Text Mining Approaches for Semantic Similarity Exploration and Metadata Enrichment of Scientific Digital Libraries

Hussein T Al-Natsheh

Thèse de doctorat de l'Université de Lyon
opérée par
l'Université Lumière Lyon 2

Laboratoire ERIC (EA 3083)
École Doctorale Informatique et Mathématiques (ED 512)

présentée pour obtenir le grade de
Docteur en Informatique

Text Mining Approaches for Semantic Similarity Exploration and Metadata Enrichment of Scientific Digital Libraries

---

Par : Hussein T. Al-Natsheh

Présentée et soutenue publiquement le 15 février 2019, devant un jury composé de :

Nathalie Aussenac-Gilles, Directrice de Recherche, IRIT Toulouse	Examinatrice
Younès Bennani, Professeur des Universités, Université Paris 13	Examinateur
Juliette Dibie, Professeure des Universités, AgroParisTech	Rapportrice
Sabine Loudcher, Professeure des Universités, Université Lumière Lyon 2	Examinatrice
Fabrice Muhlenbach, Maître de Conférences, Université Jean Monnet Saint-Etienne	Co-directeur
Gilles Venturini, Professeur des Universités, Université de Tours	Rapporteur
Djamel Abdelkader Zighed, Professeur des Universités, Université Lumière Lyon 2	Directeur

Abstract

For scientists and researchers, it is very critical to ensure knowledge is accessible for re-use and development. Moreover, the way we store and manage scientific articles and their metadata in digital libraries determines the amount of relevant articles we can discover and access depending on what is actually meant in a search query. Yet, are we able to explore all semantically relevant scientific documents with the existing keyword-based search information retrieval systems? This is the primary question addressed in this thesis. Hence, the main purpose of our work is to broaden or expand the knowledge spectrum of researchers working in an interdisciplinary domain when they use the information retrieval systems of multidisciplinary digital libraries. However, the problem raises when such researchers use community-dependent search keywords while other scientific names given to relevant concepts are being used in a different research community.

Towards proposing a solution to this semantic exploration task in multidisciplinary digital libraries, we applied several text mining approaches. First, we studied the semantic representation of words, sentences, paragraphs and documents for better semantic similarity estimation. In addition, we utilized the semantic information of words in lexical databases and knowledge graphs in order to enhance our semantic approach. Furthermore, the thesis presents a couple of use-case implementations of our proposed model. Finally, several experimental evaluations were conducted to validate the efficiency of the proposed approach. The results of the hybrid approach, based on the short text semantic representation and the word semantic information extracted from lexical databases, were very encouraging. We believe that our new proposed approaches based on text mining techniques practically achieved the expected results in addressing the limitation of semantic exploration in the classical information retrieval systems of digital libraries.

The advantage of our approach is that it is applicable to large-scale multidisciplinary digital libraries. In that sense, we use information found in the metadata of such libraries in order to enrich it with additional semantic tags. As a consequence, the enhanced and enriched metadata enable researchers to retrieve more semantically relevant documents that would have otherwise remained unexplored without the enrichment.

We think that our study and proposed approaches will provide practical solutions to knowledge access and contribute to the research communities and fields of text mining and data management in digital libraries.

Keywords: information retrieval, semantic similarity, metadata enrichment, text mining, entity name disambiguation, knowledge management, digital library.

Résumé

Pour les scientifiques et chercheurs, s'assurer que la connaissance est accessible pour pouvoir être réutilisée et développée est un point crucial. De plus, la façon dont nous stockons et gérons les articles scientifiques et leurs métadonnées dans les bibliothèques numériques détermine la quantité d'articles pertinents que nous pouvons découvrir et auxquels nous pouvons accéder en fonction de la signification réelle d'une requête de recherche. Cependant, sommes-nous en mesure d'explorer tous les documents scientifiques sémantiquement pertinents avec les systèmes existants de recherche d'information au moyen de mots-clés ? Il s'agit là de la question essentielle abordée dans cette thèse. L'objectif principal de nos travaux est d'élargir ou développer le spectre des connaissances des chercheurs travaillant dans un domaine interdisciplinaire lorsqu'ils utilisent les systèmes de recherche d'information des bibliothèques numériques multidisciplinaires. Le problème se pose cependant lorsque de tels chercheurs utilisent des mots-clés de recherche dépendant de la communauté dont ils sont issus alors que d'autres termes scientifiques sont attribués à des concepts pertinents lorsqu'ils sont utilisés dans des communautés de recherche différentes.

Afin de proposer une solution à cette tâche d'exploration sémantique dans des bibliothèques numériques multidisciplinaires, nous avons appliqué plusieurs approches de fouille de texte. Tout d'abord, nous avons étudié la représentation sémantique des mots, des phrases, des paragraphes et des documents pour une meilleure estimation de la similarité sémantique. Ensuite, nous avons utilisé les informations sémantiques des mots dans des bases de données lexicales et des graphes de connaissance afin d'améliorer notre approche sémantique. En outre, la thèse présente quelques implémentations de cas d'utilisation du modèle que nous avons proposé. Enfin, plusieurs évaluations expérimentales ont été menées afin de valider l'efficacité de notre approche. Les résultats de l'approche hybride, basée à la fois sur une représentation sémantique de petits textes et sur l'information sémantique des mots extraits de bases de données lexicales, ont été très encourageants. Nous pensons que nos nouvelles approches basées sur les techniques de

fouille de texte permettent d'obtenir en pratique les résultats escomptés en ce qui concerne la limitation de l'exploration sémantique dans les systèmes classiques de recherche d'information des bibliothèques numériques.

L'avantage de notre approche est qu'elle s'applique aux grandes bibliothèques numériques multidisciplinaires. En ce sens, nous utilisons les informations trouvées dans les métadonnées de ces bibliothèques afin de les enrichir de balises sémantiques supplémentaires. Par conséquent, les métadonnées améliorées et enrichies permettent aux chercheurs de récupérer des documents plus pertinents d'un point de vue sémantique qui seraient autrement restés inexplorés sans cet enrichissement.

Nous pensons que notre étude et les approches que nous proposons fourniront des solutions pratiques à l'accès aux connaissances et contribueront aux communautés de recherche et aux domaines de la fouille de texte et de la gestion des données dans les bibliothèques numériques.

Mots clés

Recherche d'information, similarité sémantique, enrichissement de métadonnées, fouille de texte, désambiguïsation d'entités nommées, gestion de la connaissance, bibliothèque numérique.

Acknowledgments

First off, I would like to thank the thesis directors who provided me with their wise, expertize, time, efforts as well as trust, strength and self-confidence along the whole PhD journey. I would like to name each of: My supervisor Dr. Fabrice Muhlenbach who guided me not only in the process of writing the thesis, but also though out all the research projects I worked on during the PhD studies, Prof. Djamel Abdelkader Zighed, was of a great inspiration and support as the director of the thesis, Dr. Lucie Martinet, as a supportive colleague and a co-author of most of the publications during the PhD studies as well as Dr. Fabien Rico as a supporting supervisor in the most of the research projects we worked on. I would also like to thank Dr. Patrick Fargier and Prof. Raphaël Massarelli for their kind efforts in the use case study of the scientific topic corpus expansion. I would also like to thank the reporters Prof. Juliette Dibie and Prof. Gilles Venturini for reviewing the thesis and their constructive feedback. I am also very thankful for all the the rest of the jury members; Prof. Nathalie Aussenac-Gilles, Prof. Younès Bennani and Prof. Sabine Loudcher for their time and kind efforts.

Special thanks to all of my collogues in ERIC laboratory including the director, the secretariat and the head of the DMD team as well as all the faculty members and other PhD students. Same goes to the doctoral school *infomaths* including the director and the secretariat. I would also like to thank all my colleagues in MSH Lyon St-Etienne. I would also like to thank all of my previous teachers since elementary school till today. To my precious wife, Karima and my beloved children Lya and Loay bearing with me in my home absence. To my parents, brothers and sisters. To all of my friends.

I also want to thank the project ARC6 of the French region Auvergne-Rhone-Alpes for funding my PhD studies.

Contents

List of Figures

List of Tables

Abbreviations

BoW	**B**ag **o**f **W**ords
CBOW	**C**ontinuous **B**ag **O**f **W**ords
Doc2Vec	**Doc**ument (to) **Vec**tor
GRU	**G**ated **R**ecurrent **U**nit
HCI	**H**uman-**C**omputer **I**nteraction
IR	**I**nformation **R**etrieval
ISTEX	**EX**cellence **I**nitiative of **S**cientific and **T**echnical Information
LDA	**L**atent **D**irichlet **A**llocation
LSA	**L**atent **S**emantic **A**nalysis
pLSA	**p**robabilistic **L**atent **S**emantic **A**nalysis
MSRPC	**M**icro**S**oft **R**esearch **P**araphrasing **C**orpus
MLT	**M**ore **L**ike **T**his
NER	**N**amed **E**ntity **R**ecognition
NMF	non-**N**egative **M**atrix **A**nalysis
POS or **PoS**	**P**art **o**f **S**peech
RNN	**R**ecurrent **N**eural **N**etwork
SemEval	**Sem**antic **Eval**uation
SDL	**S**cientific **D**igital **L**ibrary
SenSim	**Sen**tence **Sim**ilarity
SSbE	**S**emantic **S**earch-**b**y-**E**xamples
STS	**S**emantic **T**extual **S**imilarity
Synset	**Syn**onym **set**
SVD	**S**ingular **V**alue **D**ecomposition
TruncatedSVD	**Truncated S**ingular **V**alue **D**ecomposition
Word2Vec or **W2V**	**Word** (to) **Vec**tor

Chapter 1

Introduction

1.1 Background

Since the beginning of knowledge sharing, people have invented and used writing as the most common way to communicate and store knowledge. Most of the human accumulated knowledge has been stored in a textual format in all types of books, emails, web-pages or blogs, news articles or scientific journals. Writing text is an encoding form of the spoken language that could be considered as the easiest way to represent human communication. With the writing starts the History, the possibility to have law codification and to maintain original sacred text against distortion or infringement.

Collections of written texts (papyrus scrolls, books) have been contributing to the emergence of schools and scholars as well as intellectual development of societies. In the Ancient Times, the world had witnessed the establishment of the libraries of Alexandria and Pergamum that are considered as important centres of knowledge sharing. Later, technology has had an important influence on knowledge sharing. Thanks to the paper making techniques developed by the Chinese, financial spending decisions of rulers and the work of copyists and translators, influential libraries also appeared in the "Islamic Golden Age"[1], for instance the "House of Wisdom"[2]. Afterwards, technology continued to influence the circulation of information and the rapid transmission of new ideas with

[1] `http://islamichistory.org/islamic-golden-age/`
[2] `https://en.wikipedia.org/wiki/House_of_Wisdom`

Gutenberg printing press which led to the "European Renaissance"[3]. Knowledge sharing has been a key indicator of nations advancement. The wealth of nations is not only measured by its economy level but also by the science and knowledge development (May, 1997).

Today, most of the data are stored and indexed in the digital format and are mostly available through the World Wide Web (WWW), also called "the Web." Tim Berners-Lee[4] (currently the director of World Wide Web Consortium (W3C)) invented the Web in 1989, and was later the first to introduce the idea of "the next web", "web of data", "linked data" or the "semantic web" (Berners-Lee, Hendler, and Lassila, 2001). Linked data is currently one of the best known approaches to store, represent knowledge and have it accessible by people. Since then, the world has been facing a digital revolution of information access and knowledge sharing, thus opening a door for a "new Renaissance." The increasing capacity of storage, computing and networking in the internet is producing a massive digitalized data that humanity had never witnessed before.

The concept of digital libraries was raised in the end of the last century. It generally means a collection of books, or any type of documents, stored and indexed in its digital format. Human sciences stored in digital libraries are not an exception in the digital revolution where digital libraries collect many resources from different disciplines and publishers. The number of authors and their scientific publications is also increasing rapidly, thanks to the digitalized publication and the continuous public and private scientific research funds. However, digital libraries are now facing a lot of data management and accessibility issues. Not only specialized data management companies, but also research community in computer and data science are seeking solutions to such challenges.

1.2 Motivation and Problem Statement

Digital libraries are now becoming more interdisciplinary and their number of documents keeps on increasing rapidly. Recently, new concepts as open-access and open-knowledge have emerged where we started to witness many open-access digital libraries. However,

[3] `https://en.wikipedia.org/wiki/Renaissance`
[4] `https://en.wikipedia.org/wiki/Tim_Berners-Lee`

managing and retrieving the knowledge stored in these digital libraries poses many challenges. For instance, digital libraries are not centralized and there is not any globally uniformed data schema neither among publishers nor the different scientific disciplines.

Being a human generated text, the knowledge is represented in a natural language that is not yet understood by the machine as for programming languages. Another challenge caused by using text to represent natural language is the fact that there is no perfect match between words and thoughts (Evans, 2006; Sapir, 1985). A thought could be expressed in many ways using different possible syntax forms and different words. At the same time, one word could have several meanings. Classical information retrieval systems like typical text-matching search engines are semantically insufficient due to these later challenges.

Following in this section, we will go through some of these challenges faced by multi-disciplinary digital libraries in knowledge management and information retrieval. The main motivation behind this thesis is to solve these key challenges using data mining and knowledge management techniques.

1.2.1 Using Metadata, Keywords, Tags and Ontologies

The concept of open-access digital libraries has been adopted by many organizations as a response to the increasing demand for researchers to open-access science. Many of such open-access initiatives are aggregating publications from many publishers. However, there is no standard metadata for different publishers, even for the open-access ones. Mapping different metadata from the publisher to the combined data schema is an open problem that needs to be tackled by multi-source digital libraries. For example, the categorization of scientific subjects differs, whether in "Web of Science"[5] or "Scopus"[6]. The naming and numbering of main categories and subcategories are different.

There are different ways to provide complementary information, from unstructured data to structured data. The structuring is usually provided by a metadata that uses tags or keywords chosen by the authors of scientific papers. When submitting their scientific articles, authors are usually limited to a maximum number of tags and keywords per document. In many cases, the research work concerns many scientific disciplines and

[5] `https://clarivate.com/products/web-of-science/`
[6] `https://www.elsevier.com/solutions/scopus`

contains many variations of topic naming which all together require a larger number of keywords.

In some cases, publisher provided a reduced number of standard keywords or category names provided by the scientific societies, like in the ACM Computing Classification System[7], or the IEEE Taxonomy[8]. It becomes a critical issue when the information retrieval system uses above mentioned list of keywords to filter and rank the documents. If the querying researcher uses a single variation or related keyword that is not mentioned in the keyword set of the publication, that publication would be excluded from the results. For example, if the researcher utilize the keyword "Search Engines" while the publication is only tagged with this list of keywords: ["information retrieval," "text mining," "computational linguistics" and "content-based recommender systems"], the publication would likely not appear in the search results.

We also have to deal with this limitation in cases where diverse scientific disciplines tend to use different terms to describe the same topic. For instance, in computer science, we use the term "Machine Learning" while the same concept is found to be referred to "Multivariate Data Analysis" by the high energy physics community. Even within the same scientific community, the same topic or concept is expressed using various terms or names over time (e.g., "data mining" and "data science").

In computer science, ontology is a kind of schema free knowledge representation. Unlike relational and document databases, a network of connected facts organized in a triple format (subject, predicate, object) is used to represent the knowledge in a semantic way. Developing and using such knowledge graph in more difficult and expensive than in traditional databases. A typical example of an open-access knowledge graph is DBpedia (Lehmann et al., 2015), which was built using Resource Description Framework (RDF)(Miller, 1998), as a Semantic Web standard. DBpedia was extracted from infoboxes of the digital encyclopedia Wikipedia.

Dealing with knowledge graph or linked data (Wood et al., 2014) involves a lot of standards in defining the vocabulary, concept name conventions, semantic linkage naming, entity name disambiguation, storage, open-access data management and of course the query language and methods. A well-known and widely-used ontology is *schema.org*

[7] `https://www.acm.org/publications/class-2012`
[8] `https://www.ieee.org/content/dam/ieee-org/ieee/web/org/pubs/taxonomy_v101.pdf`

(Guha, Brickley, and Macbeth, 2016). The main challenge of linked data lies in the fact that it is being incomplete (Nickel et al., 2016). Collecting all mankind knowledge (multi-lingual, multi-domain) in a centralized knowledge graph is an infinite work itself, especially when taking into consideration its content update over time. Even if most of the process was automated, manual data curation and annotation would still be needed, which is costly and difficult to maintain. Thus, using unstructured data is still needed besides using such structured knowledge sources. In this context, text mining techniques play an important role in dealing with various source of plain text or textual data without any type of semantic annotation.

1.2.2 Promoting Interdisciplinary Research

Recently, the science community has been witnessing more and more of new emerging interdisciplinary scientific domains. In many cases, computer science becomes a cross-cutting field that has been used in many other disciplines helping in scientific discoveries as well as managing storage and computing for scientific experiments (e.g., high energy physics or gene studies in biology). Bioinformatics is an example of such cross-domain: It has been emerged out of the two domains computer science and biology.

Interdisciplinary domains have better chances for invention and innovation as they are positioned in the borders of the typical scientific domain. Many ideas and inventive patterns could be inherited from other domains bringing a successful solution (Langley et al., 1987). Staying in the core of the scientific discipline usually has much lower chances to find a breakthrough discovery comparing with research studies in the frontier of the discipline closer and sometimes overlapping with other disciplines and research communities. The invention pattern of solving some root-cause problems in a given discipline could be borrowed in another context of another discipline and successfully worked as an inventive solution on that domain. This idea of common inventive patterns was brought into a new emerged domain in systematic innovation called TRIZ (Altshuller, 2002) which means the theory of inventive problem solving[9].

An example of a useful interdisciplinary case can be seen with some mathematical models that have been borrowed from physics to data mining. Entropy and inertia mathematical

[9] `https://en.wikipedia.org/wiki/TRIZ`

models taken from energy domain are used in machine learning models like decision trees and clustering.

Another type of interdisciplinary cross-fertilization could be caused by having common concepts between two scientific domains. For instance, between Artificial Intelligence (AI) and Philosophy there are many shared concepts like "action," "consciousness," "epistemology," and "free will" (McCarthy, 2008).

Today, since the science is very wide and specific, scientists and researchers are very rarely exposed to more than one or two disciplines. In the Middle Ages, however, that was possible to find many examples of polymaths[10] who worked in many different disciplines. A couple of examples of polymaths are al-Khwarizmi[11] who worked in mathematics, astronomy, and geography, and Leonardo da Vinci[12] who worked in many fields including invention, painting, sculpting, architecture, science, music, mathematics, engineering, literature, anatomy, geology, astronomy, botany, writing, history, and cartography. Such polymaths were able to have many discoveries being exposed to many domains at the same time.

Nowadays, we only have very limited number of polymaths due to the specialization –and even hyperspecialization– of the scientists as well as a massive amount of knowledge in each scientific field. A good example however from the Modern Time is Herbert A. Simon (1916–2001), who was considered to be a very special trans-disciplinary researcher: crossing disciplinary lines in half a dozen fields (i.e., information processing, decision-making, problem-solving, organization theory, and complex systems), he formulated models in psychology to perform applications in artificial intelligence, but these models also had consequences at the economic level (Simon's theory of bounded rationality led to a Nobel Prize in economics in 1978) (Simon, 1996).

1.2.3 Dealing with Heterogeneous Sources of Metadata

A publisher can, for a certain extend, unify its data schema of publications per type. For example, the editor could define and name the required fields in which all authors must respect. Even if the schema evolves over years in order to cope with new demands,

[10] `https://en.wikipedia.org/wiki/Polymath`
[11] `https://en.wikipedia.org/wiki/Muhammad_ibn_Musa_al-Khwarizmi`
[12] `https://en.wikipedia.org/wiki/Leonardo_da_Vinci`

the schema of old publications managed by the publisher could be updated to keep consistency. The types of publication may vary however even for a single publisher. There are for example conference articles, journal articles, posters, reviews, books or book chapters.

The problem raises when it comes to a multidisciplinary digital library that usually has sources from different publishers. Here we can find a lot of differences in the schema in term of categorization, naming, hierarchy, data types and more. So, when a multi-source digital library aggregates a new publication, it goes through such challenge and usually ends up having a bigger combined schema in order to count for all new different schema. This causes a data duplication issue that adds more complexity in maintaining such combined schema for information retrieval systems.

1.2.4 Increasing Size of the Scientific Corpus

The volume and scale of data causes many technical issues. Everyday, number of publications, authors and affiliations are rapidly increasing. Big Data technologies are trying to solve this challenge in addition to variety of data forms, the analysis of streaming data as well as the uncertainty and quality of data. Information retrieval systems as a result should also tolerate to this Big Data issues. Introducing a semantic enabled information retrieval system usually comes in the expense of heavy computations that are not necessary able to scale.

1.3 Outlines and Contributions

This thesis consists in two parts. Part 1 will be the state-of-the-art in two main fields that are Information Retrieval in Digital Libraries (Chapter 2) and Semantic Text Similarity (STS) (Chapter 3).

The Information Retrieval chapter gives a background review on retrieving information from text corpus using multi-level of querying: set of keywords, sentence, document, set of documents. The Information Retrieval chapter also presents the approaches used in content-based research paper recommender systems as a specific type of information retrieval system that we are interested in.

