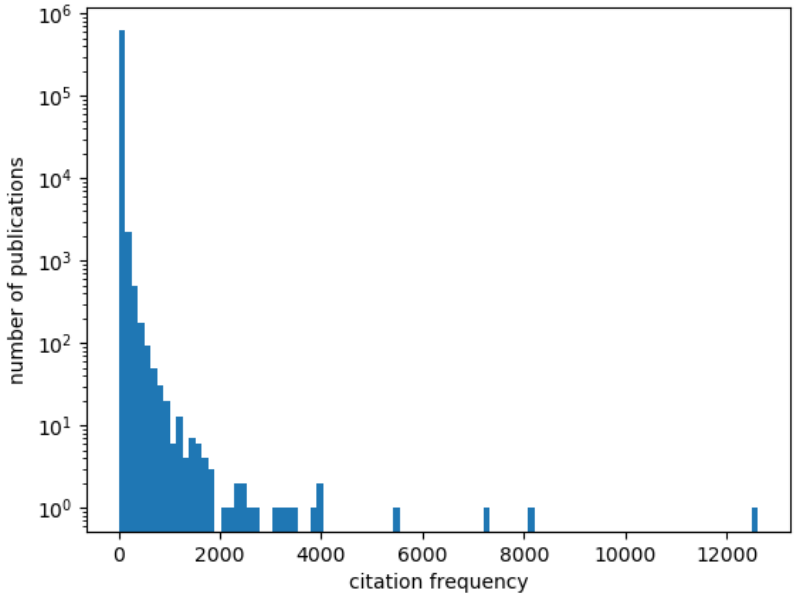

Figure 3.2: Citation Frequency of all Publications - log scale

semantics of publications required a large number of related documents, publications with low citation frequencies were not considered for experiments.

3.1.2 Database

Different aspects of the publications were collectively stored in the relational database and split into four tables:

1. **bibitem**
2. **bibitem link map**
3. **bibitem arxiv-id map**
4. **arxiv metadata**

The structures of the tables and the meaning of every column are listed in the following tables. The table REF need some more explanations. First, the full name of MAG is Microsoft Academic Graph which is a knowledge graph. It collected the metadata of a large number of scientific publications, such as author, publication venue ans so on. Second, in some cases either the citing-arxiv-id or the cited-arxiv-id column is empty. This is purely because the corresponding publication is not recorded by the arXiv.org website.

Table 3.1: bibitem link map

column name	meaning	example
id	identifier in this table	1
uuid	unique identifier of the citation	867b2e49-df59-4f65-90b4
link	website of the publication	http://stacks.iop.org/1538-3881

Table 3.2: bibitem arxiv-id map

column name	meaning	example
id	identifier in this table	1
uuid	unique identifier of the citation	17a18e49-d457-4dd5-9601
arxiv-id	identifier in the arXiv.org	2005.00707

Table 3.3: bibitem

column name	meaning	example
uuid	unique identifier of citation	adc127af-b947-444d-837c
citing-mag-id	MAG id of the citing paper	3025327795
cited-mag-id	MAG id of the cited paper	2981420411
citing-arxiv-id	arxiv-id of the citing paper	2005.05740
cited-arxiv-id	arxiv-id of the cited paper	None
bibitem-string	title of the cited paper	Mixed high-order attention...

Table 3.4: arxiv metadata

column name	meaning	example
arxiv-id	identifier in the arXiv.org	1109.3383
discipline	domain	physics:nucl-ex
date	publication date	2011-09-15

Since we are interested in the semantic stability of citations, there are two essential parts needed to be extracted from the database. The first part is the citing paper and the cited paper pairs. As the starting point, we used them to count the total number of citation frequencies of every publication in the database and set a threshold to select a set of popular publications. This is a necessary step because of two reasons. First, it is unrealistic to conduct experiments on all the 1 million publications recorded in the database. Second, only when one publication has occurred multiple times in the dataset can the Word2Vec model learn a stable representation of that publication. As was mentioned before, some records may have no arxiv-id. For this kind of records, we decided to omit them because the content of that publication was not available in the dataset. The second part is the discipline and the date information because we would like to understand the semantic of one publication in different time periods or domains.

Besides the indispensable information, other kinds of information are also of great help. For instance, the bibitem-string contains the title of that publication and thus we can have a rough idea about the content of it. In addition, we can compare it with the training result to see whether the model is useful or not.