The Semantic Text Similarity (STS) chapter provides a literature review on the topic from two main perspectives: the linguistics approach (lexical databases, classical NLP and semantic web) and statistical approach (machine learning and text mining).

Part 2 concerns the personal contribution of the thesis. It starts with Chapter 4 that talks about the challenge of evaluating semantic similarity between a pair of small texts (sentences, tweets, headlines or questions). Three main benchmarks will be discussed including an international semantic text similarity task in which we experimented our proposed model. We will also show two use cases of using our model, one for detecting different writing styles and another for linking paper content to sentences in the abstract of the paper. We will also talk about our proposed sentence semantic similarity estimator and what are the main categories of approaches dealing with this challenge (unsupervised, supervised and mixed models using part-of-speech, lexical databases and sentence embedding).

Chapter 5 will present our approach in semantic-based paper recommendation and scientific corpus expansion. The chapter also presents a detailed case study on a multidisciplinary domain.

In Chapter 6, we will talk about the semantic metadata enrichment using semantic-based scientific categorization in digital libraries and the importance of that in trans-disciplinary research. It will mainly present our approach in semantic-based metadata enrichment of digital libraries comparing to other multi-label classification methods.

Both Chapters 5 and 6, shed the light on the issue of dealing with multidisciplinary, interdisciplinary and trans-disciplinary research, and the semantic challenges that comes with it. We will also provide some work and proposed solution on how text mining and semantic-based techniques could help in solving it.

In Chapter 7, we propose a holistic approach for dealing with text similarity on many levels: Words, sentences, paragraphs and documents. It will also describe the advantage of using both semantic statistical learning and semantic networks together in solving some text mining. We will show how would such hybrid approach could enhance the prediction not only in accuracy but also in diversity and in unexpectedness.

Finally, In Chapter 8 we provide a general conclusion on our proposal in solving this digital library issues as well as our perspectives in the domain.

Our contributions could be summarized in the following items:

- A proposal of a general purpose sentence level semantic similarity using an extendable pairwise feature set (Al-Natsheh et al., 2017b). The proposed approach showed consistent evaluation results when applied in a couple of use-cases one in writing styles and another in abstract to paper content sentence semantic linkage.
- A new semantic approach for content-based recommender system for interdisciplinary scientific topic. The approach was also used for corpus expansion and further scientific topic modelling and keywords usage over time on a multidisciplinary text corpus (Al-Natsheh et al., 2017a).
- A novel scalable model for metadata enrichment and automatic semantic tagging of multidisciplinary digital library in comparision with classical topic modelling approaches and multi-labelling techniques (Al-Natsheh et al., 2018; Martinet et al., 2018).
- Diversity and unexpectedness analysis in semantic-based scientific paper recommender systems and semantic topic tagging and its role in promoting trans-disciplinary research.
- Open-source software solutions for sentence semantic similarity estimation[13] and automatic metadata semantic tagging[14] as well as reproducible use case experiments on a multidisciplinary digital library.

[13] `https://github.com/natsheh/sensim`
[14] `https://github.com/ERICUdL/`

Part I

State of the Art

Chapter 2

Information Retrieval in Digital Libraries

2.1 Chapter Overview

The information retrieval chapter gives a state of the art background on retrieving information from text corpus using multi-level of querying: set of keywords, sentence, document, set of documents. This chapter also presents the approaches used in content-base research paper recommender systems as a specific type of information retrieval system that we are interested in.

2.2 Introduction and Main Definitions

Digital Library (item ❶ in Figure 2.1)

Information retrieval is considered as one of the core components of any digital library. The definition of the concept "digital libraries" was raised mainly by the research community of computer science in the 1990's (Borgman, 2003) (same decade of the digital revolution and the invention of the web). One of a good definition of digital libraries could be found in (Kresh, 2007) as "A digital library is a library in which a significant proportion of the resources are available in machine-readable format (as opposed to print

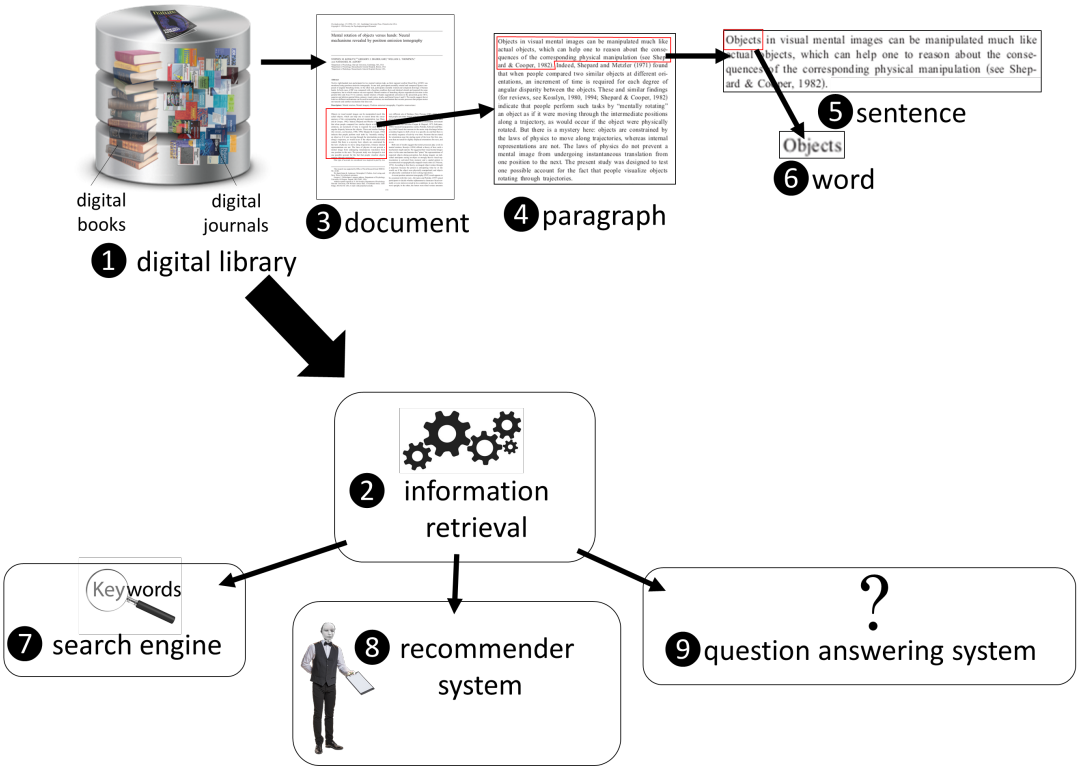


FIGURE 2.1: Information Retrieval in Digital Libraries.

or microform), accessible by means of computers. The digital content may be locally held or accessed remotely via computer networks."

Information Retrieval (item ❷ in Figure 2.1)

Information Retrieval (IR) is a technical field and active research data mining area that consists in extracting information in any form, including multimedia, text or knowledge. In this chapter, we will discuss three types of IR systems that are: search engines, recommender systems and question-answering system.

Information retrieval is the system that would allow filtering, searching or accessing a document or even a piece of information from the digital library corpus. The dominant type of digital library information retrieval systems is the search engine, which is mainly text-based. The primary role of most of the text-based search engines is to index documents based on their textual terms and the query matching to the document index. Beyond finding the matching results, the search engine also ranks the matched documents according to to weighting criteria, i.e., the Term Frequency-Inverse Document Frequency

(TF-IDF). Additionally, most of the search engines allow user-defined boosting factors to reflect the importance of query term (or word) matching with the document's section, i.e., keywords or title versus the body of the document.

As a widely used and fast retrieval engine; the traditional search engine yet encounters a primary informational retrieval issue which is counting for the meaning matching between the query and the content of the document. If there is no term match between the query and the document, the later would never appear as a matching result. Even though, the user might express his query in certain terms that are different from the ones used in the document. For example, the query "An infant sleeps on the chest of a man" and the document section "a baby having a nap on his father" would not match. Relying exclusively on term matching would solve most of the cases and would, for sure, retrieve many relevant documents. However,are these retrieved documents the only relevant ones among all available documents in the corpus? This is the primary question addressed in this thesis. In fact, the corpus contains additional relevant and related documents, but classical search engine systems, using term matching, would never find these documents that will thus be left undiscovered. We can find such case in interdisciplinary research topic for example.

This chapter presents and reviews the background work and the state-of-the-art of the following three fields in information retrieval:

- Search engines (Section 2.3)
- Recommender systems (Section 2.4)
- Question answering systems (Section 2.5)

2.3 Search Engines (item ❼ in Figure 2.1)

In this section, we will first develop some necessary basic concepts in document search engine, for instance the bag-of-words and the weighting schemes. Then, we will list different ways of querying for an information starting with a word, a sentence, a paragraph, a document or even a set of documents.

2.3.1 Basic Concepts

2.3.1.1 Bag-of-Words

Clustering textual data groups similar documents and reveals hidden connections. As textual data is more complex than numeric data, it requires to be treated differently. There are a few assumptions for processing textual data and different assumptions lead to different approaches. The "bag-of-words" assumption is one of the most popular vectorization models. It considers a piece of text (or a document) as a set of words. In this assumption, the ordering of words is ignored, only their existence matters. However, there are other assumptions that believe the ordering of words conveys necessary information, which is used as features in corresponding models.

In practice, if we have a corpus of thousands of documents that we want to represent in numbers, a bag-of-words model would have an ordered vocabulary size of hundred of thousands of words. Each document would then be represented by a one-hot-encoder style where the document words that exist in the vocabulary would have a value of 1 in the index (i.e., order sequence location) in the bag-of-words vocabulary whereas the rest of document vector elements would have 0 values. Here is a small example of denouements and their bag-of-words vector representation:

- the children are playing in the hall
- the conference is held in hall 1
- the conference is about children rights

For this small corpus of small documents (sentences in this simplified example), the set of unique words using the bag-of-words model are 12 words:

{"the", "children", "are", "playing", "in", "hall", "conference", "is", "held", "1", "about", "rights"}.

This set of unique words is called the vocabulary of the bag-of-words model. In this case the size of the vocabulary is 20. The vocabulary is then used as a list in which the order (i.e., index) of each word is preserved for the document vector representation. In our example, The bag-of-words vector representation of the documents using the vocabulary (the ordered list of unique words) would then be:

- 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0
- 1, 0, 0, 0, 1, 1, 1, 1, 1, 1, 0, 0
- 1, 1, 0, 0, 0, 0, 1, 1, 0, 0, 1, 1

We can see in the text representation model that the number of non-zero values in each document representation correspond to the number of unique tokens, words in that document. So, we can imagine that the number of zero values of a document vectors in a large corpus would be very high. Having a large number of zero values would result in having the documents similar to each others in such high dimensional space. This is what is called the sparsity problem. Accordingly, we can conclude that the bag-of-words model would work well if the size of the documents is big. A good example is a corpus of library books as documents. Small documents however, for example, tweets, would be very sparse and thus would not capture the semantics representation as we would have in the books example where a book has high number of unique words.

2.3.1.2 Bag-of-N-grams and N-gram Range

The bag-of-words is an easy and proven to work model for document retrieval for decades. However, the main draw back of this model is that it looses the word order and syntactic information. In the case of multi-word phrase, if we loose the order, the meaning of the whole sentence will be lost. For example, "cold fish" is a multi-word expression phrase composed of two adjacent elements that means a hard-hearted, unfeeling individual, one who shows no emotion. Obviously, if we split this phrase into two words, the meaning of the expression would be lost. To answer this issue, encountered in the bag-of-words model, a tweak is necessary. It consists in using "word n-grams" instead of words forming what is called a "bag-of-n-grams." A bi-gram of words, for instance, would group any frequently occurred sequence of two words and thus consider it as one phrase or concept to be used in the bag-of-words model, as if it was a single word. More examples other than "cold fish" are ("data mining", "New York", "machine learning", etc.). A tri-gram is similar but with three words. A couple of examples are "natural language processing" and "neuro-linguistic programming." A general concept to cover this number of words is what we mean by word n-gram where n is the number of sequenced words.

A more general version of word n-grams is the n-gram range. for example a range of (2,4) n-grams bag-of-words model will include all single words, all sequence of 2 words, 3 words and 4 words in the vocabulary. Here is an example of the bag-of-words vocabulary for this small document "Data mining is the process of discovering patterns in large data sets" (some n-grams are in underlined for further discussion):

["data mining", "mining is", "is the", "the process", "process of", "of discovery", "discovery patterns", "patterns in", "in large", "large data", "data sets", "data mining is", "mining is the", "is the process", "the process of", "process of discovering", "of discovering patterns", "discovering patterns in", "patterns in large", "in large data", "large data set", "Data mining is the", "mining is the process", "is the process of", "the process of discovering", "process of discovering patterns", "of discovering patterns in", "discovering patterns in large", "patterns in large data", "in large data sets"].

As we may see in this example, not all of the vocabulary items are considered as frequently used n-grams(the expected-to-be-frequent ones are underlined and listed hereinafter: "data mining", "data sets", "is the process of." If we would keep all the n-gram range vocabulary in a corpus of thousands of documents, we would end-up with a vocabulary size of billions words which, for the majority, would not be a frequently used multi-words concept. Filtering out non-frequent n-grams or only keeping very frequent ones would solve this issue. However, retaining all the vocabulary without such filtration would result in a very sparse document representation. As a consequence,all documents would be numerically very similar to each others in any document retrieval system.

2.3.1.3 Weighting Schemes

Another issue to be considered in the bag-of-words model is the importance to reflect how important a word is in the vocabulary to better capture the document semantics. For instance, when comparing two given documents, a word like "computer" should be assigned more importance (more weight) than a word like "between", because "computer" has more discriminating power than a common word like "between", for the reason that the later appear in most of the documents in a corpus. These common words are called stop-words, they are document indiscriminate words. Therefore, to obtain a good vocabulary, the tokens should be composed of more discriminating words and less non informative stop-words. The purpose of the vocabulary is to have "discriminating words"

features that will allow to differentiate semantically non-similar documents. The word "computer", for example, would only exists in computer related documents but would unlikely be mentioned in sport or political documents. This idea of assigning more weight to more discriminating words is what is called weighted bag-of-words model.

There are many weighting criteria; however, we will focus on the two most commonly used. The first criterion is the "Count-Vectorizer" which consists in using; in addition to one and zero values, an integer that represents the occurrence of words in the document. This way, a very frequently used word in a document would have higher weight than a word that only occurred once or twice. For example, the word "football" could be mentioned once or twice in a political book while it would be mentioned hundred of times in a sports book. Thus; using frequency as the value, instead of using 0-value in the one-hot-encoder method, would better capture the semantic information of the document and translate to a better vector representation.

The second and widely used weighting schema is the Term-Frequency Inverse-Document-Frequency (TF–IDF) as in Equation 2.1. It mainly increases the importance of rare words in addition to the frequency of the term. Another similar weighting schema is *BM25*, which, in this case, assigns a weight to a term within the vocabulary, reflecting its importance in the collection. As documents often contain different number of terms, the Document–Term matrix is usually normalized, standardized or scaled.

$$\text{TF-IDF}_{t,d} \;=\; tf_{t,d} \times idf_t \;=\; tf_{t,d} \times \log\left(\frac{N}{df_t}\right) \tag{2.1}$$

A recent work showed that increasing the weight of the distributional features that are decorative will achieve high accuracy of paraphrase detection (Ji and Eisenstein, 2013). Authors of that paper introduced a weighted schema called Tf-KLd based on supervised information. They developed their model on the bases of Kullback-Leibler divergence (KLd). The measure of a discriminative feature k is calculated as in equation 2.2

$$KL(pk, qk) = \sum_x p_k(x) \log \frac{p_k(x)}{q_k(x)} \tag{2.2}$$

where

$$p_k = P(w_{ik}^{(1)} | w_{ik}^{(2)} = 1, r_i = 1)$$

$$q_k = P(w_{ik}^{(1)} | w_{ik}^{(2)} = 1, r_i - 0)$$

For distance-based algorithms and when the TF-IDF scheme is applied, cosine similarity is often used to measure the proximity between two documents d_i and d_j:

$$cosine(d_i, d_j) = \frac{\langle d_i \cdot d_j \rangle}{\|d_i\| \cdot \|d_j\|}.$$

where $\langle \cdot \rangle$ indicates an inner product, and $\|d\|$ indicates the Euclidean norm of a document vector d. Under the vector space model, where documents are featured by their terms, the cosine similarity measures the angle of two document vectors in the projected space. Being 0 means that the document vectors are orthogonal, thus entirely dissimilar; and being 1 indicates that the document vectors are pointing to the same direction, they are entirely identical.

Depending on the choice of models, different proximity measures should be used. For example, when documents are represented as probability distributions over terms, Kullback-Leibler divergence (Joyce, 2011) is often used as a proximity measure. It measures how one probability distribution diverges from a second expected probability distribution.

2.3.2 Keywords or Words (item ❻ in Figure 2.1)

Using term matching is a relatively basic and straight forward approach that is widely used in text-based search engines. The main processes done on this level are stop-word filtering, TF-IDF weighting, lower-casing and stemming. Depending on the language or the field of use. As indicated previously in Section 2.3.1.3, a set of frequently occurring terms (stop-words) usually do not assist in distinguishing one document from another. do not Some examples of stop-words are: (“a”, “an”, “the”, “did”, “do”, “has”, “he”, “she”, *etc.*). Stemming, however is another important term processing approach in which the term set: (“teachers”, “teach”, “teaching”, “teacher”, *etc.*) would all be considered as equivalent terms in the term-matching.

There are, however, at least two other main issues to consider at the term level. The first issue is the fact that some concepts are expressed by more than one term, or by what we refer to as a phrase. For example, “New York” or “Natural Language Processing (NLP)”. Word n-gram is a technique that has been used to minimize the effect of such

issue. Using frequent sequence of 2 or 3 words in a corpus would distinguish, for a certain extent, a lot of these cases like in the NLP example.

The other issue is related to semantically similar terms or related terms having another meaning Yet, this problem can be tackled in two different ways: the synonym set (synset) and the word embedding. Lexicon databases are defined as a linguistically rich knowledge graph from which we can extract sets of synonyms, i.e., synsets. The synset contains variations of terms that are strongly related in the meaning; (e.g., "new", "modern", "recent", "updated"). There is still an issue in lexicon databases, for the reason that they are expensive to develop and maintain. As a consequence, it is never considered as completely covering all special domains, i.e., science. With new scientific discoveries, the number of scientific terms is increasing accordingly and the same word could have different meanings depending on the context. Another issue related to terms is the concept disambiguation. An example of that is the disambiguation between the synset of "Orange", the telecommunication company and an "orange", the fruit, or "AI", the scientific field and "AI", the movie.

Word embedding (Collobert et al., 2011) is another approach that provides a solution for matching semantically related terms. As we will present in more details later in Section 3.2.1, word embedding is a numerical representation approach in the semantic space using dense vectors of hundreds of dimensions. These semantic space dimensions encode the context which comes with the frequent term.

There are many ways to find the word semantic representation in such dense space. It could be statistically derived from a big corpus; starting by encoding the word by its frequent neighbours in a sparse representation and then applying a matrix decomposition to obtain the dense representation where, the matrix represents the fixed vector representation of the vocabulary in a bag-of-words fashion and where each row represents a term. Another way to obtain the word embedding is by learning its semantic features using a supervised neural network. It is obtained either by optimizing the prediction of the word vector, given its frequent neighbours, or by optimizing the prediction of the vector representation of its frequent neighbours, given the word.

Applying the word embedding approach to find synonyms in an information retrieval system like a search engine will have some limitation. For instance, it is not trivial to determine the semantic similarity threshold; Using a high threshold similarity value

means too few possible synonyms whereas using low threshold similarity value might result in including irrelevant terms out of the actual synonym set of the word.

Another issue is to disambiguate the sense and the phrases like "New York" and "Machine Learning." This issue has been somehow worked out using sensEmbed (Iacobacci, Pilehvar, and Navigli, 2015) and Sense2Vec (Trask, Michalak, and Liu, 2015). However, such approaches still did not provide a sufficient solution in practice because they needs huge text corpus than the one required to find a good word embedding representation of single terms.

2.3.3 Sentences (item ❺ in Figure 2.1)

Sentence essentially a segment of text that represents a meaningful thought or a fact. Also a question could be considered as a type of a sentence with relatively similar number of terms but in a different syntax. The sentence as a text representation of a thought is the closest concept of utterance in linguistics but not equivalent to it. Similarly to terms or words, the same thought could be expressed in many sentences that are not necessary similar in words or even characters. A sentence pair example on that case:

- "The queen patronages the graduation ceremony of Jordanian military forces"
- "Rania Al-Abdullah attended the awarding of the new solders"

Sometimes, in the other hand, two sentences could have many common words but are semantically not similar at all. A sentence pair example on that case:

- "Good chief executives like Steve Job makes Apple great for everyone"
- "Steve makes a good pine apple that everyone likes, so his job makes him a great chief"

The difficulty for a term matching-based system to recognize the semantic irrelevancy in the sentence pair of the later example could be reduced to a certain extent by using NLP techniques including Named Entity Recognition (NER). However, using such NLP techniques is usually not practical in search engine systems that generally index large

number of documents. The query analysis using NER and other NLP tools would also delay the fast response time of the search engine system.

The number of possible sentences is far to be bounded; as in the case of words and its synonym set (synset). Lexical databases provide a much closer realization of such boundaries where the number of possible words in a language is somehow bounded and thus its synonym sets. That is to say that there could be a finite number of synset even very high. However, the number of unique sentences is considered infinite even for a language with 28 characters and hundreds of thousands of unique words. The possible number of word combinations forming different meaningful sentences could not be bounded. We will further discuss this sentence semantic representation issue in Chapters 2 and 4. We will show that transferring the problem into modelling sentence pair semantic similarity instead of trying to model the representation of each sentence alone is a practically working solution for sentence representation.

Back to the example of Queen Rania of Jordan; breaking down the sentence into words and then use the word embedding as a base to find the sentence embedding could also be a solution direction of the semantic representation of the sentence as we would also see in Chapters 3 and 4. However, it is not a straightforward usage of word embedding as the number of words on the compared sentences could be significantly different. Using the word embedding for sentence pair example of Steve Jobs would be problematic. For instance, if we get the semantic vector representation of each word of the sentence and compute the average of the word vectors of each sentence, the result will be very similar for both sentences. However, the two sentences are semantically not similar on a sentence level.

We will list and discuss several approaches in Chapter 3 like sentence embedding (i.e., Sent2Vec and skip-thought vectors) and paragraph vectors, many of which showed good experimental results and some performed very well in sentence semantic similarity tasks for question answering systems and sentiments analysis. Most of these techniques try to apply unsupervised feature learning approach as in the word embedding but sometimes guided by a training set of sentence pairs annotated as similar or not resulting in a better performing semi-supervised models.

2.3.4 Documents and Paragraphs (items ❸ and ❹ in Figure 2.1)

A document conceptually is a container of an article, or even a web page, as a collection of sections or paragraphs. Technically speaking, it could range from a single sentence or a paragraph to a book with many chapters. Representing a document for information retrieval purposes was the first granularity of text that was well studied and used in practical search engine. A bag-of-words (BoW) matrix representation where rows are documents and columns are words is the typical way to index and represent the documents and the queries.

The queries are not limited to a set of keywords; they could also be a document asking the search engine system to find similar documents. Such document query type is usually referred to as (Query by Example) or in some search engine systems as (More Like This query) (Dixit, 2017). This could be extended to a small set of documents instead of one only. These types of queries could be utilized as a context-based Recommender System However, when we need to have a bigger set of documents as a query, we may rather refer to such system as a corpus expansion.

Similar to the semantic representation challenge we introduced for words and sentences, documents that are not well represented in a semantic space would limit the accessibility of relevant and related documents. This challenge is valid not only for search engine systems but also for recommender systems and corpus expansion systems. Thus, building semantic-enabled information retrieval systems would require dense semantic representation methods of documents as BoW sparse representation would not be sufficient.

If documents contain too many terms, feature reduction or dimensionality reduction is necessary. It helps to reduce the number of features and to transform the features into a different space, where document similarity or other text mining tasks can be performed with better effectiveness. There are a few document semantic representation and dimensionality reduction techniques that have been introduced like learned semantic feature approaches (i.e., Paragraph vectors, document co-clustering), topic modelling approaches (i.e., LDA) and a family of LSA decomposition method of BoW representation which we will introduce in Chapter 3 section 3.2.4.

2.4 Recommender Systems (item ❽ in Figure 2.1)

The recommender system (Ricci, Rokach, and Shapira, 2015) is a computer science multidisciplinary field that overlaps with Human-Computer Interaction (HCI), data mining, machine learning and information retrieval. The recommender systems have gotten more attention since Netflix, an online DVD-rental and video streaming service which had its US$ 1 million cash prize[1] that was awarded back to 2009. The competition was for designing the best user rating prediction model to be used in its film recommendation system. Besides finding the best predictor, the methods used by winning system (Koren, 2009) like singular value decomposition (SVD) for data dimensionality reduction and gradient boosting decision trees (GPDT) for regression model was a key winning factor. Since then, the field had an obvious spike in related publications. A lot of research efforts have been done in the recommender systems, especially in the field of collaborative filtering (Schafer et al., 2007) in addition to other research problems like cold-start (Schein et al., 2002), active learning (Rubens, Kaplan, and Sugiyama, 2011), preference elicitation and similarity search (Zezula et al., 2006) (Lee, Lakshmanan, and Yu, 2012).

2.4.1 Collaborative Filtering

Collaborative filtering is mainly based on a matrix that is similar to bag-of-words described in section 5.1. However, instead of having documents and words, the collaborative filtering matrix has users and items for headers of columns and rows. The values in such matrix could be the user rating, number of time the user interact with the item (i.e., buy, view, click..). As in the bag-of-word, the purpose is to have a vector representation of the users, or items, in which we could capture similarities between these users. A recommender system would then assumes that a user would probably be interested to interact with new items that similar users have interacted with. We can see in Table 2.1 a simplified example of an item recommendation (i.e., headset) to a user based on the purchase history of all users. The example shows a user, Sarah, who bought a laptop, keyboard, mouse, webcam but not yet a headset while most of other users who bought similar items did also buy a headset. Similar concept could be used to recommend, for example, news articles or scientific papers.

[1] `https://www.netflixprize.com/`

TABLE 2.1: A simplified example of collaborative filtering user-item recommender system.

user/item	laptop	mouse	keyboard	webcam	**headset**	bag
Tom	1	1	1	1	1	0
Ali	1	1	1	1	1	1
Sarah	1	1	1	1	**0**	0
Lya	1	1	1	1	1	1
Adam	1	1	1	1	1	0

Of course, the collaborative filtering matrix is much larger in practical examples. It usually has high sparsity, again, similar to the bag-of-words. The user could have additional columns other than items, like demographics information (gender, age, location...) which provide more similarity features. Similarly, in the news article recommendation, textual similarity features could also be added to the columns.

There are other forms of such matrix for the purpose of recommendation that are item-item and user-user. These matrix formats are referred to as co-occurrence matrices in which the values in the matrix indicate the co-occurrence of the user pair, or item pair, in a certain context (i.e., bought same item, read same article, attended same class...). Same co-occurrence approach was also used for word-word where the co-occurrence context could be the same sentence, paragraph or document. As in user-item collaborative filtering we can infer the similarity feature or representation of the user in user-user and similarly for item-item, word-word or abstract-abstract co-occurrence matrix.

One known open research issue in collaborative filtering is the cold-start problem where the recommender system needs to have a sufficient amount of data to create a useful user-item like matrix. The cold-start problem is mainly related to recommending items to new users or recommending new items. In case of new users, the system does not have enough historical data to predict their preferences in order to send recommendations to them. However there were many works that try to address this problem (Schein et al., 2002) (Lam et al., 2008) (Lika, Kolomvatsos, and Hadjiefthymiades, 2014).

2.4.2 Context-based Recommender Systems

IR systems are not limited to typical search engine systems. Recommender systems are another type of information retrieval systems that could also be used in digital libraries.

For example, if a researcher wants to expand his literature review citations for a given topic, he would need to provide a set of publication examples that he wants to expand. The aimed expansion is to explore other semantically relevant and related publications from other scientific disciplines that might use different keywords to describe similar topics. This obviously requires a good document semantic representation as introduced earlier in this chapter. In addition to relevancy, such recommender systems should also be able to scale, serving millions for documents and providing quick query responses. We will propose a semantic based approach for content-based paper recommendation in Chapter 5.

Context-based recommender systems are based on recommending items that are similar in the textual content as well. It is similar to *More-Like-This* query in search engines where the query is a document, or a small set of documents, that the user would like to explore similar ones. Unlike collaborative filtering, such content-based recommender system does not suffer from the cold-start issue in recommender system as there is no need for item-user like matrix to extract the document similarity features.

2.5 Question Answering Systems (item ❾ in Figure 2.1)

Question answering systems are also another important type of information retrieval systems. Instead of looking for documents, the user might be interested in getting direct answers that are usually in the form of a single or few sentences. This facilitates the knowledge access instead of the need of reading many publications looking for an answer. It requires, however, a good semantic representation of sentences as also described earlier. However, millions of documents would include billions of sentences. Representing all of these sentences in their semantic space and indexing them in a semantic-enabled information retrieval system is another big challenge. Another way to represent data in a semantic-enabled question answering system is the knowledge graphs or semantic web ontology. These however, are considered expensive to build and maintain. Moreover, It is relatively difficult to be queried using a special query language (i.e., SPARQL) which is more complex than hitting a question as a search keywords query.

A couple of good examples of quetion answering systems are; DBpedia Bot[2] that retrieves answers from DBpedia (a knowledge graph extracted from Wikipedia articles and infoboxes), and QAKiS[3] (a multi-lingual question answering system that supports French, Italian, English and German (Cojan, Cabrio, and Gandon, 2013)).

2.6 Conclusion

In this chapter, we introduced a brief literature review on the domain of information retrieval including search engines, recommender systems and question answering system. We discussed the various query size level starting from keywords, sentences until documents, which are used nowadays in knowledge access as we introduced in Chapter 1. The purpose of this chapter is also to define some basic concepts to which we refer in our proposed approaches in semantic based information retrieval systems in digital libraries that will be discussed in Chapter 5. In the next chapter, we go a bit deeper in the direction of semantic short text representation, listing the state-of-the-art work on that domain.

[2] `http://chat.dbpedia.org/`
[3] `http://qakis.org`

Chapter 3

Semantic Text Similarity

3.1 Chapter Overview

In the chapter, we will present the concepts and the existing state of the art approaches and techniques that concern the semantic text similarity. The chapter starts with fundamental knowledge on this subject. This includes the basic assumption, the base model, and commonly-used proximity measures for semantic text similarity. After presenting these concepts, we propose an overview of popular and recent text similarity approaches. The chapter also contains comparisons and discussions that are indispensable for readers to understand the content of semantic text similarity. Since this subject also concerns the text numerical representation, we also introduce some fundamental knowledge on dense semantic space text representation. The chapter consists in two main sections. The first one will list several text mining and machine learning based approaches in semantic text representation and similarity The second one, however, provides a set of approaches that could be categorized as computational linguistics approaches.

3.2 Text Semantics Using Mathematical Approaches

3.2.1 Word Embedding

"You shall know a word by the company it keeps" (Firth, 1957). This famous quotation by an English linguist has become the most cited one wherever introducing the word

embedding concept. It summarizes the basic phenomena behind encoding a word with a dense vector representation in the semantic space. The context-dependent nature of meaning "context of situation" could be explain by this example:

- I will travel to *Barcelona* for this summer vacation.
- I will travel to *Rome* for this summer vacation.

We can see that the two cites Barcelona and Rome could be seen as interchangeable words. These two interchangeable words occur in similar context. This could be justified by the linguistic assumption of that words that often occur together are often related. These two words in the example have common semantics information, they are both cities, both are considered as a touristic destination, etc. The concept of interchangeability could be compared to the synonym set where we could interchange the synonyms in a sentence without changing the general meaning of the sentence. Accordingly, we can consider the interchangeable words as semantically similar words.

Mathematical speaking, we can measure the co-occurred neighbouring words by counting the times they came together in the same context. A context, for instance, could be a sentence or a paragraph or even a window of size of 7 for example (3 words before and 3 words after). Scanning a big text corpus by counting the neighbouring words of each unique word in the corpus would result in a histogram of neighbouring words for each word. When applying the idea of interchangeable words occurring in similar context (similar to synonym set concept), the semantically similar words should have similar histograms. That's to say, two similar word histograms of two similar words should be similar. applying that to the cities example:

$$\text{Hist}(\text{“Barcelona”}) : \text{Hist}(\text{“Rome”})$$

The histograms have a very nice mathematical property; since they can represent log probabilities of word co-occurrences. Accordingly, the addition of histograms is equivalent to the product of distributions. Since the product is an intersection, the similarity between two histograms could be calculated by the dot product of the log of the two histogram. Let us take another explanatory example showing the nice mathematical property:

- assuming that Hist("Germany") is the histogram representation of the word 'Germany' (i.e., the log of the histogram distribution of words that occurs near it) where near means for example a distance of 5 words.
- In Hist("Germany"), we have high probability to words that co-occur with "Germany"
- Hist("Germany") + Hist("capital") , adding the log, means multiplying the distributions, which means taking the intersection
- So, the sum would be the distribution of words that occur with both "Germany" and "capital" (related to both words)
- The resulted distribution will be very close to Hist("Berlin")!

However, using the histogram to represent the word as a vector would mean that each vector would have the size of the vocabulary (hundreds of thousands). So, histograms are very large objects that also suffer from sparsity. Another drawback of using histograms is the poor representation of rare words. This raises a very important questions that would be the entry to the word embedding which is: Can we compress the histogram into a lower dimensional 'dense' vector while keeping this nice property? The answer to this question was the key for several word embedding models, the most famous model is called Word2Vec model (Mikolov et al., 2013b) where, instead of capturing the co-occurrence histogram directly, the model predicts neighbouring words of each word using one of these possible methods:

- Continuous Bag-of-Words Model (CBOW): predicting current word based on the context (i.e., surrounding words).
- Continuous Skip-gram Model (Skip-gram): maximizing classification of word based on another word within a certain range surrounding it.

It appeared that such predication model is not only faster but also easier to incorporate new documents or words. The skip-gram method could be briefly described as follows: Take a vector, multiply it to a matrix A, pass it to a *softmax* function and get the histogram in other words, modelling the distribution with a maximum likelihood estimation. The *softmax* function, also called normalized exponential is used in various probabilistic

multi-class classification methods including multinomial (or dynamic) logistic regression as in Equation 3.1.

$$\text{Hist}(\text{“Germany”}) = \text{softmax}(A \times \text{Vec}(\text{“Germany”})) \tag{3.1}$$

Now, giving a lot of histograms, we can estimate all the above parameters (matrix A). Interestingly, similar words are found to have similar skip-gram vector representations. Another nice property of the skip-gram model is that it can estimate rare words more accurately than using histograms.

However, do such prediction based models preserve the linear relationship we found in the histogram? The answer is yes! Because adding the input of the *softmax* (normalized exponential) is equivalent to multiplying the distributions:

$$\text{If:} \begin{cases} \text{Hist}(\text{“Germany”}) = \text{softmax}(A \times \text{Vec}(\text{“Germany”})), \text{and} \\ \text{Hist}(\text{“capital”}) = \text{softmax}(A \times \text{Vec}(\text{“capital”})), \text{and} \\ \text{Hist}(\text{“Germany”}) \times \text{Hist}(\text{“capital”}) = \text{Hist}(\text{“Berlin”}) \end{cases},$$
$$\text{then: } \text{Vec}(\text{“Germany”}) + \text{Vec}(\text{“capital”}) = \text{Vec}(\text{“Berlin”})$$

Finally, the cost of computing the softmax is proportional to the number of words in the vocabulary. The two possible implementation inventive steps proposed by (Mikolov et al., 2013b) are hierarchical softmax and negative sampling. In practice, another way to perform the skip-gram model could be applying Stochastic Gradient Decent (SGD) using neural network (auto-encoder model) (Mikolov et al., 2013a).

Another word vector representation method that is frequently used due to its good performance in many tasks and the availability of its pre-trained vectors is GloVe (Pennington, Socher, and Manning, 2014). The main difference between GloVe and Word2Vec (CBOW or skip-gram word embedding) is that the word features are not learned but rather obtained using mathematical matrix operations on the word co-occurrence matrix of the large text corpus. Another extra work that was useful in this domain is sense embedding, or Sense2Vec (Trask, Michalak, and Liu, 2015), where the term is syntactically and semantically annotated before adding it to the vocabulary of words. For example, the

term “duck” could be a verb or a noun where, in each case, POS case has semantically different meaning. Another example but with named entities is Apple as an organization or apple as a fruit. Similar work on that direction is Sensembed (Iacobacci, Pilehvar, and Navigli, 2015) where the concept name got semantically disambiguated before finding its word vector representation.

3.2.2 Sentence Embedding

Words combine in order to produce units of discourse: an utterance. Words do not ‘carry’ or encode meaning. Rather, meaning is a property associated with a complete utterance (Evans, 2006) .Utterances do not exist in written language, only their representations do. For written language, the closest concept to utterance is sentence, knowing that they are not the same thing. Many successful models have been developed for sentence semantic embedding or sentence dense vector representation after the wide adaptation and the success of the word dense vector representation. However, most of such techniques use deep learning techniques learned on very large text corpus; and, in many cases, reuse the word vectors as input to such deep learning models. Example of these deep learning techniques are convolutional neural network (Collobert et al., 2011), recurrent neural networks (Mikolov et al., 2010) using many architectures like long-short term memory (LSTM) (Gers, Schmidhuber, and Cummins, 2000), bidirectional LSTM (Graves and Schmidhuber, 2005) and GRU units (Kiros et al., 2015). All of such neural network s are mainly used to learn the dense vector representation or the semantic features of the sentence in an unsupervised learning approach.

Sent2Vec is one of the recent practical open-source models that has performed very well in semantic similarity tasks (Pagliardini, Gupta, and Jaggi, 2018). Before that the same name, Sent2Vec, was used by Microsoft Research for one of their sentence embedding models that performs the mapping using the Deep Structured Semantic Model (DSSM) proposed in (Huang et al., 2013), or the DSSM with convolutional-pooling structure (CDSSM) (Shen et al., 2014) (Gao et al., 2014).

3.2.3 Sentence Semantic Similarity

3.2.3.1 Overview

In this section we will present the state-of-the-art review on evaluating sentence semantic similarity. First, we will list a few available benchmark in the domain then provide the proposed models and their evaluation results on the corresponding benchmark.

3.2.3.2 Benchmarks

There are a few data-sets available to test the sentences similarity measures. "Microsoft Research Paraphrase Corpus" (MSRPC)[1] was used in (Shin et al., 2015). MSRPC contains 5800 pairs of sentences, which have been extracted from news sources on the web, along with human annotations indicating whether each pair captures a paraphrase/semantic equivalence relationship (Dolan, Quirk, and Brockett, 2004). So the main use of this data set is the paraphrase detection in which sentence similarity measure could be utilized and evaluated against other models.

A frequently used benchmark in the field is introduced in (Li et al., 2006). It is composed of 65 sentence pairs that were scored by people expressing how similar are the sentences of each pair are. As a benchmark, 30 particular pairs are used for training and the rest for testing. More recent benchmark was published by the same co-authors (Li et al., 2006) a long with others which is the "131 sentence pairs" (O'Shea, Bandar, and Crockett, 2013). Some of the state-of-the-art results of this banchmark were reported in (Croft et al., 2013; Islam and Inkpen, 2008; Tsatsaronis, Varlamis, and Vazirgiannis, 2010)

Semantic Textual Similarity (STS) benchmark comprises a selection of the English datasets used in the STS tasks organized in the context of the International Workshop on Semantic Evaluation (SemEval) between 2012 and 2017. The selection of datasets include text from image captions, news headlines and user forums.

There are other related benchmarks and datasets, but less adopted by the research community, "WordSim-353" (Agirre et al., 2009), and "SimLex-999" (Hill, Reichart, and Korhonen, 2015; Chiarello et al., 1990)). Other used datasets for paraphrasing include

[1] `https://www.microsoft.com/en-us/download/details.aspx?id=52398`

Penn Paraphrase Database (PPDB) (Ganitkevitch, Durme, and Callison-Burch, 2013) and the PPBD subset, human scored paraphrasing database (Pavlick et al., 2015).

3.2.3.3 State-of-the-art Results over MSRPC Dataset

The state of the art results of MSRPC paraphrase identification benchmark are listed in Table 3.1 (as listed in ACL[2]).

Algorithm	**Reference**	**Description**	**Supervision**	**Accuracy**	**F**
Vector Based Similarity (Baseline)	Mihalcea et al. (2006)	cosine similarity with tf-idf weighting	unsupervised	65.4%	75.3%
ESA	Hassan (2011)	explicit semantic space	unsupervised	67.0%	79.3%
KM	Kozareva and Montoyo (2006)	combination of lexical and semantic features	supervised	76.6%	79.6%
LSA	Hassan (2011)	latent semantic space	unsupervised	68.8%	79.9%
RMLMG	Rus et al. (2008)	graph subsumption	unsupervised	70.6%	80.5%
MCS	Mihalcea et al. (2006)	combination of several word similarity measures	unsupervised	70.3%	81.3%
STS	Islam and Inkpen (2007)	combination of semantic and string similarity	unsupervised	72.6%	81.3%
SSA	Hassan (2011)	salient semantic space	unsupervised	72.5%	81.4%

[2] `https://aclweb.org/aclwiki/Paraphrase_Identification_(State_of_the_art)`

QKC	Qiu et al. (2006)	sentence dissimilarity classification	supervised	72.0%	81.6%
ParaDetect	Zia and Wasif (2012)	PI using semantic heuristic features	supervised	74.7%	81.8%
Vector-based similarity	Milajevs et al. (2014)	Additive composition of vectors and cosine distance	unsupervised	73.0%	82.0%
SDS	Blacoe and Lapata (2012)	simple distributional semantic space	supervised	73.0%	82.3%
matrixJcn	Fernando and Stevenson (2008)	JCN WordNet similarity with matrix	unsupervised	74.1%	82.4%
FHS	Finch et al. (2005)	combination of MT evaluation measures as features	supervised	75.0%	82.7%
PE	Das and Smith (2009)	product of experts	supervised	76.1%	82.7%
WDDP	Wan et al. (2006)	dependency-based features	supervised	75.6%	83.0%
SHPNM	Socher et al. (2011)	recursive autoencoder with dynamic pooling	supervised	76.8%	83.6%

MTMETRICS	Madnani et al. (2012)	combination of eight machine translation metrics	supervised	77.4%	84.1%
L.D.C Model	Wang et al. (2016)	Sentence Similarity Learning by Lexical Decomposition and Composition	supervised	78.4%	84.7%
Multi-Perspective CNN	He et al. (2015)	Multi-perspective Convolutional NNs and structured similarity layer	supervised	78.6%	84.7%
REL-TK	Filice et al. (2015)	Combination of Convolution Kernels and similarity scores	supervised	79.1%	85.2%
SAMS-RecNN	Cheng and Kartsaklis (2015)	Recursive NNs using syntax-aware multi-sense word embeddings	supervised	78.6%	85.3%

TF-KLD	Ji and Eisenstein (2013)	Matrix factorization with supervised reweighting	supervised	80.4%	85.9%

TABLE 3.1: The state of the art results of MSRPC paraphrase identification benchmark. The results are listed in order of increasing F score.