3.2 GLUE

The introduction of the GLUE [21] dataset originated from the results of our experiments. In short, we would like to enhance the performance of our model by feeding general English to the model. The details will be left to the experiment section. Here we first introduced this widely used dataset.

The full name of GLUE is General Language Understanding Evaluation. It consists of multiple learning material designed for different NLP tasks. Those tasks covered a diverse range of text genres and degrees of difficulty, including CoLA, SST, QQP, MRPC, MNLI, RTE, WNLI, QNLI. We can use all the data from these tasks to train our model. Since different corpora had their own structures (e.g. some used sentence pairs, some had labels), we need to write independent codes for all of the tasks to extract our required English sentences.

4 Experiments and Results

In this section, we would like to describe the procedure of training the model and present the results of our experiments.

4.1 Training

The training procedure can be divided into two steps:

1. use the temporal referencing model to inject new words
2. use the skip-gram model with negative-sampling [13] to learn vector representations of the words

Before training the model, we first need to select several popular publications to conduct early-stage experiments. The concrete practice is to query the bibitem table, load all the (citing-arxiv-id, cited-arxiv-id) pairs into the memory, and construct a hashmap to count the total number of citation frequencies of every publication. One minor flaw is that this procedure took a long time. We speculated that there were over 29.2 million records in this table and frequently querying the traditional relational database may become the bottleneck of the algorithm. In addition, the (citing-arxiv-id, cited-arxiv-id) pairs might be reused for subsequent tasks. So we extracted all the pairs from the database and saved them on the disk before constructing the hashmap.

The criterion of selecting publications for experiments was simple: the publications must be cited at least 1 thousand times in total and at least 10 times every year. There were 19 papers qualified.

4.1.1 New Citation Representation

Next, we need to emulate what temporal referencing did to create our custom new words. First, we used the arXiv-id of every publication to represent itself. This is because all

the arXiv-ids are unique and does conflict with existing English words. To avoid being separating to two parts by the sentence tokenizer and the word tokenizer, all the dots and hyphens were removed from the arXiv-ids (e.g. "1603.04467" became "160304467", "hep-th0405231" became "hepth0405231").

Second, we concatenated the processed arXiv-id of the cited paper with the publication year of the citing paper. By this practice, we merged the time-specific information into the publication. Since the the arXiv-ids ended with numbers and the years are pure numbers, we added a string "yyyy" between the arXiv-id and the year when concatenating these two parts. Thus, we were able to separate these two kinds of information when analyzing the results of the experiments. In addition, since "yyyy" is not an English word, it can serve as the marker of a citation and help us quickly locate the corresponding citation in the document. In that, the processed citations were of the following form: 160304467yyyy2018.

We were also interested in the semantic stability of publications across different domains because researchers from different domains were likely to cite the same publication for distinct purposes. This aim can be achieved by concatenating the arXiv-id of the cited paper with the discipline of the citing paper, rather than the publication date.

The reason why this straightforward process could take effect was that the Word2Vec model did not really understand the meanings of English words. In other words, it treated both valid English words and meaningless strings as normal words remained to be learned.

4.1.2 Processing Documents

The next step was to substitute the original citations with the new citations in the documents. At the same time, we need to process the documents to be the input of the Word2Vec model. The details were can be split in the following steps:

1. We utilized the Word2Vec model from the Gensim library [15] and the model required that the words in one sentence should be separated and ordered in a list. So we first used the sentence tokenizer from the nltk library [2] to separate sentences in documents.

2. We searched the citation pattern "{{cite:uuid}}" by a regular expression in every sentence. As long as a citation was found, our algorithm extracted the uuid from the citation string and query the simplified database for the corresponding (citing-arXiv-id, cited-arXiv-id)) pair.

3. The year information of the citing paper was retrieved from the database and then the new citation representation was generated as was described before.

4. The matched citation string was replaced by its new representation.

Finally, we utilized the word tokenizer to identify words in every sentence and filter out useless words to keep the corpus as clean as possible. After all theses steps, the document were ready for the Word2Vec model.

4.1.3 Removing Unnecessary Words

For every publication, most of the original content was preserved in its document: the title, the authors, the abstract and the body. However, all the citations were removed. Instead, they were transformed into the following form: {{cite:b8a5f32c-7c79-41b6-a6b4-fe9a24f22d16}}, where the string of numbers represents the uuid of the citation. In addition, all the mathematics formulas were replaced by the word "FORMULA".

Although most of the plain-text was identical to the original version, there were still some problems in the document and thus pre-processing was required.

Mistakes: The content of every publication was downloaded from the website and the document was generated by the Python scripts. Some problems originated from the imperfection of the Python scripts, such as spelling mistakes and remained latex codes such as "=1pt".

Redundant Information: One the other hand, some valid but not useful words should also be removed to keep the dataset as clean as possible. For instance, pure numbers and dates were of little meanings. In addition, punctuation symbols and special symbols (e.g. '+', '-') were also not helpful to understand the meanings of citations.

Unrelated Publications When processing documents, we started from one popular publication, queried the citation database, and process the document corresponding to the publication citing this popular publication. However, all of the citing papers did not only cite the popular paper, but also cited many other papers. However, these publications were not helpful in understanding the meaning of the popular publication. So we removed all the unrelated publications from the documents.

There is no straightforward method to deal with the spelling mistakes (e.g. distinguish between "spelling" and "spleling"). However, For the rest of the problems, we created several regular expressions to precisely find out target patterns without touching other words.

Authors: The author's information also belongs to redundant information but was especially difficult to tackle. The first possible solution was to resort to the bibitem-string in the bibitem table because the author information was already included in it. However, this method was infeasible because other kinds information such as the discipline and the title were also mixed in the bibitem-string, separated by different delimiters (e.g. colons, dots).

Considering only the authors of the cited paper will repeatedly appeared in the documents of the citing papers, the second option was to directly read the bibitem-string of the cited paper from the database and manually extract the author information from it. Since the total number of publications was decreased to a relatively low number, this unscalable compromise was acceptable. However, it turned out that the results of the experiments were still noisy because many authors' names were identical. In other words, if we did not remove all the authors in the documents, we effectively introduced many polysemous words into the dataset, which made the model more difficult to train.

Finally, we found out that we had access to the arxiv-id, the title, the author, the venue, the url, and the category of every publication available on the website by visiting the URL https://arxiv.org/abs/arXiv-id. In that,as the ultimate solution, we wrote a Python script to collect all the author information of the publications. However, after retrieving about several hundred records, our visits were permanently forbidden by the website. By visiting the website in person, we found out that our behavior had triggered the automatic robot detection system. One good news was that arXiv.org provided

https://export.arxiv.org/ to the public for automated programmatic harvesting purpose. After switching to the backup website, our script could work normally again.

4.2 Experiment Settings

4.2.1 Building Corpus

Ideally, the Word2Vec model should be trained on all of 1 million documents to achieve the theoretically best performance. However, this was infeasible because both pre-processing that many documents and training a Word2Vec model based on these documents were too time-consuming. Alternatively, we came up with an idea to build a smaller corpus and then conduct our experiments.

First, we selected a set of publications by the criterion described before (cited at least 1,000 times). Then, for every popular publication, all the publications that had cited it before were included into the corpus. After enumerating all the popular publications, we obtained a medium-sized corpus, where all the documents in the corpus were closely related to at least one of the popular publications. In addition, we were able to learn the word vectors of the popular publications at the same time, which was convenient when we had verified that our algorithm was correct and planned to conduct experiments on more publications. However, one potential problem was that we cannot tell how much we changed the distribution of English in advance because we introduced many "fade" words (i.e. processed citations such as 160304467yyyy2018) into the corpus.

4.2.2 Pre-trained Model

After running experiments on over twenty publications, we did not find any obvious semantic shifts in any of the publications. The phenomenon may originate from the two reasons. First, the semantic of the publications has not changed in the first place. For instance, that paper was recently published, which means there is not enough time for semantic shift to happen because it typically takes a long time to happen. On the other hand, our model may fail to capture to semantic shifts. After a series of consideration,

we came up with an idea to enhance the capability of our model: we decided to first feed general English sentences to the model. The idea is to let the model learn English and thus to make sure that the model is able to understand the content of the publications correctly. Then we feed the academic sentences to the model.

Up to now, the corpus was completely constructed by academic documents. As we all know, the expression of academic articles was very different from the daily expression and the words allowed to be used in scientific publications were very limited. In that, many daily words seldom appeared in the corpus and thus the model may have little chance to learn the meanings of them.

To alleviate this phenomenon, we considered adding some common English material to the existing corpus. When searching for suitable English dataset, the criterion was that the dataset should cover as many topics as possible. By a series of comparisons, we finally adopted the GLUE dataset to build an initial model and continued training the model by feeding scientific documents to it.