Looking to the results and the techniques reported in Table 3.1, we may conclude that recent supervised learning approaches that utilize both semantic statistical learning and computational linguistics methodologies like dependencies parser are achieving best evaluation results for this benchmark.

3.2.3.4 State-of-the-art results over SemEval STS Benchmark

The state of the art results of SemEval STS benchmark are listed in Table 3.2 (As in the benchmark official webpage[3]) where: *Sentence representation* model used in the system:

- *Independent*: systems that are solely based on a pair of sentence representations that are computed independently of one another
- *Other*: systems that also use interactions between sentences (e.g. alignments, attention or other features like word overlap)

Amount of *supervision* used by the system:

- *Unsupervised*: systems that do not use any STS train or development data (can include transfer learning, or resources like WordNet or PPDB)
- *Dev*: systems that only use the STS benchmark development data (weakly supervised)
- *Train*: systems that only use the STS benchmark training and development data (fully supervised)

[3] `http://ixa2.si.ehu.es/stswiki/index.php/STSbenchmark`

Sentence representation	**Supervision**	**Paper**	**Comments**	**Dev**	**Test**
Independent	Unsupervised	(Pennington, Socher, and Manning, 2014)	Glove	52.4	40.6
Independent	Unsupervised	(Joulin et al., 2016)	Fastext	65.3	53.6
Independent	Unsupervised	(Salle, Idiart, and Villavicencio, 2016)	LexVec	68.9	55.8
Independent	Unsupervised	(Mikolov et al., 2013b)	Word2vec skipgram	70	56.5
Independent	Unsupervised	(Duma and Menzel, 2017) (More details in section 3.2.4)	Doc2Vec (SEF@UHH3)	61.6	59.2
Independent	Unsupervised	(Kruszewski, Lazaridou, and Baroni, 2015)	C-PHRASE	74.3	63.9
Independent	Unsupervised	(Le and Mikolov, 2014; Lau and Baldwin, 2016) (More details in section 3.2.4)	PV-DBOW Paragraph vectors, Doc2Vec DBOW	72.2	64.9
Independent	Unsupervised	(Wieting et al., 2016)	Charagram (uses PPDB)	76.8	71.6

Independent	Unsupervised	(Wieting et al., 2015)	Paragram-Phrase (uses PPDB)	73.9	73.2
Independent	Unsupervised	(Conneau et al., 2017)	InferSent (bi-LSTM trained on SNLI)	80.1	75.8
Independent	Unsupervised	(Pagliardini, Gupta, and Jaggi, 2018)	Sent2vec	78.7	75.5
Independent	Unsupervised	(Yang et al., 2018)	Conversation response prediction + SNLI	81.4	78.2
Independent	Dev	(Wieting and Gimpel, 2017)	GRAN (uses SimpWiki)	81.8	76.4
Independent	Train	(Tai, Socher, and Manning, 2015)	LSTM	75	70.5
Independent	Train	(Tai, Socher, and Manning, 2015)	BiLSTM	76	71.1
Independent	Train	(Tai, Socher, and Manning, 2015)	Dependency Tree-LSTM	76	71.2
Independent	Train	(Tai, Socher, and Manning, 2015)	Constituency Tree-LSTM	77	71.9

Independent	Train	(Yang et al., 2018)	Conversation response prediction + SNLI	83.5	80.8
Other	Train	(Shao, 2017)	CNN (HCTI)	83.4	78.4
Other	Train	(Al-Natsheh et al., 2017b)	mixed ensemble (UDL)	79	72.4
Other	Train	(Maharjan et al., 2017)	mixed ensemble (DT_TEAM)	83	79.2
Other	Train	(Wu et al., 2017)	WordNet + Alignment + Embeddings (BIT)	82.9	80.9
Other	Train	(Tian et al., 2017)	mixed ensemble (ECNU)	84.7	81

TABLE 3.2: SemEval STS benchmark results

We may conclude from the results of the approaches listed in Table 3.2 that relying only on unsupervised learning of the sentence representation does not lead to relatively high evaluation results. Other approaches, however, that have a mix of features (i.e., semantic networks or other external resources like WordNet) with unsupervised sentence representation features using supervised learning on the training set resulted in better results.

3.2.3.5 Unsupervised representation learning and Text Sequence Based Model

There are other methods to extract feature which could be classified as generic ones. A recent example is skip-thought vectors(Kiros et al., 2015) in which the model can be used to any text corpus without tuning and still perform very well. The skip-thought model is inspired by skip-gram model (Kiros et al., 2015) that uses a word to predict the surrounding context. Being an encoder-decoder framework model, skip-thought encodes

a sentence to predict the sentences around it. They use Recurrent Neural Network (RNN) (Cho et al., 2014) encoder with GRU (Chung et al., 2014) activation.

Another recent sequence based technique is AdaSent (Zhao, Lu, and Poupart, 2015), which is a self-adaptive hierarchical sentence model. AdaSent shows high accuracy values in many benchmarks. For such group of techniques, the text sequence is very important. It is the base of such vectorization methods. These methods can be considered as off-the-shelf generic sentence representation models; however, there are a lot of sub embeded mathematical models which brings complexity in implementation. They also need a large and diverse enough data-set to be generic, which requires weeks in the training process.

Most of these text sequence based recent models are using RNN as algorithm to model the sequence. As we will see in section 3.3.3, extracting sentence main components could, somehow, count for the sequence knowledge.

We can find some of these category of unsupervised and Neural representation models in Table 3.2 under the supervision type (*Unsupervised*).

3.2.3.6 Feature Engineered and Mixed Systems

There are a few models that performed well in the STS bechmark in which they combine more than one system and apply feature engineering and supervised learning. Among them are the one listed in Table 3.2 with a sentence representation type (Other).

3.2.4 Document and Paragraph Semantic Representation

In this section we will provide a basic literature review on a set of semantic feature extraction approach for the level of paragraphs or short documents.

3.2.4.1 Semantic Latent Analysis and Matrix Factorization

Applying dimensionality reduction techniques like Latent Semantic Analysis (LSA) (Landauer, Foltz, and Laham, 1998) and probabilistic Latent Semantic Analysis (pLSA) (Hofmann, 1999) to the bag-of-words provides better semantic feature vector for documents in the semantic space. Matrix factorization techniques like Singular Value Decomposition

(SVD) (Golub and Reinsch, 1971) and non-Negative Matrix Factorization (NMF) (Lee and Seung, 2001) are two very good factorization techniques that some software packages like Scikit-Learn (Pedregosa et al., 2011) managed to use in a way that can be scale to big text corpus, as in the case of, for example, TruncatedSVD (Halko, Martinsson, and Tropp, 2011) that can be applied to very big sparse TF-IDF weighted bag-of-words matrices.

3.2.4.2 Topic Modelling as a Document Level Semantic Information Extraction

Topic modelling is a way to group semantically similar documents under a topic. The document could however belong to more than one topic but with different degree of membership. So, topic modeling could be seems as a text fuzzy clustering method. The most famous method for topic model is Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan, 2003). LDA is a generative probabilistic model for a text corpus that is somehow similar to pLSA. It is based on a simple exchangeability assumption for the topics and terms in a document where the topics are distributions over words and this discrete distribution generates observations (words in documents) (Blei, 2012).

Tagging a document with a ranked list of semantic topics could be observed as a semantic information extraction. That is to say, the grouped documents per topic are semantically similar as they share common semantically related terms over the text corpus of what can be generally called discrete data collection where the probabilistic topic model was built on. Therefore, we think that LDA is a comparative technique for document semantic representation. LDA will be presented as a related work of Chapter 6 which takes about semantic tagging and meta data enrichment.

3.2.4.3 Feature Learning Approaches

Paragraph vector or Doc2Vec (Le and Mikolov, 2014) applies very similar methodology of Word2Vec (Skip-gram and CBOW models) using the frequent neighbouring words to predict the document features and vice versa. Another top cited model is Skip-thought vector (Kiros et al., 2015). It is also inspired by skip-gram model that uses a word to predict the surrounding context. Skip-thought uses Recurrent Neural Network

(RNN) encoder with GRU activation encoder-decoder to produce a generic model that can be used to any text corpus without tuning and still perform well. Both of Doc2Vec and Skip-thought vector could also be used for sentence embedding if we consider a document having only one sentence. They have been used for such sentence level tasks as unsupervised approach for sentence semantic evaluation tasks. For instance, they perform well on benchmarks like SemEval STS Benchmark as shown earlier in Section 3.2.3.4.

3.3 Computational Linguistics Approaches

3.3.1 Semantic Networks and Lexical databases

Lexical databases such as WordNet is widely used to get synonyms of words. WordNet has a concept hierarchy that could be used to compute semantic similarity of words. DBpedia is another knowledge base that is extracted from Wikipedia and contains many languages. It is multilingual and can be accessible using standers of Linked Data or Semantic Web.

Another conceptual knowledge-base which is tailored for Chinese language is HowNet. HowNet is an on-line common-sense knowledge-base unveiling inter-conceptual relationships and inter-attribute relationships of concepts as connoting in lexicons of the Chinese and their English equivalents (Dong, Dong, and Hao, 2010).

BabelNet (Navigli and Ponzetto, 2012) is a multilingual encyclopedic dictionary and semantic network covering 284 languages and containing 16 million entries. It combines many sources of knowledge including Wikipedia, WordNet, Open multilingual WordNet, OmegaWiki, Wiktionary and Wikidata. Babelnet allows extracting synonym sets as well as semantic domains, categories and neighbour concepts. Yago (Suchanek, Kasneci, and Weikum, 2007) is another comprehensive knowledge graph that also combines many knowledge sources. It has a recognized and widely used anthology for many other knowledge graphs and linked data projects.

3.3.2 Exploiting Synonymy

A recent work (Shin et al., 2015) exploited synonymy to measure semantic similarity of sentences. Authors of this work have utilized WordNet to estimate the similarity between words by the minimum number of synonyms chain between two nodes in a constructed synonymy graph. This is based on a social network concept that a friend of a friend is likely to be a friend. Extending this to the sentence scale, the similarity of two sentences is calculated by summing all the similarities between meaningful words. These meaningful words are selected by having POS tags. So, only nouns, adjectives and verbs were selected. The algorithm then generates word pairs from the same POS tag of the two sentences. As stated in equations 3.2 and 3.3, the maximum pair-wise similarity per word will be summed and then normalized by dividing the summation by the maximum number of words participating in the pairwise calculation.

$$Sim_{syn}(S_1, S_2) = \frac{\sum_{w_i \in S_1, w_j \in S_2} maxSim_w(w_i, w_j)}{max(|S_1|, |S_2|)} \tag{3.2}$$

$$Sim_w = \begin{cases} 1 & \text{if } w_i, w_j \text{ exactly matched} \\ (1 + dist(w_i, w_j))^{-1} & \text{otherwise} \end{cases} \tag{3.3}$$

Results of this similarity measure which makes use of WordNet show that the F-measure on MSRPC data set was 0.807 comparing with 0.753 utilizing the commonly used TF–IDF vector model similarity measure.

3.3.3 Syntactic Structure

Part-of-Speech (POS) is a very important methods when we talk about feature extraction. Tagging sentence main components eases the sentence matching task. So, the sequence of text considering POS feature would be indirectly used as most of POS algorithms actually base their prediction on the text sequence. We will see in sections 3.3.2 and 3.3.3 how POS is used in the task of semantic-based sentence similarity measures.

A sentence-based similarity measure could be defined by also counting for the syntactic structure of the sentence (Li and Li, 2015). Authors of this paper divide the sentence into

3 components: subject, predicate and object as the key components. However, they also include some modifier components which are attributive, adverbial, and complement. Adverbs modifier was not included in their calculation as they showed that it could not be calculated.

Practical speaking, there are text parsers that are utilized for tagging the sentence components. A recommended one is *Stanford dependencies parser* (Marneffe, MacCartney, and Manning, 2006) which is an English language dependencies representation parser. Stanford parsers also have versions for other languages like Chinese, Arabic, French and German. There is a Chinese language special semantic and syntactic parser called Language Technology Platform (LTP) (Che, Li, and Liu, 2010). It has been utilized in a related work by (Li et al., 2006) among with HowNet which is presented in Section 3.3.1. What is interesting in the syntactic-based algorithm proposed by (Li et al., 2006) was counting for negativity and antonym by having the similarity value pounded the the range [-1.0, 1.0].

3.3.3.1 Other Lexicon based Related Work in Sentence Similarity

Here is a list of some related work that could be categorised as syntactical and semantic networks based methods:

- Exploiting Synonyms to Measure Semantic Similarity of Sentence (Shin et al., 2015)
- Sentence Similarity Based on Semantic Nets and Corpus Statistics (Li et al., 2006)
- Calculation of Sentence Semantic Similarity based on Syntactic structure (Li and Li, 2015)
- A new benchmark dataset with production methodology for short text semantic similarity algorithms (O'Shea, Bandar, and Crockett, 2013)
- Text Relatedness based on a Word Thesaurus (Tsatsaronis, Varlamis, and Vazirgiannis, 2010)
- Sentence similarity based on semantic kernels for intelligent text retrieval (Amir, Tanasescu, and Zighed, 2017)

3.4 Conclusion

In this chapter, we presented some of the state-of-the-art approached in sentence representation and sentence semantic similarity. The evaluation results on many benchmarks motivate further work and improvements as many applications strongly need more powerful models to enhance their usability and performance. Methods vary based on the ultimate usage and the main category of techniques used to model the text sequence and/or the part of speech. The overall sentence modelling system is a pipeline of stages in which the designer have several optional algorithm to use for each stage. Having a state-of-the-art overview and comparison helps in guiding such design choices towards enhancing any future model. We have also presented some techniques for semantic representation of small size documents (a range of one to a few paragraphs) which are comparable to the size of a web-page, paper abstract, scientific article, or a book chapter. These techniques go beyond the classical BoW model in order to capture document level semantics versus only relying on term matching in classical search engines.

The chapter also summaries the related work that are mainly based on lexical databases dealing with terms as concepts name in a knowledge graph where concepts are connected using semantic links. These semantic links would be, for example, a linguistic relation like synonym, hyponym or hypernym as we can find in WordNet. The graph of concepts would be wider and include linked data or semantic relations whereas domains, categories or even neighbouring concepts could provide another semantic information that would help in the short text semantic representation.

Part II

Personal Contributions

Chapter 4

Evaluating Semantic Similarity between Sentences

4.1 Chapter Overview

In this chapter we talk about the challenge of evaluating semantic similarity between a pair of small texts (sentences, tweets, questions). Three main benchmarks will be presented where we will talk about our experiments with "Exercices de styles" (Raymond Queneau) and the ACL SemEval workshop. We will also talk about our model (SenSim) and what are the main categories of approaches dealing with this challenge (unsupervised, supervised and feature engineering and mixed models using Part-of-Speech (PoS), lexical databases and sentence embedding).

4.2 Introduction: Problem Statement and Applications

The semantic representation of a short text, like a sentence, is a special case comparing with a document of a few paragraphs like web-pages, articles or even books. This is mainly due to the limited number of terms in a sentence. To overcome this limitation of number of terms, we could enrich such short text with synonyms or link its terms with other semantically related concepts. Another direction towards solving this issue is to use a semantic space similar to the word embedding approach or derived from the

embedding of the words in the sentence. A third approach could use both techniques of synonym enrichment as well as semantic space representation of the sentence.

There are many useful applications of sentence representation. For example, question answering systems where the system should understand that a single question could come with many semantically equivalent variations. Another application is to cluster the top ranked results of a search engine. Sentiments analysis like in product reviews or tweets is another important application of sentence representation. In the context of digital library or scientific corpus, finding semantically similar titles or finding corresponding sections to the abstract sentences could be another useful application. It can be also used in machine translation across languages or dialects. It is also useful in recommender systems, machine translation and sentiments analysis. Finally, paraphrasing is another useful application to either generate or detect in the text.

We have mentioned a lot of related work and state of the art approaches in Chapter 2. In this chapter, we will describe a model we proposed which is a pair-wise sentence representation approach with expandable feature set that can vary from PoS and NE based alignment to sentence level representation features. The proposed approach was compared with other existing approaches on STS SemEval benchmark[1]. It has also been published as an open source software for developers and researches who want to experiment any expanded feature set.

4.3 Exercise-in-Style by Raymond Queneau

Towards exploring and designing our approach of sentence or short text semantic representation, we introduced a challenging and exciting problem to experiment with. There is a special French book by a well-known French writer and linguist called *Raymond Queneau* (1903-1976) which is entitled "*Exercices de style*" (Queneau, 1947). The book provides a very short story but in 99 different writing styles. This book has also an English version translated by Barbara Wright in 1986 (Queneau, 1986). We have experimented our sentence pairwise representation model with the English version. In this section, we will show a few examples of these styles and how did our model perform.

[1] `http://ixa2.si.ehu.es/stswiki/index.php/STSbenchmark`

4.3.1 Writing Styles and a few Samples

The book provides 99 styles. Here are a few styles of them from the English translated book (Queneau, 1986):

- **Double-entry style:**
  Towards the middle of the day and at midday I happened to be on and got onto the platform and the balcony at the back of an S-line and of a Contrescarpe-Champerret bus and passenger-transport vehicle which was packed and to all intents and purposes full. I saw and noticed a young man and an old adolescent who was rather ridiculous and pretty grotesque: thin neck and skinny windpipe, string and cord round his hat and headgear. After a scrimmage and scuffle he says and states in a lachrymose and snivelling voice and tone that his neighbour and fellow traveller is deliberately trying and doing his utmost to push him and inconvenience him every time anyone gets off and makes an exit. This having been declared and having spoken he rushes and goes towards a vacant and a free place and seat. After two hours, and a hundred and twenty minutes later, I see him and come across him again in the Cour de Rome and in front of the Gare Saint-Lazare. He is with and in the company of a friend and pal who is advising and urging him to have a button and corozo disc added and sewn onto his overcoat and mantle.
- **Dream style:**
  I had the impression that everything was misty and pearly around me, with multiple and indistinct apparitions, amongst whom however was one figure that stood out fairly clearly, which was that of a young man whose overly long neck in itself seemed to proclaim the character at once cowardly and quarrelsome of the individual. The ribbon of his hat had been replaced by a piece of plaited string. Later he was having an argument with a person whom I couldn't see, and then, as if suddenly afraid, he threw himself into the shadow of a corridor. Another part of the dream showed him walking in bright sunshine in front of the Gare Saint-Lazare. He was with a companion who was saying: "You ought to have another button put on your overcoat." Whereupon I woke up.
- **Retrograde style:** You ought to put another button on your overcoat, his friend told him. I came across him in the middle of the Cour de Rome, after having left him rushing avidly towards a seat. He had just protested against being pushed by another passenger who, he said, was jostling him every time anyone got off. This scraggy young man was wearing a ridiculous hat. This took place on the platform of an S bus which was full at noon that day.
- **Surprise style:** How tightly packed in we were on that bus platform! And how stupid and ridiculous that young man looked! And what was he doing? Well, if he wasn't actually trying to pick a quarrel with a chap who so he claimed, the young fop! kept on pushing him! And then he didn't find anything better to do than to rush off and grab a seat which had become free! Instead of leaving it for a lady! Two hours later, guess whom I came across in front of the Saint-Lazare! The same fancy-pants! Being given some sartorial advice! By a friend! You'd never believe it!
- **Litotes style:** A few of us were travelling together. A young man, who didn't look very intelligent, spoke to the man next to him for a few moments, then he went and sat down. Two hours later I met him again; he was with a friend and was discussing clothing matters.

- **Synchysis style:** Ridiculous young man, as I was on an S bus one day chock-full by traction perhaps whose neck was elongated, round his hat and who had a cord, I noticed a. Arrogant and sniveling in a tone, who happened to be next to him, with the man to remonstrate he started. Because that he pushed him he claimed, time every that got off anyone. Vacant he sat down and made a dash towards a seat, having said this. Rome (Cour de) in the I met him later two hours to his overcoat a button to add a friend was advising him.