Except for the GLUE dataset, We also considered using pre-trained word vectors to build the initial model because there were many excellent works. However, we need both the learned word vectors and the weights of the neural network to continue training the model, but the weights were not available from the Internet. So this idea was inapplicable.

4.2.3 Other Settings

We created four groups of experiments to explore the semantic stability of citations in different situations:

Table 4.1: Settings of Word2Vec

Group Number	Pre-trained Model	Aspect
I	Yes	over time
II	Yes	across discipline
IV	No	over time
V	No	across discipline

Finally, the settings of the Word2Vec model were almost same across different experiment setting. The details were listed in the following table:

Table 4.2: Settings of Word2Vec

Parameter	Meaning	Value
epochs	training iterations	5
min-count	minimum number of occurrence	50
dimension	dimension of word vector	300
window	max distance from predicted word	5

4.3 Results

In this part, we presented the results of our experiments. We selected the 19 publications such as 1603.04467, 1312.6114, 1512.03385, and 1502.03167 as the target publications. They were cited over 1,000 times in total and cited by other publications from both computer science and physics over 100 times. By including and processing all the publications that cited one of these popular papers, we built a corpus consisting of over 30,000 documents. In addition, 1312.6114 and 1512.03385 were cited by papers from physics, math, statistics, and computer science over 100 times, so we would like to mainly present the results of these two publications and skip the rest of the publications for the sake of concision. Finally, it was worthy to mention that the similarity between two word vectors were measured by cosine distance.

4.3.1 Semantic shifts over Time

First we took a look at how the publications themselves changed as time went by. To be more specific, we selected the word vector corresponding to the publication year of one publication, and then compared it with all the word vectors in other years. The results were plot in figure 4.1.

Next, we tried to understand and analyze the semantic of citations by their neighbors. The TOP 5 closest word vectors to the one of publications were listed in the tables 4.3, 4.4, 4.5, and 4.6.

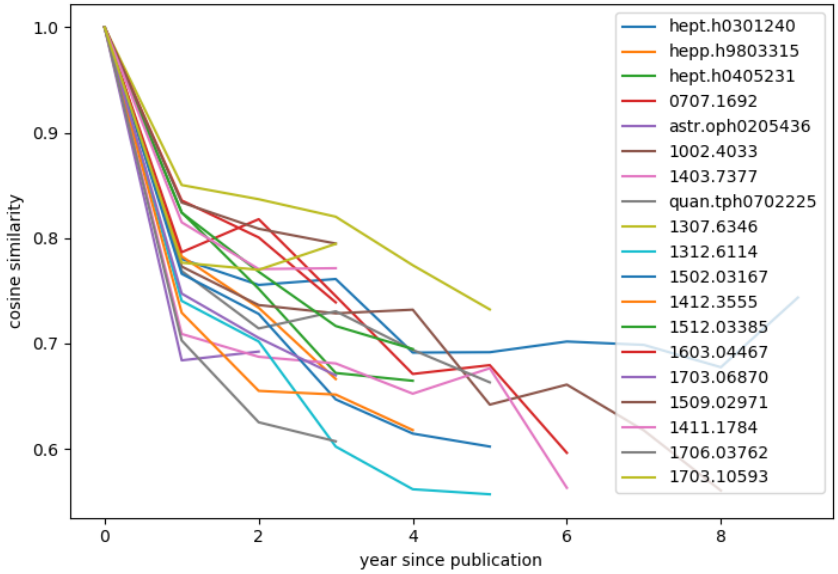

Figure 4.1: Changes of Cosine Similarity since Publication

Table 4.3: 1312.6114 without Pre-training

Year	First	Second	Third	Fourth	Fifth
2015	kingma	nvil	welling	importance-weighted	sgvb
2016	aaes	importance-weighted	kingma	vampprior	nvil
2017	importance-weighted	kingma	variational	vae	vampprior
2018	variational	importance-weighted	kingma	vaes	vae
2019	variational	vae	kingma	importance-weighted	vaes
2020	variational	importance-weighted	kingma	vaes	vae

Table 4.4: 1312.6114 with Pre-training

Year	First	Second	Third	Fourth	Fifth
2015	nvil	sgvb	bbvi	bbb	kingma
2016	aaes	kingma	vampprior	importance-weighted	vaes
2017	kingma	vaes	variational	importance-weighted	vae
2018	importance-weighted	kingma	variational	vaes	aaes
2019	variational	kingma	vaes	vae	importance-weighted
2020	variational	vaes	vae	importance-weighted	aaes