As you may notice, all the above styles, 99 in total in the book, are talking about the same story but are written in a different way. As human beings, we can reason and relate these writing styles; understanding that they are the same story but written in different phrases. If a machine-based representation could find such semantic similarity, then we could somehow be sure that we have reached a good semantic representation model.

4.3.2 Feature Extraction and Evaluation Results on the few Styles

4.3.2.1 Corpus Generation and Feature Extraction

We mixed the corpus of the 99 styles with irrelevant short stories. We started our work efforts by developing a proof of concept using PoS parser to find (subject, verb, object, adverb) quadruples extracted from syntactical dependencies. It also uses WordNet on words to aggregate sentence similarities. This primary approach will be denoted by (SVOA) referring to the Subject-Verb-Object-Adverb quadruple.The lexical database (i.e., semantic network) WordNet was used in our SVOA approach in order to extract synonyms and antonyms (in case of a negation term like "not" are used before the PoS tagged term). The sentence PoS parser model used in our experiment was the one developed by (Amir, Tanasescu, and Zighed, 2017) which used Stanford dependency parser (Chen and Manning, 2014). The code of our SVOA approach is published on GitHub[2].

We have tried two versions of the SVOA model. In the first one we give equal weight to the subject, verb, object as well as adverb by provide the average score of the 4 part-of-speech (PoS) tags semantic similarity using WordNet. We called this first model version (SVOA-Average). In the other one, we provided an experimentally tuned weighting for these 4 PoS tags which we called (SVOA-Custom). The model that we compare with

[2] `https://github.com/natsheh/qbe`

was a TF–IDF model that takes a word ngram range of 2 to 3. The result summary of this primary model experiment is shown in Figure 4.2.

The experiment data was generated by mixing the 50 randomly selected writing styles (from Raymond Queneau 99-styles book) with 50 irrelevant short stories documents gathered from the web (resulting in a corpus of 100 documents). These irrelevant stories were selected using a Google search for short stories webpages. We have randomly selected the 50 with one condition of having similar average length to the ones of Raymond Queneau 99-styles. This way, we would avoid any possible text length related bias in the experiment.

The experiment was to use each of the writing styles documents as a query and measure how many of the relevant documents were ranked in the top. As a query, we provide one of the style document then we see how much of the 99 style documents came as top ranked versus other random stories of similar size range. In order to rank the query results, we have used *cosine similarity* between the query and the documents. In Figure 4.1 we can see a screenshot of a web demo of that experiment.

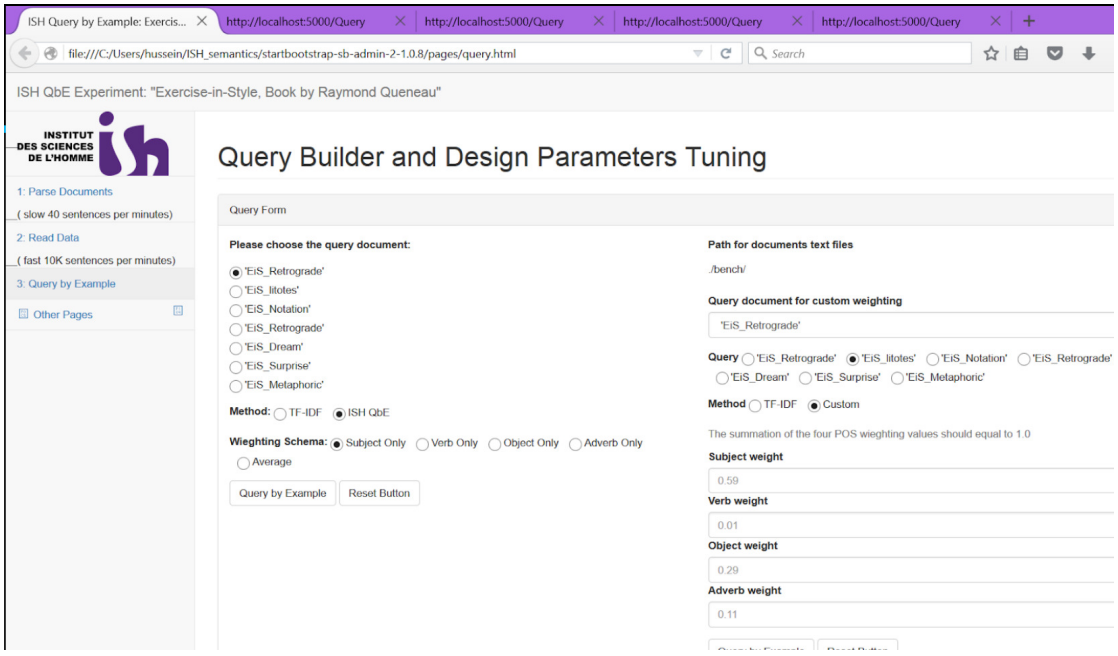


FIGURE 4.1: SVOA experiment web demo where the user can choose one of the available style documents as a query. The good model should rank the other 99 style documents in the top while in the bottom on the ranked search results they should be the irrelevant documents (other randomly selected short stories)

The evaluation was done based on information retrieval evaluation metrics which are F1-measure and area under the curve (AUC). These evaluation measure were computed out of the information retrieval precision and recall assuming a perfect case of having all 50 relevant style documents in the top of the results ranked list and all other irrelevant mixed stories in the bottom. The results were calculated by averaging evaluation score of 50 search result ranked list in which each were produced by one of the style document as a query to the corpus.

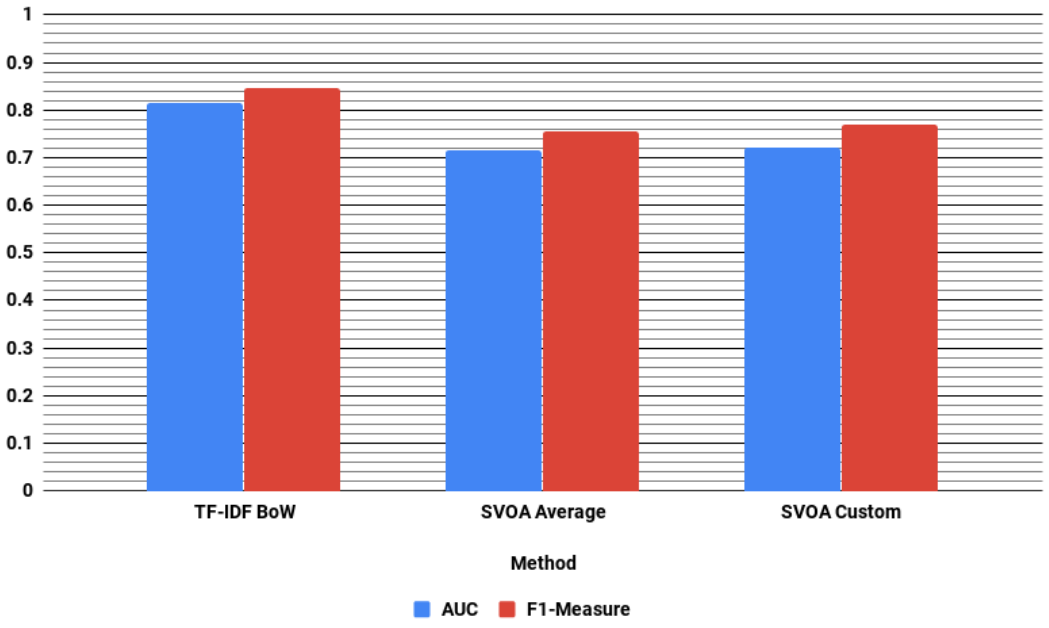


FIGURE 4.2: SVOA experiment evaluation results showing that TF–IDF model was better comparing to the SVOA approach using F1-measure and Area Under the Curve (AUC) score as two information retrieval evaluation metrics

What we can conclude from the experiment of the primary tried model SVOA tis as follow:

- Having many documents and enough vocabulary size to use (TF–IDF BoW on word n-gram range) to have very good information retrieval results.
- It is better to mix features (custom weights) since we noticed in few cases of the ranked relevant docs that TF–IDF failed to rank well compared to SVOA-Custom model

- Metaphoric style was always the style that all the primary models failed to rank high. This is kind of expected as such style is even complex to understand by a human.

4.4 SenSim model

SenSim (Al-Natsheh et al., 2017b) is a sentence pair semantic similarity estimator tool that we have developed based on a set of pair-wise linguistic features. It was developed and made accessible to the public through a workshop associated with the Association of Computational Linguistics (ACL) called SemEval for the Semantic Text Similarity (STS) shared task of the year 2017.

4.4.1 Transfer Learning Approach

According to a comparable transfer learning strategy (Bottou, 2014), the trained pair-wise transformer (i.e., features extractor) of a model, whose comparator can well predict the similarity of a sample pairs, can be reused to easily train a classifier for the annotation of a single sample. if we are able to build a model consisting in (1) a pairwise transformer (i.e., feature extractor), and (2) a comparator that can well-predict if the two elements of the input are of the same class or not, then the learned transformer could be reused to easily train a classifier to label a single element. A good example to understand such system is face recognition. It is considered impossible to have all human faces images to train the best features set of a face; however, a learned model that can tell if two given face-images are of the same person or not could guide us to define a set of good representative features to recognize a person given one face image. We can generate $\frac{2^n}{2}$ comparative pairs from n examples. Similarly, we cannot have all possible sentences to identify the sentence semantics, but we can generate a lot of comparative sentence pairs to learn the best semantics features set, i.e., sentence dense vector representation. Thus we consider our pairwise feature-based model as an initial step to build a sentence dense vector semantics representation that can perform very well in many applications like semantics highlighter, question answering system and semantics-based information retrieval system.

Algorithm 1 The pairwise features extraction process of aligned PoS and NE tagged tokens.

Input: Sentence pair
Extract a PoS type or a NE type word tokens from both sentences
- Pair each tagged word-token in one sentence to all same tagged tokens in the other sentence
- Get the word vector representations of both tokens of each paired tokens
- Compute the vector representations of both tokens of each paired tokens
- Align words if the cosine similarity (CS) is above a threshold value
- Solve alignment conflicts, if any, based on the higher CS value

Compute the average CS of the aligned tokens and use it as the pairwised feature value

4.4.2 Feature Engineering and Model Description

Our approach is based on the comparable transfer learning systems discussed in section 4.4.1. Accordingly, our model pipeline mainly consists in 2 phases: (1) pairwise feature extraction, i.e., feature transformer, and (2) regression estimator. While many related work either use words embedding as an input for learning the sentence semantics representation or learning such semantics features directly, our model is able to reuse both types as input for the pairwise feature transformer. For example, as listed in Table 4.1, we used features that is based on word vectors similarity of aligned words while we also have a feature that considers the whole sentence vector, i.e., sparse BoW. The model can also use, but not yet used in this paper, unsupervised learned sentence representation out of methods like BoW *matrix decomposition*, *paragraph vector*, or *sent2vec* methods as input to our pairwise features transformer.

4.4.2.1 Pairwise Feature Extraction

We used different feature types as in Table 4.1. The first two types are based on aligning PoS and NE tagged words and then compute the average word vectors cosine similarity (CS) of the paired tags. The process of extracting these type of pairwise features are resumed in the algorithm 1.

The third feature is extracted by transforming each sentence to its BoW vector representation. This sparse vector representation is weighted by TF–IDF. The vocabulary of the BoW is the character grams range between 2 and 3. This BoW vocabulary source is only the data set of the task itself and not a general large text corpus like the ones usually used for word embedding. We are planning to try out a similar feature, but

	Feature	Pair Combiner	Importance
1	Aligned PoS tags (17 tags)	Average of w2v CS of all PoS tag pairs	0.113
2	Aligned NE tags (10 tags)	Average of w2v CS of all NE tag pairs	0.003
3	TF–IDF char ngrams BoW	Cosine similarity of the sentence BoW vector pair	0.847
4	Numbers	Absolute difference of the number summation	0.006
5	Sentence length	Absolute difference of the number of characters	0.032

TABLE 4.1: Pairwise features set.

unsupervised, where we consider a corpus like Wikipedia dump as a source for the BoW. Another feature we plan to consider as a future work is the dense decomposed BoW using SVD or NMF. Finally, we can also consider unsupervised sentence vectors using *paragraph vectors* or *sent2vec* methods.

Features number 4 is extracted by computing the absolute difference of the summation of all numbers in each sentence. To achieve that, we transferred any spelled number, e.g., "sixty-five", to its numerical value, e.g., 65. The fifth pairwise feature we used was simply based on the sentence length.

4.4.2.2 Learning Model

We have mainly evaluated two regression estimators for this task. The first estimator was random forests (RF) and the other was Lasso (least absolute shrinkage and selection operator). Based on a 10-fold cross-validation (CV), we set the number of estimators of 1024 for RF and a maximum depth of 8. For Lasso CV, we finally set the number of iterations to 512.

4.4.3 The Generic Model in Practice (Demo Examples)

In this section, we will show a few examples where we have applied and experimented our model "SenSim" to evaluate the sentence semantic similarity. Two main demo applications are presented: First, finding the top semantically similar sentence among the styles of "Exercise-in-Style" book. The second is a sentence semantic highlighter application that pairs each sentence in the paper abstract to its corresponding semantically related sentence in the paper content (e.g., paper introduction). The score is bounded between [0 and 5], where 5 is the highest possible score that means the two sentences are semantically identical.

4.4.3.1 Exercise-in-Style

Sent1: "This scraggy young man was wearing a ridiculous hat."

Sent2: "And how stupid and ridiculous that young man looked!"

Sentencerelatedness score: **3.11**

Sent1: "I came across him in the middle of the Cour de Rome, after having left him rushing avidly towards a seat."

Sent2: "After two hours, and a hundred and twenty minutes later, I see him and come across him again in the Cour de Rome and in front of the Gare Saint-Lazare."

Sentence relatedness score: **3.55**

Query: "This scraggy young man was wearing a ridiculous hat."

"How tightly packed in we were on that bus platform!" *score*: 0.68

"And how stupid and ridiculous that young man looked!" *score*: **3.11**

"And what was he doing?" *score*: 1.24

"Well, if he wasn't actually trying to pick a quarrel with a chap who so he claimed, the young fop" *score*:0.68

"kept on pushing him!" *score*: 1.37

"And then he didn't find anything better to do than to rush off and grab a seat which had become free!" *score*: 0.75

"Instead of leaving it for a lady" *score*: 0.79

"Two hours later, guess whom I came across in front of the Saint-Lazare!" *score*: 1.22

"The same fancy-pants!" *score*: 1.17

"Being given some sartorial advice! " *score*: 0.54

"By a friend " *score*: 1.52

"You'd never believe it!" *score*: 1.45

Query: "I came across him in the middle of the Cour de Rome, after having left him rushing avidly towards a seat."
"Towards the middle of the day and at midday I happened to be on and got onto the platform and the balcony at the back of an S-line and of a Contrescarpe-Champerret bus and passenger-transport vehicle which was packed and to all intents and purposes full." *score*: 1.94
"I saw and noticed a young man and an old adolescent who was rather ridiculous and pretty grotesque: thin neck and skinny windpipe, string and cord round his hat and headgear." *score*: 0.41
"After a scrimmage and scuffle he says and states in a lachrymose and snivelling voice and tone that his neighbour and fellow traveller is deliberately trying and doing his utmost to push him and inconvenience him every time anyone gets off and makes an exit." *score*: 1.15
"This having been declared and having spoken he rushes and goes towards a vacant and a free place and seat." *score*: **2.48**
"After two hours, and a hundred and twenty minutes later, I see him and come across him again in the Cour de Rome and in front of the Gare Saint-Lazare." *score*:**3.55**
"He is with and in the company of a friend and pal who is advising and urging him to have a button and corozo disc added and sewn onto his overcoat and mantle. " *score*: 0.94

Noteworthy, the sentence-level semantic similarity could be used to find the document-level similarity. This could be achieved for example by aggregating the similarity score of the semantically paired sentences between each document pair. This approach was also applied and validated to the "Exercise-in-Style" book where the results were aligned as expected. However, among the experimented styles, only the results of the style "Metaphorically" were not encouraging. This is due to the challenging task of this style in which even a human would find it difficult to semantically relate metaphoric forms.

In order to have an idea of the semantic relatedness difficultly of this single case, below we can see the metaphorically style from the book. Even for a human reader, it is very challenging to semantically relate it to the other styles of the same short story (check the other style examples in Section 4.3.1):

In the centre o f the day, tossed among the shoal of travelling sardines in a coleopter with a big white carapace, a chicken with a long, featherless neck suddenly harangued one, a peace-abiding one, of their number, and its parlance, moist with protest, was unfolded upon the airs. Then, attracted by avoid, the fledgling precipitated itself thereunto. In a bleak, urban desert, I saw it again that self-same day, drinking the cup of humiliation offered by a lowly button.

4.4.3.2 Abstract-Introduction Sentence Semantic Highlighter

Here, we took a sample paper (Zighed, Lallich, and Muhlenbach, 2005) in order to evaluate our sentence semantic similarity model as a semantic highlighter between the sentences in the paper abstract and their corresponding sentences in the paper content.

Sent1: "First, we build a geometrical connected graph like Toussaint's Relative Neighbourhood Graph on all examples of the learning set."
Sent2: "At first, they build a multidimensional neighbourhood structure by using some particular models like the Toussaint's Relative Neighbourhood Graph (Toussaint 1980)."
Sentence relatedness score: **3.49**

Abstract – Introduction (Top 3 out of 23):
"At first, they build a multidimensional neighbourhood structure by using some particular models like the Toussaint's Relative Neighbourhood Graph (Toussaint 1980)." *score*: **3.49**
"Recently, Sebban (Sebban 1996) and Zighed (Zighed and Sebban 1999) have proposed a test based on the number of edges that connect examples of different classes in a geometrical neighbourhood." *score*: 2.44
"They calculate thereafter the number of edges that must be removed from the neighbourhood graph to obtain clusters of homogeneous points in a given class." *score*: 2.44
. . .
"Finally, they have established the law of the edge proportion that must be removed under the null hypothesis, denoted A 1/4 , of a random distribution of the labels." *score*: 0.47
"Kruskal and Wallis have defined a nonparametric test based on an equality hypothesis of the scale parameters (Aivazian, Enukov, and Mechalkine 1986)." *score*: 0.25
"This reliability is generally evaluated with a posteriori test sample a O. " *score*: 0.22

4.4.4 SemEval STS Benchmarking

Semantic Textual Similarity (STS) is a shared task that has been running every year by SemEval workshop since 2012. Each year, the participating teams are encouraged to utilize the previous years data sets as a training set for their models. The teams are then ranked by their test score on a hidden human annotated pairs of sentences. After the end of the competition, the organizers publish the gold standards and ask the teams of the coming year task to use it as a training set and so on. The description of STS2017 task is reported in (Cer et al., 2017). In STS2017 , the primary task consisted in 6 tracks covering both monolingual and cross-lingual sentence pairs for the languages Spanish, English, Arabic, and Turkish. Our team, UdL, only participated in the English monolingual track (Track 5).

The data consisted in thousands of pairs of sentences from various resources like (Twitter news, image captions, news headline, questions, answers, paraphrasing, post-editing...). For each pair, a human annotated score (from 0 to 5) is assigned and indicates the semantic similarity values of the two sentences. The challenge is then to estimate the semantic similarity of 250 sentence pairs with hidden similarity values. The quality of the proposed models would then be evaluated by the Pearson correlation between the estimated and the human annotated hidden values.

We experimented different settings varying the feature transformation design parameters and trying out three different training set versions for RF. We show the 3 selected settings for submission and the test score of a few evaluation data-sets from previous years in Table 4.2.

Model	dataset	DF	PoS	vectors	images	AS	H16	AA	QQ	plagiarism	mean
-	small	no	polyglot	spaCy	0.85	0.77	0.80	0.47	0.54	0.82	0.71
-	small	yes	polyglot	spaCy	0.82	0.75	0.79	0.53	0.56	0.84	0.72
Run2	big	yes	spaCy	spaCy	0.82	0.74	0.79	0.54	0.61	0.84	0.72
-	big	yes	polyglot	spaCy	0.82	0.75	0.79	0.52	0.55	0.84	0.71
-	big	no	spaCy	spaCy	0.82	0.78	0.80	0.46	0.60	0.82	0.71
-	big	no	polyglot	spaCy	0.85	0.77	0.80	0.51	0.56	0.82	0.72
Run1	big	no	polyglot	spaCy	0.85	0.77	0.80	0.46	0.54	0.82	0.71
Run3	BH	no	polyglot	spaCy	0.85	0.77	0.80	0.51	0.58	0.82	0.72
-	BH	no	polyglot	GloVe	0.85	0.77	0.80	0.46	0.57	0.81	0.71

TABLE 4.2: Evaluation 2-decimal-rounded score on some testsets. DF: domain feature, AA:answer-answer, AS:answers_students, H16:headlines_2016, QQ:question-question, BH:bigger data set size where hash-tags are filtered

The evaluation results shows that Random Forest regression estimator on our extracted pairwise features provided 80% of Pearson correlation with hidden human annotation values. The model was implemented in an able to scale pipeline architecture and is now made available to the public where the user can add and experiment any additional features or even any other regression models.