Table 4.5: 1512.03385 without Pre-training

Year	First	Second	Third	Fourth	Fifth
2016	googlenet	50-layer	110-layer	vggnet	inception-v4
2017	vggnet	googlenet	50-layer	vgg-net	inception-v4
2018	vggnet	50-layer	googlenet	18-layer	vgg-net
2019	vggnet	50-layer	googlenet	18-layer	resnet
2020	vggnet	50-layer	resnet	18-layer	googlenet

Table 4.6: 1512.03385 with Pre-training

Year	First	Second	Third	Fourth	Fifth
2016	152-layer	inception-v4	50-layer	110-layer	34-layer
2017	vggnet	googlenet	inception-v4	50-layer	152-layer
2018	vggnet	50-layer	152-layer	101-layer	googlenet
2019	vggnet	50-layer	googlenet	152-layer	vgg-net
2020	vggnet	50-layer	resnet34	resnet	googlenet

4.3.2 Semantic shifts across Disciplines

In this section, we represented the results of how semantic shifted across different domains in the tables 4.7, 4.9.

Table 4.7: 1312.6114 without Pre-training

Discipline	First	Second	Third	Fourth	Fifth
stat	variational	sgvb	importance-weighted	kingma	non-conjugate
math	aaes	vampprior	avb	auto-encoders	vaes
cs	vae	auto-encoders	vaes	aaes	kingma
phy	variational	importance-weighted	aaes	kingma	auto-encoders

4.3.3 Analysis and Discussion

By looking at how one publication itself changed over time, we can observe that the semantics of publications did shifted. For instance, the cosine similarities of 1312.6114 changed from 1.0 to 0.6. In addition, the overall tendencies of all publications were decreasing every year. Note that all the lines started from point 1 because the cosine similarity between

Table 4.8: 1312.6114 with Pre-training

Discipline	First	Second	Third	Fourth	Fifth
stat	variational	sgvb	importance-weighted	kingma	non-conjugate
math	aaes	odeformula	lgae	sivae	jvae
cs	vaes	auto-encoders	vae	aaes	variational
phy	kingma	variational	aaes	welling	sivae

Table 4.9: 1512.03385 without Pre-training

Discipline	First	Second	Third	Fourth	Fifth
cs	vggnet	resnet	googlenet	vgg-net	vgg
math	50-layer	18-layer	34-layer	110-layer	resnet32
stat	50-layer	18-layer	resnet32	resnet	110-layer
phy	50-layer	18-layer	34-layer	110-layer	resnet

Table 4.10: 1512.03385 with Pre-training

Discipline	First	Second	Third	Fourth	Fifth
cs	vggnet	googlenet	resnet	50-layer	152-layer
math	50-layer	18-layer	32-layer	101-layer	152-layer
stat	50-layer	resnet34	18-layer	34-layer	vggnet
phy	50-layer	152-layer	18-layer	34-layer	101-layer

one vector and itself is always 1.

Although we achieved positive results from the previous analysis, the outcome of the closest neighbors seemed to provide the opposite evidence. No matter we focused on the semantic shifts over time or across domains, there were no obvious difference. For instance, the top 5 closest neighbors of 1312.6114 in different years were almost identical and the slight difference was that the order of these words was changing every year.

The results seemed to be problematic, but we can prove that they should be correct. In fact, the closest neighbors in the tables were proceeded. In the first place, many of the words close to our target publications were not English words, but publications. Furthermore, in most cases, the closest word to a certain publication was itself in another year. This was an surprising but reasonable observation because the meaning of a publication was so complex that no words other than itself could summarize it meaning.

On one hand, this observation indicated that the model actually captured the meanings of publications. On the other hand, this result was not very meaningful for interpreting the results. As was mentioned before, a word vectors was just an array of numbers so we cannot directly interpret the outcome of the Word2Vec model. Instead, we could try understanding a word vector by its similar neighbors. However, we can hardly learn anything about a publication by its publication neighbors. In that, when presenting the results of the experiments, we skipped all the publication and only retained valid English words.

One possible explanation to these seemly conflicting outcomes was that the semantics of publications did shift to some extent but the closest neighbors did not capture this aspect. In addition, one potential drawback of our conclusion is that we did not analyze papers that were published over ten years so it may only apply to recent publications.

Finally, by comparing the results between not-pre-trained and pre-trained models of the same publications, we can find out that the results were still similar. One slight difference of 1512.03385 was that the pre-trained model paid more attention to the structure of the neural network.

5 Conclusion

In this thesis, we combined the temporal referencing strategy and the skip-gram model to encode the semantics of citations, and utilized the cosine distance to evaluate the semantic shifts in citations. On one hand, by looking at how citations themselves changed over time, we found out that the semantics of citations did change as time went by. However, on the other hand, the results of the closest neighbors to citations indicated that the semantics of citations tend to be stable over time, which was similar to the conclusion drawn in [8]. In addition, the same phenomenon also happened to the cross-domain citations. We speculated that closest neighbors might not be the optimal method to evaluate semantic shifts in citations, which need to be verified in the future. Furthermore, all of our analyses are based on cosine distance, on possible improvement for future work is to leverage other measurements and see whether our conclusion is still valid.

Bibliography

[1] M. Berger, K. McDonough, and L. M. Seversky. "cite2vec: Citation-driven document exploration via word embeddings". In: *IEEE transactions on visualization and computer graphics* 23.1 (2016), pp. 691–700.

[2] S. Bird, E. Klein, and E. Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

[3] J. A. Bullinaria and J. P. Levy. "Extracting semantic representations from word co-occurrence statistics: stop-lists, stemming, and SVD". In: *Behavior research methods* 44.3 (2012), pp. 890–907.

[4] H. Dubossarsky, S. Hengchen, N. Tahmasebi, and D. Schlechtweg. "Time-out: Temporal referencing for robust modeling of lexical semantic change". In: *arXiv preprint arXiv:1906.01688* (2019).

[5] S. Ganguly and V. Pudi. "Paper2vec: Combining graph and text information for scientific paper representation". In: *European conference on information retrieval*. Springer. 2017, pp. 383–395.

[6] W. L. Hamilton, J. Leskovec, and D. Jurafsky. "Cultural shift or linguistic drift? comparing two computational measures of semantic change". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*. Vol. 2016. NIH Public Access. 2016, p. 2116.

[7] W. L. Hamilton, J. Leskovec, and D. Jurafsky. "Diachronic word embeddings reveal statistical laws of semantic change". In: *arXiv preprint arXiv:1605.09096* (2016).

[8] J. He and C. Chen. "Understanding the changing roles of scientific publications via citation embeddings". In: *arXiv preprint arXiv:1711.05822* (2017).

[9] P. Koch. "Meaning change and semantic shifts". In: *The Lexical Typology of Semantic Shifts* 58 (2016), p. 21.

[10] V. Kulkarni, R. Al-Rfou, B. Perozzi, and S. Skiena. "Statistically significant detection of linguistic change". In: *Proceedings of the 24th International Conference on World Wide Web*. 2015, pp. 625–635.

[11] O. Levy, Y. Goldberg, and I. Dagan. “Improving distributional similarity with lessons learned from word embeddings”. In: *Transactions of the Association for Computational Linguistics* 3 (2015), pp. 211–225.

[12] T. Mikolov, K. Chen, G. Corrado, and J. Dean. “Efficient estimation of word representations in vector space”. In: *arXiv preprint arXiv:1301.3781* (2013).

[13] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. “Distributed representations of words and phrases and their compositionality”. In: *arXiv preprint arXiv:1310.4546* (2013).

[14] H. Peng, J. Liu, and C.-Y. Lin. “News citation recommendation with implicit and explicit semantics”. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2016, pp. 388–398.

[15] R. Rehurek and P. Sojka. “Gensim–python framework for vector space modelling”. In: *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic* 3.2 (2011).

[16] T. Saier and M. Färber. “unarXive: a large scholarly data set with publications’ full-text, annotated in-text citations, and links to metadata”. In: *Scientometrics* 125.3 (2020), pp. 3085–3108.

[17] P. H. Schönemann. “A generalized solution of the orthogonal procrustes problem”. In: *Psychometrika* 31.1 (1966), pp. 1–10.

[18] U. S. Semantics. *An introduction to the science of meaning*. 1962.

[19] J. Sivic and A. Zisserman. “Video Google: A text retrieval approach to object matching in videos”. In: *Computer Vision, IEEE International Conference on*. Vol. 3. IEEE Computer Society. 2003, pp. 1470–1470.

[20] H. Tian and H. H. Zhuo. “Paper2vec: Citation-context based document distributed representation for scholar recommendation”. In: *arXiv preprint arXiv:1703.06587* (2017).

[21] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman. “GLUE: A multi-task benchmark and analysis platform for natural language understanding”. In: *arXiv preprint arXiv:1804.07461* (2018).