In SemEval shared task, our team was named as UdL. According to the task regulations, we have submitted three runs of our model UdL for the task official evaluation (similar to the other participating team). The settings of these three runs are shown in Table 4.2. The summary of the evaluation score with the baseline (0.7278), the best score run model (0.8547), the least (0.0069), the median (0.7775) and the mean (0.7082) are shown in Figure 4.3. Run1 was our best run with Pearson correlation score of (0.8004), At this run, we used RF for regression estimator on our all extracted pairwise features except the domain class feature. Run2 (0.7805) was same as Run1 except that we used the domain class feature. Finally, Run3, submission correction phase (0.7901), used a different data set were we filtered-out hash-tag symbol from Twitter-news sentence pairs.

4.4.5 Open-source and Future Work

We published the software with the dataset of SenSim as an open-source tool [3] in order to allow further feature experiment by interested researchers. It also allows to increase the dataset and to benchmark on SemEval STS task as well as to show the examples of Raymond Queneau's exercises in style documents experiment as well as an experimented example of the abstract-to-paper-content sentence semantic highlighter that we showed earlier in this chapter.

4.5 Conclusion

In this chapter, we proposed SenSim, a model for estimating sentence pair semantic similarity. The model mainly utilizes two types of pairwise features which are (1) the aligned part-of-speech and named-entities tags and (2) the TF–IDF weighted BoW vector model of character-based n-gram range instead of words. The evaluation results shows that Random Forest regression estimator on our extracted pairwise features provided 80% of

[3] `https://github.com/natsheh/sensim`

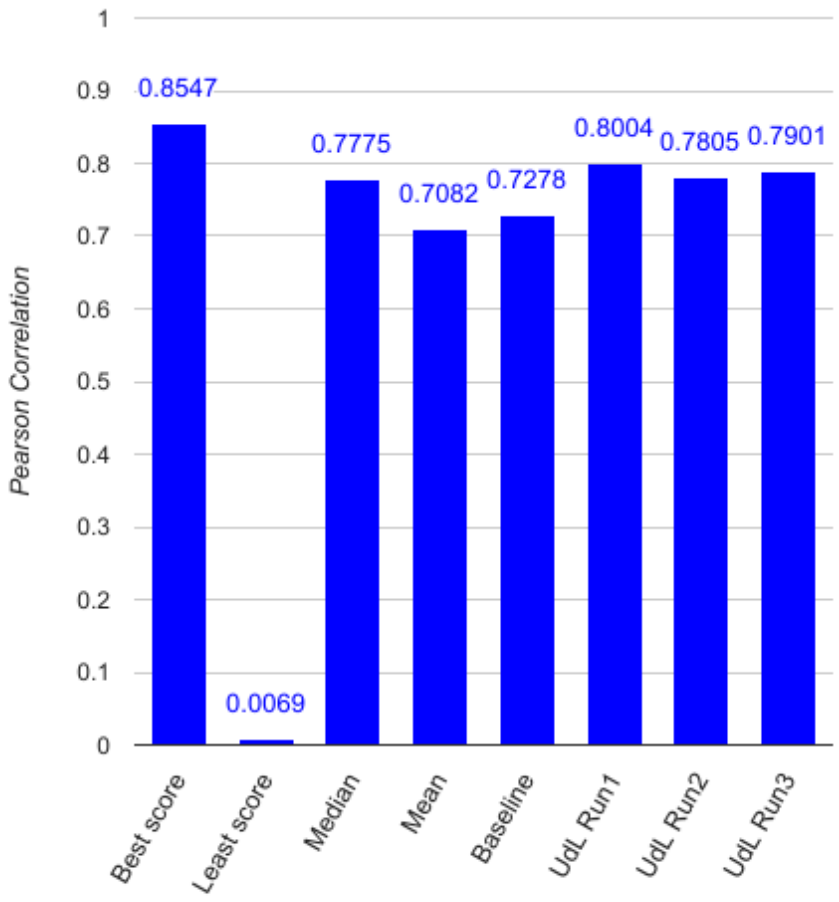


FIGURE 4.3: Track 5 results summary in comparison to UdL three runs;*: submission correction.

Pearson correlation with hidden human annotation values. The model was implemented in a scalable pipeline architecture and is now made available to the public where the user can add and experiment any additional features or even any other regression models. Since the sentence vector BoW-based pairwise feature showed high percentage in the feature importances analysis of the Random Forest estimator, we are going to try other, but dense, sentence vector representation, e.g., in (Shen et al., 2014; Le and Mikolov, 2014). We are also planning to use and evaluate the model in some related applications including a semantic sentences highlighter, a topic-diversified document recommender system as well as a question-answering system.

Chapter 5

Semantic-based Paper Recommendation and Scientific Corpus Expansion

5.1 Chapter Overview

Chapter 5 will present our approach in semantic-based paper recommendation and scientific corpus expansion. The chapter takes as a case study our work on interdisciplinary research topic (mental rotation). This use case was part of a project that was performed in collaboration with the STAPS[1] in Claude Bernard University Lyon 1 (UCBL).

5.2 Introduction

Scientific research carried out in academic or industrial circles produces ever more numerous knowledge. In order to make their original contributions to this knowledge in their daily activity, researchers must be able to access existing knowledge (publications, patents) and to find among them the relevant elements (theories, demonstrations, methodologies, experimental results, etc.) which serve as a basis for their work and make it possible to delimit the frameworks of their own scientific contributions. It is now possible to benefit from the mass of articles on the web or specialized scientific digital

[1]The Training and Research Unit in Science and Technology of Physical and Sports Activities

libraries (SDLs). Examples include web portals on general publications or specific domains (e.g., *PubMed* for references and abstracts on life sciences and biomedical topics, *CiteSeer*^{X} in computer science and information science), with subscription (e.g., Elsevier *ScienceDirect* and *Scopus*, *Web of Science*, Springer *Online access Journals*, *ACM Digital Library*, *Google Scholar*) or without subscription (e.g., *arXiv* or the *Directory of Open Access Journals*). These SDLs are both the instrument and the raw material for scientific research and innovation. Therefore, the control and effective use of these sources is a strategic challenge for the development of science, the increase of economic wealth (May, 1997), and more broadly the evolution of society.

The problem is that this search for relevant scientific articles is essentially done, either by entering specific keywords in search engines as well as database query systems or by studying recent articles published in referenced journals, major conference proceedings in a given field, or using the citation network study of the bibliographic references cited in articles considered to be of interest.

As a result, the exploration of these gigantic SDLs is not effective. On the one hand, this exploration is limited to a focus on the most recent articles, whereas old articles could prove to be relevant. On the other hand, the articles returned by these interrogation systems are most often limited to the scientific community belonging to the researcher, whereas articles coming from complementary disciplines could be interesting. For example, when the data miners and computer scientists became interested in the field of social network analysis with the arrival of major social networks such as *Facebook* after the mid-2000s, most of them were surprised to discover that it was essential to take into account the work carried out in physics in the field of the complex systems study (Girvan and Newman, 2002).

Thus, with the current interrogation techniques of the SDLs of scientific articles, the field of exploration of researchers is very restricted: researchers are trapped in a filter bubble (Pariser, 2011) which limits them to what they know, to what they expect to find, leaving no space for diversity or surprise.

5.3 Related Work

According to Beel *et al.*'s initial study in 2013 ("Research Paper Recommender System Evaluation: A Quantitative Literature Survey"), recommender systems for research papers are becoming increasingly popular. In continuing their exploration in 2016 (Beel et al., 2016), they studied 200 research articles on research-paper recommender-system domain, and identified 7 main approaches in this field:

- Stereotyping,
- Content-based Filtering,
- Collaborative Filtering,
- Co-Occurrence,
- Graph-based,
- Global Relevance, and
- Hybrid.

Despite having different approaches, the research-paper recommender-system community focuses almost exclusively on accuracy. The implicit assumption is that an accurate recommender system will lead to high user satisfaction. In other areas of research on recommendation systems (e.g., in music recommender systems), criteria other than accuracy are sought. One such criterion that can make users unsatisfied is the lack of diversity (Ziegler et al., 2005). For a music streaming service, there is diversity when the list of recommended music includes songs of different music styles rather than different songs of styles which the user is already used to listening to (Castells, Hurley, and Vargas, 2015). In the field of recommending research papers in multidisciplinary digital libraries, we can extend this idea of diversity to disciplines that are not necessarily those to which the user of the system belongs.

In order to carry out research, it is important for a new researcher to know well the field he wants to contribute to. Nevertheless the amount of literature and approaches represents a problem for him because it is difficult to identify which of the articles are most relevant so as not to make a discovery that has already been made elsewhere.

To facilitate the exploration of the articles in the scientific digital libraries, numerous works were carried out following several tracks. Most often they rely on topic modeling realized with latent Dirichlet allocation (Blei, Ng, and Jordan, 2003). This modeling is used to establish a similarity between the documents, this similarity is then used to link the documents together in different ways, such as a graph, and then allow a graphical exploration of this graph (Klein, Eisenstein, and Sun, 2015; He et al., 2016; Le and Lauw, 2016). Some approaches focus on the human aspects of the document exploration interface (Gretarsson et al., 2012), others tend to detect the evolution of scientific topics in the time (He et al., 2009), or try to promote serendipity (Alexander et al., 2014).

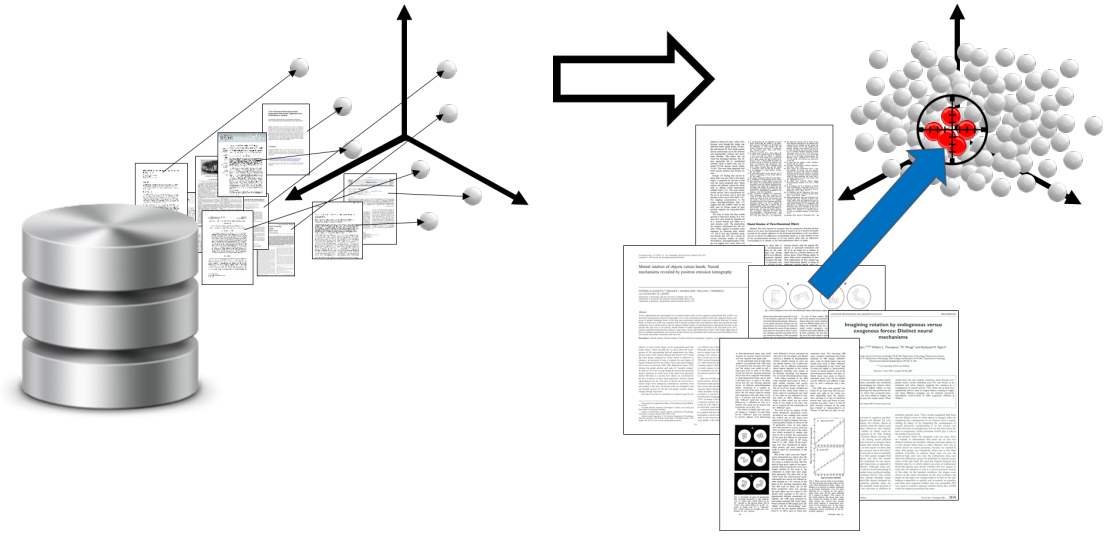
The concept of query-by-examples (QbE) search was mainly used in multimedia information retrieval systems (Kabary et al., 2013). QbE and More-like-this query (MLT) (Hagen and Glimm, 2014; Dong and Smyth, 2016) may address the task we are trying to solve which is providing the user with similar documents to the one he or she provides as a query. MLT exists as a feature in Apache Lucene used by ElasticSearch (Dixit, 2017). This special type of query finds documents that are "like" a given set of documents. In order to do so, MLT selects a set of representative terms from these input documents, forms a query using these terms, executes the query and returns the results. The user controls the input documents, how the terms should be selected and how the query is formed. MLT is part of a family of similarity supporting queries that provide the ability of searching for similar documents to the one(s) passed to the query.

5.4 SSbE Model

5.4.1 Semantic-Similarity Shadow Hunter

We propose a content-based research-paper recommender system called *3SH* for "Semantic-Similarity Shadow Hunter." This model is based on the transformation of documents (here, scientific articles of SDLs) into a vector form that emphasizes the semantic similarity between the texts. Indeed, semantic relatedness between units of language (e.g., words, sentences, or texts) can be estimated using a vector space model. For that it is necessary to construct the vectorized representation of the documents of the multidisciplinary SDL, as shown on the left part of the Figure 5.1. The advantage of such

FIGURE 5.1: Construction of the vectorized representation of the documents from the multidisciplinary SDL. Once the representation space is built, we can use 3SH model: some specific scientific papers are in the viewfinder of the researcher and projected in the documents vectorized representation (red dots) for highlighting some interesting topics.

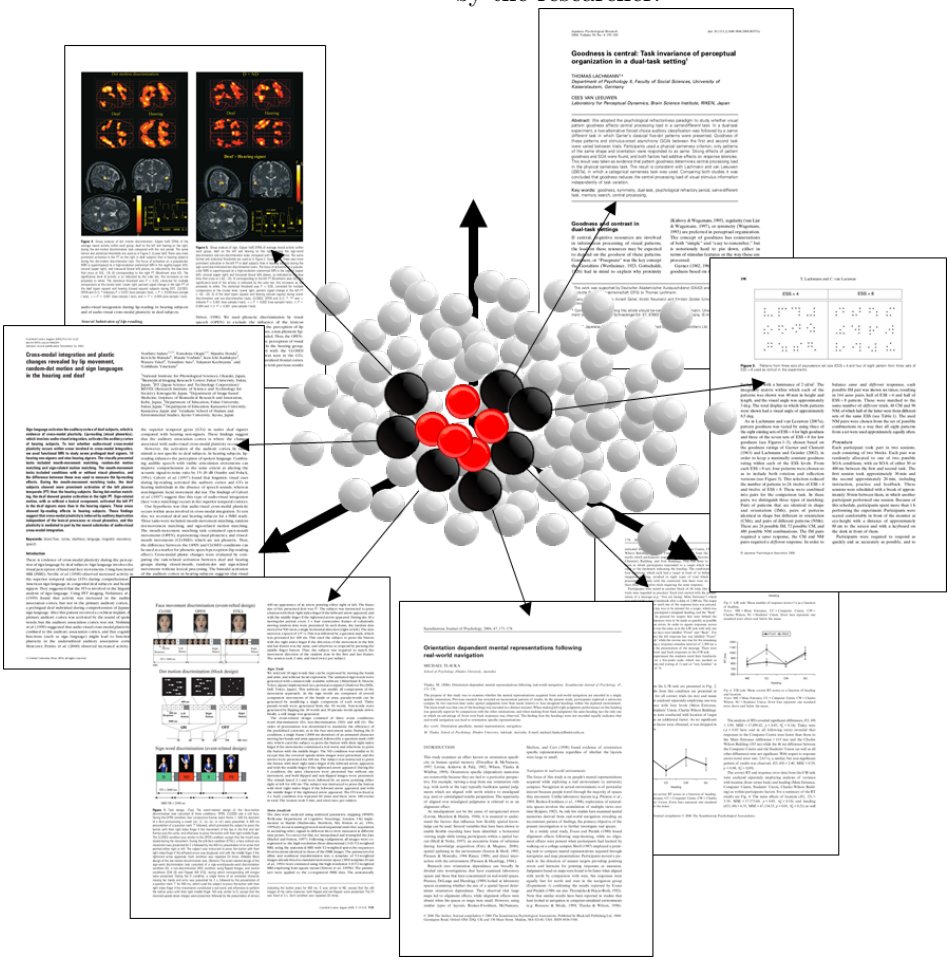


a representation is that the semantically close articles will be close in this vectorized representation space, whether these articles come from the same discipline or not.

The user of the model is a researcher who wants to work on a specific topic. His/her approach is analogous to a hunt: finding relevant articles related to the topic that he or she is interested in but lost in the mass of papers present in digital libraries is like tracking wildlife animals hidden in a natural environment. Instead of entering keywords into the system, the researcher presents some example papers that are interesting for him/her, i.e., the topics he or she has in the viewfinder, as shown on the right part of the Figure 5.1.

With this initial scientific corpus, our system is able to find and recommend semantically similar papers that do not necessary contain same terminologies but similar concepts. Highlighting some specific articles (the target articles supplied by the system user as input) allows to "hunt" other relevant articles, present in the semantic neighbourhood of the latter, but lying in the shadows and not accessible in the usual way (Figure 5.2). The use of some keywords restricts indeed the viewfinder too much and induces a lack of diversity, which means that the researchers are unable to reach articles that are interesting for their research topic.

FIGURE 5.2: Extension of the scientific corpus by recommending the semantically close papers located in the shadows (black dots) of the target articles (red dots) enlightened by the researcher.



"Semantic Search-by-Examples", or *SSbE*, is the name we give to the method we propose to solve the problem of scientific domain expansion. In the following sections we will present each element in the pipeline of the model. We will show the model in two stages. The first stage would be denoted by $SSbE_p$, which is the **p**artial pipeline (without the active learning process). The second stage, denoted by *SSbE*, would be the completed pipeline (with the active learning process).

5.4.2 Model Overview

The purpose of this model is to expand a bibliography of a certain *focus scientific topic*. Such a topic is defined by a set of articles and possibly a topic label that we denote as a *topic key-phrase*, e.g., "human machine interface," "breast cancer," or "biological water treatment."

As it is presented in Figure 5.3, we define the input of the model from two main sources: the *scientific corpus* and the *seed articles* belonging to the topic. The scientific corpus should have a big amount of articles from many disciplines. Those articles may be extracted from a scientific digital library using the classical search engine (i.e., term matching search engine that is based on TF–IDF scoring) used by the domain experts. It could be considered as the maximum number of articles they were able to find using such method. They aim, the domain experts, to expand the set of these founded articles ,i.e., seed articles, to more relevant results. In order to be used in our approach, the seed articles data must have a minimum set of metadata like the title, the abstract, and a unique index to be retrievable. It is possible to benefit from other metadata fields like the set of keywords, authors, references but they are not required for the model. In order to be able to evaluate the model, the content of the articles would be necessary, so that the expert annotator could provide their feedback. The set of *seed articles* consists in a few number of examples, preferably between 100 and 300, with the same requirements as those of the scientific corpus, i.e., a title, an abstract and a unique index. These seed articles are provided as a kind of query-of-examples in which the user aims to find semantically similar articles possibly from other disciplines. Practically, seed articles are articles belonging to the *scientific corpus*, or which can be added to it, and that are annotated as *focused topic*. This set of articles is retrieved by matching the *topic key-phrase* with the metadata of the articles of the *scientific corpus*. We will denote this set of articles as *extended positive articles*.

The SSbE model consists in a few high-level phases illustrated in Figure 5.3. The output is the ranked list of recommended articles that may extend the knowledge about the focus scientific domain by including semantically relevant articles from other disciplines. The first process in the model is to vectorize the whole corpus in addition to the seed articles using the bag-of-words (BoW) method. The second step is to transform the BoW into the vector semantics dense representation. Next, a balanced dataset will be generated from both positive examples, that are the *seed* and the *extended positive articles*, and negative examples, that are randomly selected from the *scientific corpus* other than the *matched key-phrase* articles. This dataset is then used to train a supervised binary classifier. The trained classifier is finally used to rank all the articles of the *scientific corpus* with the probability of belonging to the *focused topic*. A complementary enhancing step is the active learning process where the user feedback is used to regenerate the balanced

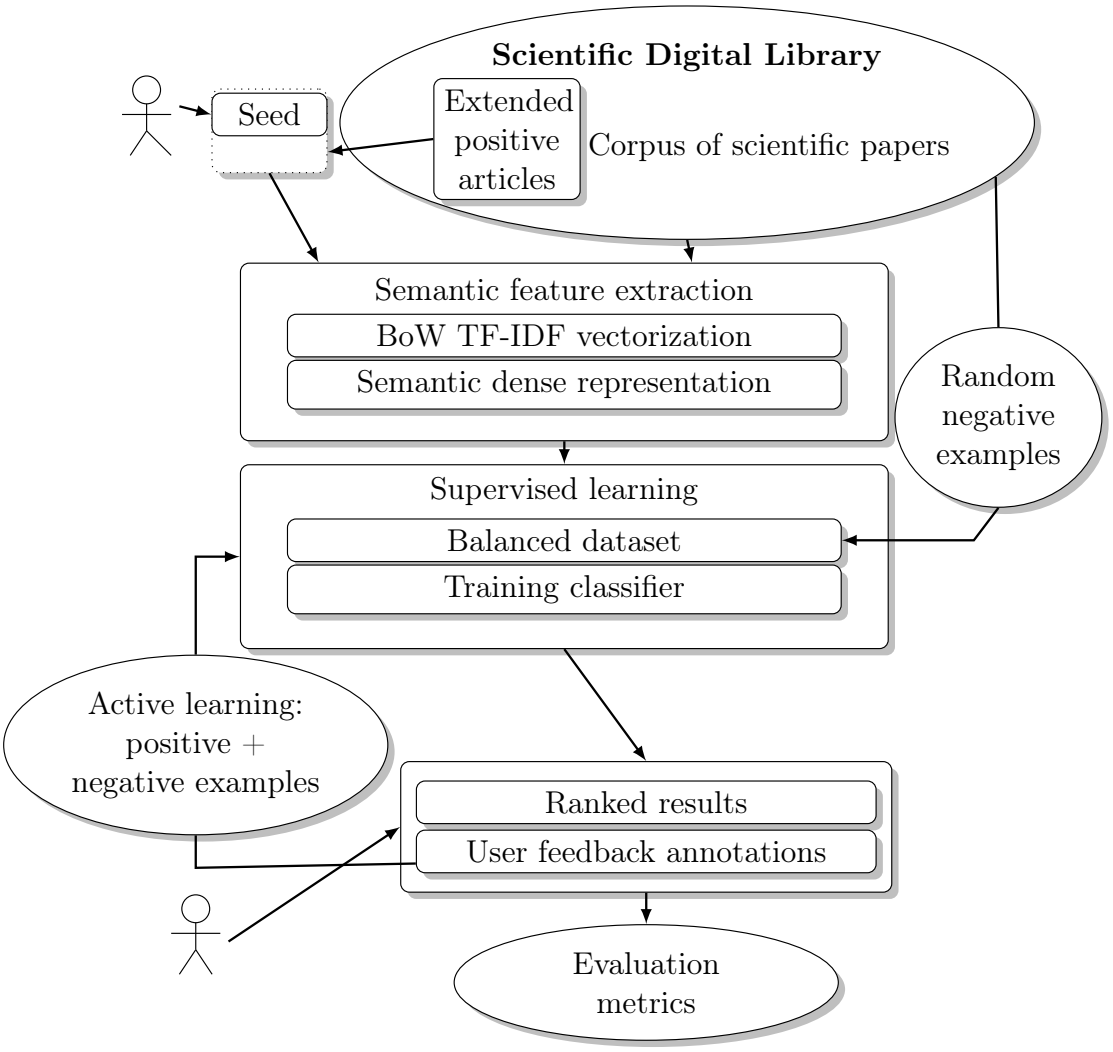


FIGURE 5.3: The *SSbE* Model Pipeline. The input of the system is an initial corpus that consists in the seed articles and the extended positive examples which are search key-phrase matches to the focus scientific topic. After transforming all the articles into their semantic feature representations, a supervised learning classifier is trained on a balanced set of positive (initial corpus) and negative (randomly selected) article examples. The results are then ranked by the probability value that the trained binary classifier predicted each article in the digital library as the positive class. Finally the user provides his annotation on the top results which are used to regenerate a new training set with negative examples with the active learning process to enhance the results in which the top ranked results would be the output scientific topic expanded corpus

training dataset.

5.4.3 Vectorization

Our system uses the common TF–IDF weighted BoW method to initially obtain a vectorized representation of the documents. The main drawback with BoW vectorization

is that the information of the order of the words in the text is lost. Although there are few techniques to overcome this issue, i.e., n-gram, using BoW alone would still require to encode the semantic and syntactic information (Le and Mikolov, 2014).

The system then extracts the dense semantic representation from the weighted BoW. This could be done either by Latent Semantics Analysis (LSA), based on decomposition, or by a technique based on learning semantic features (Le and Mikolov, 2014). In order to find a good semantic representation space, we computed the average inner cosine similarity (*AICS*) as in Equation 5.1 for two lists of documents: the positive ones, and negatives ones. The negative list is constructed by randomly selecting the same number of documents from the corpus. We maximize the function that is given in Equation 5.2 searching for a good semantics space transformer.

$$\text{AICS} = \frac{\sum_{i=1}^{n-1} \text{cosine\_similarity}(\text{list}[i], \text{list}[i+1:n])}{\text{number\_of\_comparisons}} \tag{5.1}$$

$$\text{argmax}_{\text{transformer}}(\text{AICS}_{\text{positive\_list}} - \text{AICS}_{\text{random\_list}}) \tag{5.2}$$

The vector semantics transformation is constitutionally a long and expensive process, however it is luckily needed to be run only once. This is true not only for a certain *focused topic* use case but also for any other *focused topic* that the users would like to apply later on if the *seed articles* are found in the same corpus.

5.4.4 Learning Process

5.4.4.1 Balanced Training Set Generation

After transforming all the articles corpus (i.e., title + abstract) into its semantic vectorized representation, our method relies on building a classifier that would be able to predict if a given example is part of the *focused topic* or not. To build such classifier, we built a balanced training set of both positive and negative examples. At the beginning, the negative examples are randomly sampled from the corpus excluding positive examples. This is of course based on the assumption that a uniformly randomly picked sample of examples from such corpus would less likely be positive examples. In case the

number of positive examples is small, even after adding the *extended positive articles*, the system randomly duplicates some positive examples to match the experimented size of the dataset. At the stage of active learning, this balanced dataset would be regenerated with better negative examples provided by the user feedback.

5.4.4.2 Supervised Learning

In order to generate the aimed results, our method uses a binary classifier trained on the generated balanced dataset in order to compute the prediction probability of each article in the *scientific corpus* to belong to the *focused topic* class. A ranked list of all the corpus articles, sorted by that probability value as a score, is finally considered as the system output. This result excludes all the positive examples used in the learning process, as the aim is to find any unexpected relevant article with our semantic-based recommendation approach.

Choosing the type of the classifier is a design parameter of the model and could be decided experimentally. We recommend ensemble learning methods like gradient boosting or random forest because of the ability of such methods to provide the predicted probability value of a document to belong to the class. Otherwise, regression could also be used.

Unlike the vector semantics transformation phase, the supervised learning is a very fast and repeatable process which is practically very useful for the *active learning* process.

5.4.4.3 Active Learning

In this complementary but important process, the user feedback is used to regenerate the balanced training dataset. This aims to extend the negative examples with the related but marked-irrelevant results by the user. The positive examples will also be enriched by providing marked-relevant articles, but from different disciplines. Accordingly, the classifier will continuously learn how to semantically separate the articles in a better way than only using randomly sampled negative examples as in the first generated dataset.

In case of many users providing relevancy annotation to the results, the model computes the average score for each annotation. The numeric value used to indicate relevance is 1, 0 in the case of an irrelevant document, and 0.5 when experts can not decide.

5.4.4.4 Using Sentence Semantic Relatedness for Evaluation

In order to avoid asking the user to provide his feedback annotation, which is not an easy task, we introduced an automated comparative evaluation criterion of the results after applying the active learning process. This criterion is based on sentence semantic relatedness (Agirre et al., 2016) between the titles of the seed articles and the titles of the results. We first use the Cartesian product composed of the titles of the seed and the results articles set to generate the set of titles pairs. Then, we use a pre-trained model that takes a set of sentences (i.e., title pairs) as input and that estimates the semantic relatedness score for each pair as an output. Our proposed evaluation method is then to count the title pairs that exceed a semantic relatedness score threshold.

5.5 Experimentation

5.5.1 Use Case from Sports Science: Mental Rotation

For our experiments, we chose to focus on a field of research far from our own field (i.e., computer science) for which there were possibilities of trans-disciplinary inputs because this field is already in close connection with related disciplines. This research discipline is *sports science*. This field is interconnected with other scientific domains (e.g., physiology, psychology, anatomy, biomechanics, biochemistry and biokinetics).

We have been able to find sports science specialists who proposed a research topic of their interest and for which they are experts: the *mental rotation* domain. "Mental rotation" is a psychological task proposed for the first time by Shepard and Metzler in 1971 to account for certain mental abilities to manipulate images. As presented in Figure 5.4, this experimental task was designed to measure the time that human subjects require to determine the shape identity between figures presented in different orientations. They discovered that the response time is a function of the angular difference in the portrayed orientations of the two three-dimensional objects.

Interest for sports science researcher in mental rotation is manifold: mental image transformations sometimes implicate motor processes and sometimes not (Kosslyn et al., 2001), mental rotation performances are connected to motor abilities and experimental studies suggest that improved mental rotation performance would promote the ability

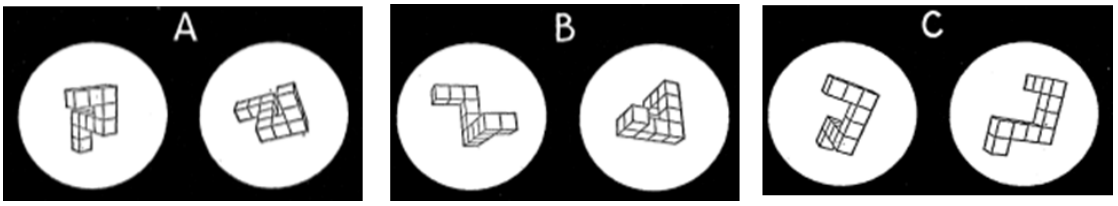


FIGURE 5.4: Mental Rotation: Examples of pairs of perspective line drawings presented to the subjects. (A) A "same" pair, which differs by an 80° rotation in the picture plane; (B) a "same" pair, which differs by an 80° rotation in depth; and (C) a "different" pair, which cannot be brought into congruence by *any* rotation (Shepard and Metzler, 1971).

to quickly perform complex motor rotations, as can be seen in the body movements of professional athletes.

The research topic "Mental Rotation", is studied by many disciplines and research communities as a research problem but each of them consider it from their own perspective. Some of such disciplines are cognitive sciences, aerodynamics and sport sciences. Researchers have an issue in the different terms used by different research community in their publications. Thus, it causes a barrier for them to retrieve all related publication using a limited number of search keywords. For instance, in sport sciences research they use the terminology "Mental Rotation", however, the same concept is sometimes used as "Spatial Abilities" in cognitive sciences. This interdisciplinary property of the field of "Mental Rotation" makes is a very interesting use case for our proposed model towards solving the trans-disciplinary research challenge.

5.5.2 Data Description

To test the model above, we chose to apply it to a very concrete question. How can we enlarge the bibliography about a given topic when the vocabulary is not stable (difficult to use specific key words), given some sample articles by experts of the domain? We applied this question to the specific domain of "Mental Rotation" which is fully across disciplines. The goal for the expert is to benefit from the diversity of new references that are transversal to different domains. We use two kinds of input data to propose our solution. The first one is a big base of articles, later referred to as the *corpus*, that the model has to learn and classify, guided by the second input which is a set of articles examples we are looking for, denoted. We start by describing the two data bases we built. Then, we explain our approach using these data sets.

In this experiment we are interested in the domain of "mental rotation" which is a good case of study because it rises interest in different disciplines such as education, social sciences, psychology or medical science. The data we use in our SSbE model is based on the *metadata* describing the articles, which are composed of the DOI, the title, the authors, the key-words when they exist and the abstract of the documents.

The documents composing the *scientific corpus* set come from ISTEX (see section 6.6.1) scientific digital library (SDL) whose aims is, first to gather the publications of different publishers of the last decades, second to offer an interface to access this large amount of research documents, and third to develop some useful statistical and research functions in order to exploit the available documents.

This SDL provides a large amount of scientific documents with different formats and language standards such as slides, posters, conference articles, and others. We chose to focus on the research papers written in English with complete metadata as required for our system. We also added some specific requirements since the research is only based on the abstracts of the papers: we chose articles containing an abstract with 35 to 500 words and being long enough without being a book (3 to 60 pages).

We limit our experimented corpus to articles published after 1990 extracting the following tuple of information: metadata and source (i.e., which database: *seed articles*, *scientific corpus*, or both). The distribution of the articles over time is non uniform due to the access conditions to publications and editors constraints negotiated by the SDL. In the end, we were able to benefit from a number of useful articles of 4,174,559 documents.

Out of many document types (e.g., slides, posters and conference articles) we only considered English research papers that were published after 1990 with sufficient abstract size (35 to 500 words). The extracted metadata dataset contains more than 4.17 millions articles. The distribution per year is illustrated in Figure 5.5 where we can observe two main drops in the number of articles. The first drop is in the year 2001 and 2002 and the second one is after the year 2010 until we have almost no articles in the year 2015. This observation must be taken into consideration for any analysis we would discuss in the results section. Therefore, we will also present the ration of the number of articles to the global number of articles in topic modelling analysis results that will be presented later in Section 5.6.4.

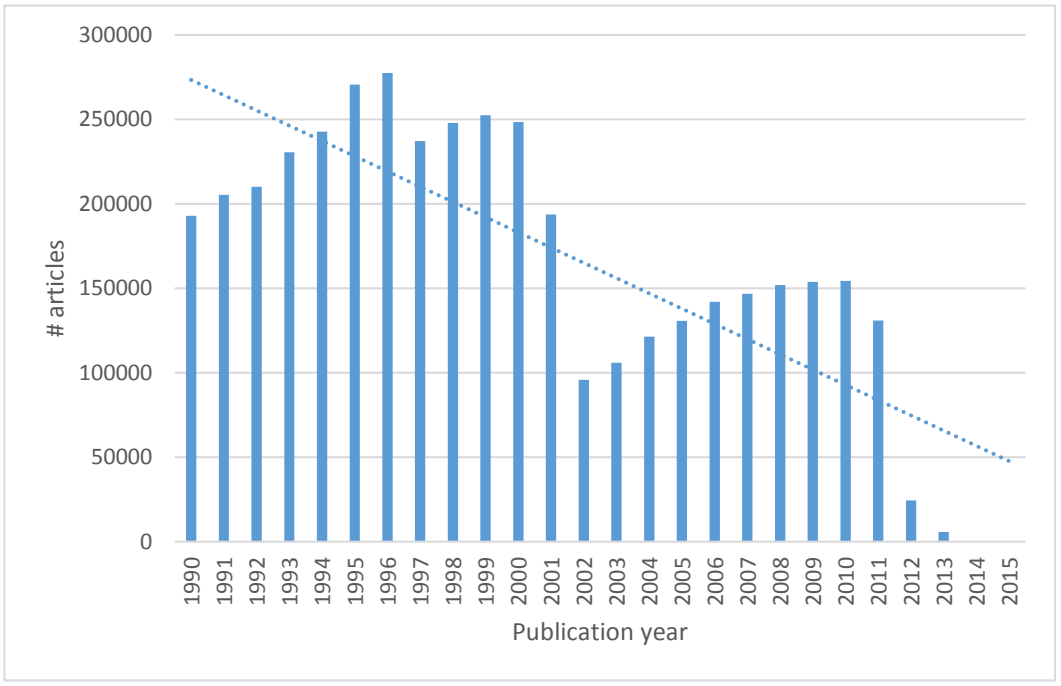


FIGURE 5.5: Distribution of the number of selected articles in English from the *corpus*, according to their publication date.

5.5.2.1 Data Preparation and Full-Text Versus Abstracts

As in any data mining task, data preparation is a crucial phase towards an accurate prediction model. Data loading and cleaning, text pre-processing, feature extraction and more are preparation steps that come with challenges in design and implementation. In many cases, we do not have a full access to all that data we expect to have. It might be missing due to a metadata multi sources integration issue as described earlier or due to an access limitation. Sometimes the cost of preparing full data is high. For example, it is mostly easy to extract a cleaned text content from the publication abstract being well structured in the metadata, that is, in many cases, an open access metadata. However, using the full-text of the publication is more difficult due to many constraints (e.g., when containing mathematical expressions, being only available in PDF format where automating full-text extracting resulted in non-cleaned data because of having the page footers and headers mixed with the content, scanned old publications format that require optical character recognition solutions, 2-columns format of the article with small space between the columns resulting in mixing lines, having tables and charts with inner text captions, *etc.*).

5.5.2.2 Construction of the Seed Article Set

The number of the seed articles of our use case experiments was 182 articles. They are all annotated by the focus domain experts as the focused topic: mental rotation. In this seed article set, 29 tagged articles do not even contain the *topic key-phrase* in their metadata. Only 25 documents tagged by the specialists are also part of *the scientific corpus*. For each article, we extracted the same metadata as in the SDL: DOI, authors, title, abstract, and keywords (when exist).

5.5.2.3 Expansion of the *Seed Articles* Set

We increase the number of the positive examples by extracting from the SDL database the research articles containing the expression "mental rotation" in the metadata. Thanks to this strategy, we extracted 199 additional documents out of the SDL and consider them as positive examples. We will denote these 199 additional articles as *extended positive articles*.

5.5.3 Model Experimental Design

Truncated Randomized SVD (Halko, Martinsson, and Tropp, 2011) and *Paragraph Vector* (Le and Mikolov, 2014) are two examples of vector semantics transformation techniques we considered in our experiments. The choice of these two methods among others was based on the availability of a scalable implementation in addition to the recent claimed efficiency. We first run comparative experiments of the two transformers based on Equation 5.2. Unexpectedly, the Paragraph Vector transformer did not result in any good vector representation using our experimented corpus. This could be due to the size and the speciality of such text corpus. However, the SVD transformer showed good results. Accordingly, we focused on finding a good design parameters of the SVD transformer. The parameter values we found the best among several experiments are listed in Table 5.1:

The result of the cosine difference of Equation 5.2 on these parameters was 0.31 as detailed in the following:

- Average cosine similarity within seed articles: 0.4;

TABLE 5.1: Best parameter settings found by our experimental design

Parameter	**Best value**
Minimum term frequency	20
Maximum term frequency percentage to keep	0.95
N-gram range (constrained by the memory size)	range (1,2)
Filtering out stop-words or not	yes
Using lemmatization or not	no
Dense semantic space dimension	150
Using TF–IDF transformer or not	yes

- Average cosine similarity within articles randomly selected from corpus articles other than the positive ones (same set size of mental rotation ones): 0.09.

As a binary classifier, we used the ensemble learning method that is random-forest classifier. This choice was based on the performance of such type of machine learning in many applications reported recently in many publications. Another important feature of this method is the ability to provide the probability score of class prediction which is needed for our method. The design parameters were set as the defaults of the classifier implementation of scikit-learn machine learning library (python2 version 0.18.1). In order to decide on the best number of estimators to use and to validate the accuracy of the classifier, we used cross validation and 30:70 test-training dataset splitting. Using 500 estimators for that classifier, the prediction accuracy was higher than 0.95. This accuracy value was the average of several runs with different randomly sampled negative examples. The number of runs were 100 so that we can somehow neutralize our assumption that the randomly selected samples from the scientific corpus are more likely to be negative examples. This assumption will be also handled in the active learning process as we will discuss later in this chapter.

After obtaining our trained classifier, we apply it to all the documents, more than 4 millions, predicting the probability for each document to be classified as a mental-rotation article. We used this probability value as a score value in which we ranked all the documents in a descending order. The top few thousands documents can then be evaluated and thus considered as a potential expanded scientific corpus of the topic "mental rotation."

In order to decide on the top-N ranked results, we can reuse our 199 *extended positive articles* so that we can compromise between the score value and the number of included *matched key-phrase articles* that we used as a test set. Our criteria of deciding on N

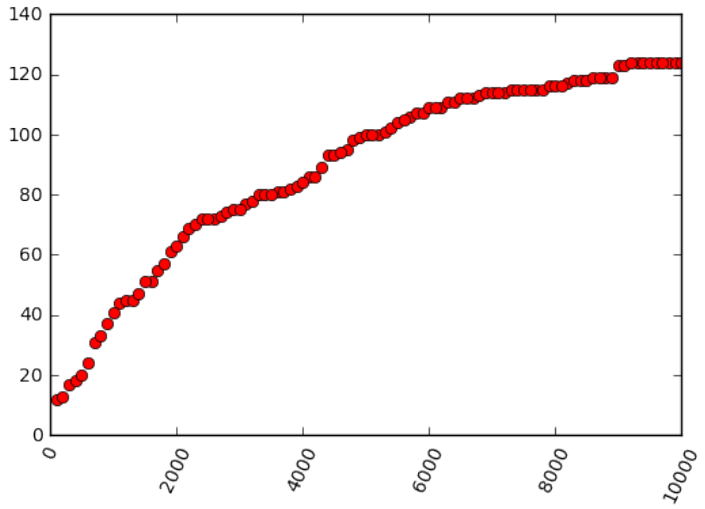


FIGURE 5.6: number of test set matched in the top 10K results

value was a compromise between the size of the corpus, the score, and the number of matched results with the *extended positive articles* articles, we will take the top 10K as our predicted expanded corpus. In top 10K results, we have 128 matched out of *extended positive articles*, 199 which is about 2 thirds of the *extended positive articles* articles, and the model minimum score of 0.71, i.e., the probability of being classified as "mental rotation" related article. We also think that a corpus size that is 1:400 portion of a multidisciplinary scientific corpus is somehow representative. In other words, if we have balanced distribution of articles over assumed 400 other scientific disciplines or topics, which we think is a fair number of scientific topics.

The cure of number of test set matched in the top 10 thousands results is shown in Figure 5.6

5.5.4 Computational Time of the Proposed Approach

Using a scientific corpus of 4 million, the most time consuming phase is to vectorize it. based on our experiment,it takes up to 22 hours on a 64 GB RAM 16 core CPU machine. However, this process is only required once and in an offline mode of the information retrieval system. It could be run periodically based on the number of the new articles coming to the corpus.

5.5.5 Active Learning

For the active learning process of the SSbE model, we generate a balanced dataset as follows:

- Negative examples that are composed of 2 sets:
  - the annotated results by the domain experts in which at least one of them marked it as irrelevant
  - In case the number of positive examples are higher than the annotated irrelevant articles, we randomly extract articles from the digital library corpus other than the positive ones in order to have a balanced dataset
- Positive examples that are composed of 3 sets:
  - Seed articles (182)
  - Extended positive articles (199)
  - The annotated results by the domain experts in which at least one of them marked it as relevant while the other could not decide

This new balanced dataset is then used to re-train the classifier we used in SSbE. We then use this newly trained classier to predict the probability of each article in the digital library corpus. Finally, we sort all the articles by the score value with descending order to form the new results of the model. The new results should not have any of the irrelevant-annotated articles in the top results. This would be verified in section 5.6.2

5.5.6 Sentence Semantic Relatedness Measure

We extracted the titles of the top 200 results for each of:

- More-Like-This method: MLT
- Partial SSbE model (without active learning): $SSbE_p$
- SSbE model with active learning: $SSbE$

We then generated 3 sets of pairs from the seed titles and the titles of each method. The size of each set was $200 \times 182 = 36,400$ pairs. In order to estimate the semantic relatedness score of each pair for each of the 3 sets, we used a pre-trained model[2] (Al-Natsheh et al., 2017b) which provides an estimation score between 0.0 to 5.0. This model was trained on an open access datasets[3]. We finally counted the pairs with the semantic relatedness score above a threshold $t = 3.0$ in order to compare the results of the 3 methods.

5.5.7 Diversity Analysis

Exploring relevant articles from different disciplines, by definition, should lead to a higher diversity and related topic overlapping in the expanded scientific topic corpus. Accordingly, we need to identify and define measures that quantify the rate of diversity and relevancy in order to compare the results obtained by different methods. This task is not that simple due to the large possible number of parameters even with simple aggregation like counting or averaging. In our case, we proposed to base the statistics on the words appearing in the title of the article, the author affiliations, the journal names, the keywords, or even a compilation of the keywords appearing in one or more of these fields. We should keep in mind that any derived diversity measure must maintain the results relevancy, e.g., such diversity indicators should still be relevant to the studied scientific topic expansion. We can assume that we achieve this purpose if we extract such keywords from the relevant articles in the results.

Our proposed diversity measure compares the distribution of the vocabulary extracted from the titles, author affiliation, and other interesting elements of the articles. We can base these analysis on all the top ranked articles of the two compared methods. Considering only relevant articles, according to the feedback annotation of the domain experts, could be risky in case the amount of relevant articles is not balanced between the two methods. To overcome this risk, we extract diversity indicator keywords from equivalent number of relevant articles of both methods. This means applying a kind of random sub-sampling from the method that has bigger set of results. So, we run the experiment a certain number of times and then we apply the Wilcoxon test (Wilcoxon, 1945) that counts the number of times that a method has a higher results than the other.

[2] https://github.com/natsheh/sensim
[3] http://ixa2.si.ehu.es/stswiki/index.php/STSbenchmark

5.5.8 Repeatability

The developed code for all conducted experiments of this article is available as open-access in a github repository[4]. We think that this shared code would be useful for the repeatability and further comparative research. The dataset generation script is also included in the repository.

5.6 Results and Discussion

In this section we will show and discuss the evaluation results of our proposed model SSbE with and without the active learning process in comparison with another method that is the More-Like-This query method (MLT).

5.6.1 Model Result Evaluation without Active Learning

In order to generate comparative results to partial $SSbE_p$ model (without active learning), we passed to the MLT query the seed articles and the 199 articles that we found "mental rotation" in its metadata, i.e., extended positive examples. Using the default parameters of this query in ElasticSearch resulted into low number of results. So, we looked into these parameters and tuned them according to the design parameters of our method in order to return a sufficient number of results for our evaluation experiment. The number of results we achieved was 391 articles.

The tuned MLT query parameters were:

- *max query terms* was set to 150 to reflect the vector size we have in our method;
- *min term freq.* was set to 20;
- *max doc freq.* was set to 0.95 × number of articles;
- and not providing a list of stop-words.

For a comparative evaluation, we took 100 articles from the top results of our SSbE method and another 100 articles from the top results of MLT method. The resulted 200

[4]https://github.com/ERICUdL/ISTEX_MentalRotation

TABLE 5.2: Confusion matrix of the two domain expert judgment of both of $SSbE_p$ (SSbE without active learning) and MLT method on 100 results randomly picked from the top 200. S corresponds to $SSbE_p$ and M corresponds to MLT. CND indicates that the expert Can Not Decide

	relevant		CND		irrelevant		**Total**	
Method	S	M	S	M	S	M	S	M
relevant	8	2	3	3	0	0	11	5
cannot decide	10	1	10	4	17	4	37	9
irrelevant	2	0	13	5	37	81	52	86
Total	20	3	26	12	54	85	100	100

articles were then shuffled and blindly handed to two experts, same who provided the initial corpus, to manually annotate each article if it is relevant or not to the focused topic, i.e., "mental rotation." Keeping in mind that none of these articles have "mental rotation" in their metadata, the experts needed to look carefully through the whole article content to give their annotation. Inexpert annotators would be inadequate as the task requires deep understanding of the research topic to decide whether an article from different discipline is relevant. Accessing to more 2 experts in such rare domain was not easy but we think it would be sufficient for a fair comparison. Given that the annotation efforts were big, we could barely reach our minimal target of 200 annotated articles. In addition to [relevant, irrelevant], we found a third case in which the domain expert find the recommended article related and useful being partially relevant. So, they cannot label it as relevant nor as irrelevant. After discussion with the experts, we decide to include such a case that would be denoted as *cannot decide*.

A confusion matrix for each method were then computed in order to check the agreement level between the two experts. These confusion matrices are shown in Table 5.2.

We also computed the Cohen's kappa coefficient (Cohen, 1960) that measures the agreement between the two domain experts with their annotation labels. As we can see in Table 5.3, the kappa score is very high for the labels [*relevant, irrelevant*]. It means that they were mostly agreed on the extreme judgment on the resulted articles. However, The two domain experts have less agreement when one of them use the label *cannot decide*.

To come up with a relevancy score for each article in the list of ranked results, we assign a numeric value for each expert annotation (i.e., 1.0 for relevant, 0.5 when experts can not decide, and 0.0 for irrelevant). The final score of each item is then the average of

TABLE 5.3: Cohen's kappa scores for annotation of the two domain experts. The table shows results for different combination of annotation labels. The scores are rounded to 4 decimals

Labels	**Cohen's kappa score**
[relevant, irrelevant]	0.9008
[relevant, cannot decide]	0.1751
[cannot decide, irrelevant]	0.2764
[relevant, irrelevant, cannot decide]	0.3797

TABLE 5.4: Frequencies of the evaluation scores values for both the $SSbE_p$ method and the MLT method. The blue score labels are good while the red score labels are bad

Score	1	0.75	0.5	0.25	0
$SSbE_p$	**8**	13	12	30	37
MLT	2	4	4	9	**81**

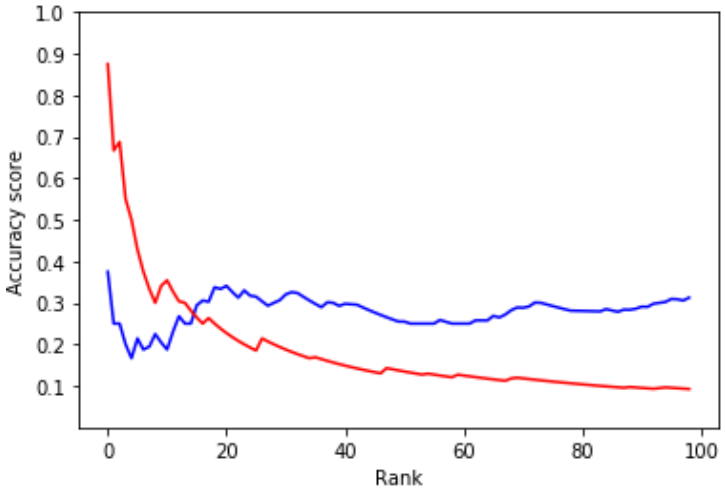


FIGURE 5.7: Accuracy curves of $SSbE_p$ method in blue and MLT method in red. Considering the top 100 $SSbE_p$ scored 0.3125 while MLT scored 0.09. At the very top results, MLT has better score but with very few total number of relevant results.

two expert scores. Thus, the possible score values we have for each result item are {0.0, 0.25, 0.5, 0.75, 1.0}. The corresponding score results are listed in Table 5.4.

Afterwards, we computed the accuracy of both methods at the top n, such that n is the number of results, based on Equation 5.3. Iterating over the ranked results from 1 to the number of annotated articles, we could see in Figure 5.7 the accuracy curves of our SSbE method and MLT method.

$$\frac{\sum_{rank=1}^{top_n} score_{rank}}{top_n} \tag{5.3}$$

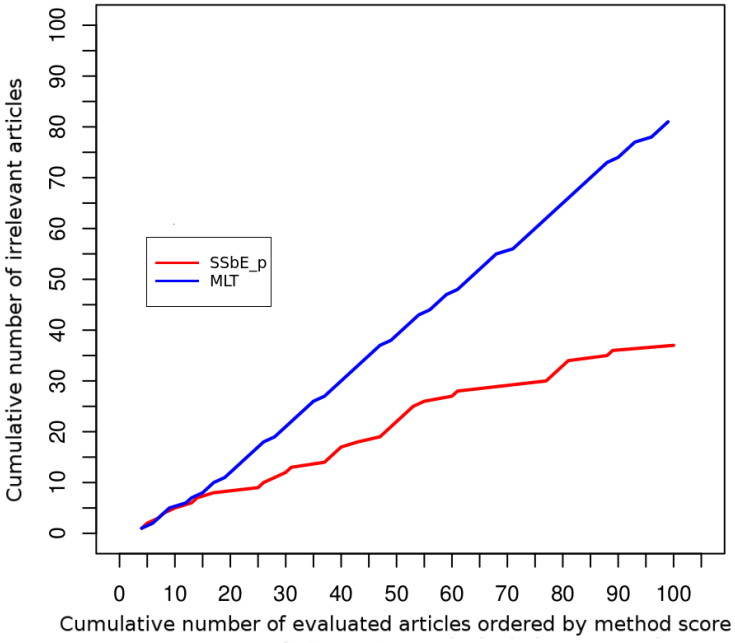


FIGURE 5.8: Number of irrelevant documents proposed to be in the field asked by the user (here, "mental rotation"), that have a lower rank than the value in abscissa using MLT or $SSbE_p$ method.

Another point of view for comparing the quality of the results of the two methods is to use very simple measures: Count of really irrelevant articles, i.e., scored 0, in the set of documents proposed to the reader by the classification method. We notice that our method obtain better results on a long ranked list of articles than the MLT method which is a little bit more efficient on the first results as we can see on figure 5.8.

In order to generate ROC curves, shown on Figure 5.9, we calculate the X- and Y-axis as follows:

- X : $\frac{\text{number of relevant documents at rank i}}{\text{number of relevant documents in top n}}$
- Y : $\frac{\text{number of irrelevant documents at rank i}}{\text{number of irrelevant documents in top n}}$
- where i is the rank from 1 to n.

We can notice that the area under the ROC curve (Flach, Hernández-Orallo, and Ramirez, 2011) for our $SSbE_p$ method is bigger than the one of MLT method.

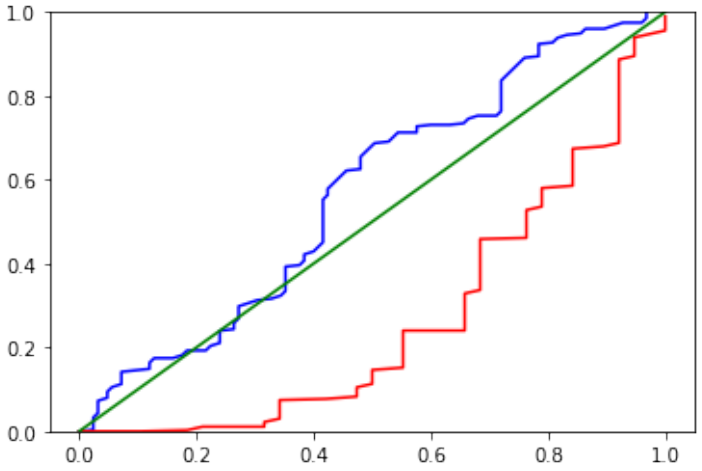


FIGURE 5.9: ROC curves for $SSbE_p$ method in blue and MLT method in red. The diagonal green line shows the goodness boundary where the curves should lay above.

TABLE 5.5: Comparative results of the 3 methods using sentence semantic relatedness measure based on count of pairs with score higher than 3.0 out of 5.0

Method	MLT	$SSbE_p$	$SSbE$
Count of pairs	124	217	382

5.6.2 Evaluation of the Model with Active Learning

The first results verification step we did was to make sure that the new results after applying the active learning process do not contain any of the irrelevant-user-annotated results. We verified that the top results of the new ranked list of articles does not contain any. The second step is then to find a way to evaluate the new results. For that, we will show two evaluation criteria: first, by using the sentence semantic relatedness measure on the article titles, and second, by using a test set generated from the metadata annotation of the digital library corpus that was hidden from our experiment.

5.6.2.1 Evaluation using sentence semantic relatedness

As described in section 5.5.6, we want to evaluate 3 models: MLT, $SSbE_p$ and $SSbE$ by pairing the titles of top 200 results of each method with the titles of the seed articles. Using the introduced evaluation measure in section 5.5.6 and a threshold semantic relatedness score value of 3.0, we obtained the results in Table 5.5.

We can see from Table 5.5 that the results of the introduced measure correlate with the evaluation results of the two domain experts showing that the $SSbE_p$ method is better

TABLE 5.6: Comparative results of the 3 methods on the top 959 results of each method using a test set extracted from the digital library metadata that was hidden from our experiment. The number of 959 results were selected as a result of excluding extended positive articles, which have been used in the training phase, from the top 1000 results of $SSbE_p$. *:The total number of the MLT results is 391 articles

Method	MLT^*	$SSbE_p$	$SSbE$
matches count	1	1	6
rank of them	1	851	$24, 82, 227, 567, 699, 929$

than the MLT method. Using the same measure, we can observe that we achieved a higher evaluation value of $SSbE$ than $SSbE_p$ thanks to the active learning process.

5.6.2.2 Evaluation using a test-set

In this measure, we checked how many matches and at which ranks we can find a test set of articles. These test set articles were hidden from the experiment and were extracted from the metadata of the digital library corpus. The query criteria we used to extract this test set was finding the phrase *"mental rotation"* in the list of *subjects* or *keywords* but not mentioned in the *abstract* nor the *title*. The results of this test set evaluation are shown in Table 5.6.

We can notice in Table 5.6 that we have 6 matches for $SSbE$ comparing to only 1 match for the other two methods. Looking to the rank of these matches, we observe that MLT method was better than $SSbE$ method using this type of evaluation. However, by using the domain experts annotation as shown in section 5.6.1, we could find much more relevant articles using $SSbE$. We can also see that 5 out of 6 ranks of matches for the $SSbE$ method were higher than the rank of the $SSbE_p$ method.

5.6.3 Results of Diversity Analysis

As introduced in section 5.5.7, we propose to observe the diversity of the documents, using indicators like journal names, departments of the authors, assigned topics or keywords. We may have a clue of the coherence of the results looking at the number of articles concerned by categories of subjects.

The initial idea was to simply observe the amount of vocabulary we can extract from titles, affiliations or scientific categories of the articles that is ranked in the top 200 results

TABLE 5.7: Amount of distinct vocabulary over the first 200 articles ranked by the three systems, based on categories : MLT, $SSbE_p$, $SSbE$

System	MLT	$SSbE_p$	SSbE
	76	57	57

of each method, without taking into account the relevancy of this documents. We first considered single-word tokens extraction from the titles of the articles, author affiliations and journals. This single-word tokens strategy produced a set of vocabulary that are noised with a lot of irrelevant vocabulary. We also face the problem of very generic words that can be applied in a lot of domains, especially with the MLT system. On the contrary, working on key-phrases such as domain categories given in the metadata of the articles seems to give very good and relevant results. Simply by counting the number of different phrases that we can find in the articles excluding the completely marked irrelevant ones produced the results summarized in Table 5.7.

The illustrated diversity analysis in Figure 5.10 presents the number of documents containing the vocabulary in abscissa for $SSbE_p$, compared to MLT system. We select only the vocabulary that appears enough time (20 at least) to be considered as relevant because well used in the domain. In addition of a slightly better diversity of the $SSbE_p$ method, shown in with more blue color than red, we also notice that the vocabulary of our method includes more related vocabulary of the target domain, i.e. mental rotation. Most of the MLT vocabulary we get is very generic and can be applied to various domains. For example, the three category names in the top of the figure that have more articles produced by MLT are actually irrelevant to the studied domain. On the other hand, the key-phrases extracted from the results of our systems are more precise such as "biological psychology" or "physiology" compared to "gerontology" or "psychiatry." So, we could conclude that the amount of diversity is not enough to judge; We should additionally take into account the relevancy of this diversity.

5.6.4 Topic Modelling Analysis

Before running an experiment, we need to better understand the input itself to interpret the results properly. One question we address is the organization in term of topics of the articles sets. We need to identify the topics present in the corpus and the sub-topics of the main domain of the seed articles set. To have a good overview of the organization

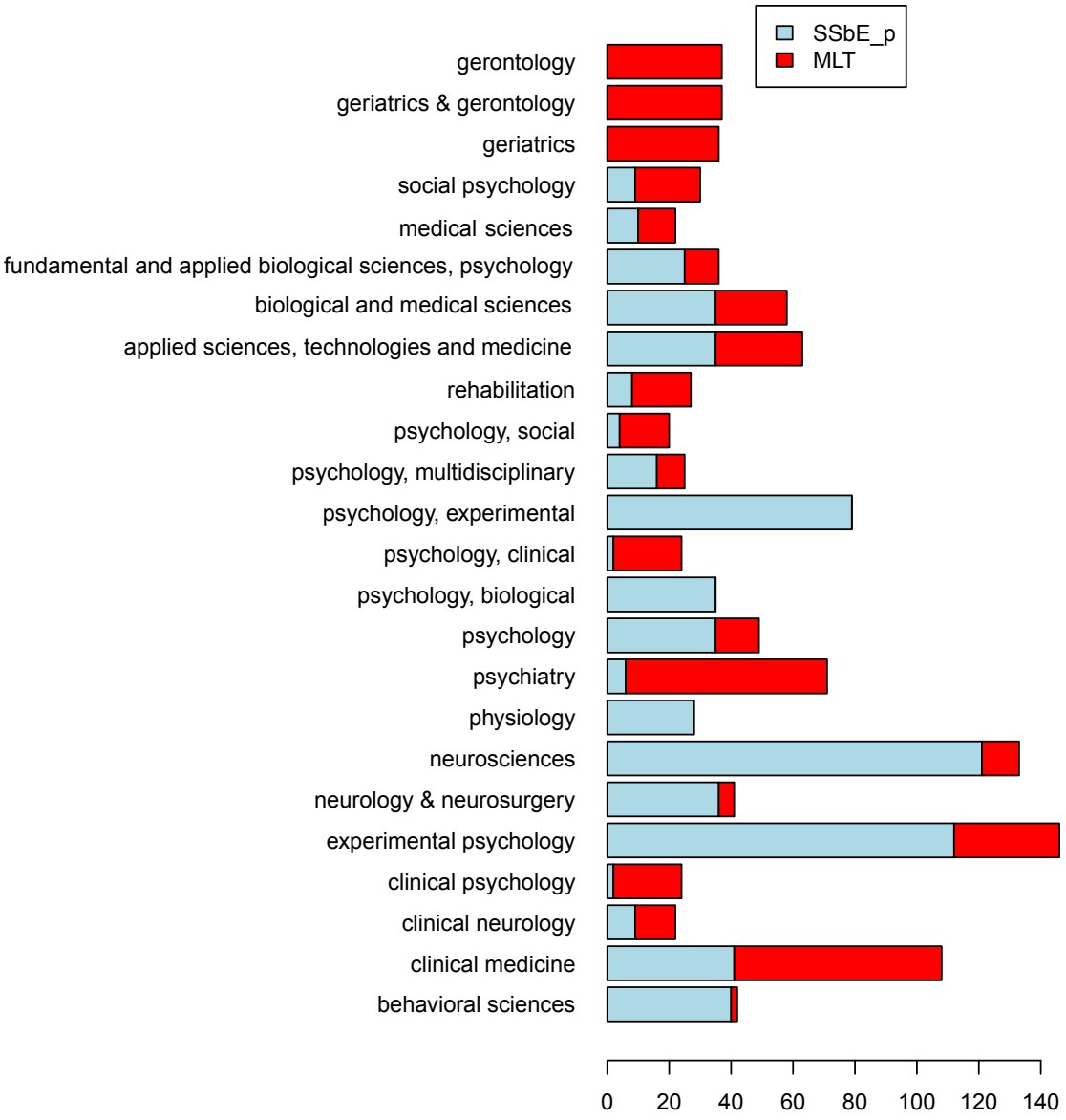


FIGURE 5.10: Distribution of the vocabulary of each documents over the global vocabulary based on categories discovered in the 200 best ranked documents for each the MLT and $SSbE_p$ systems. We show here only the vocabulary appearing more than 20 times globally for the two methods.

of our data, and especially identify the sub-topics of the mental rotation domain we use two complementary strategies described in the following sub-sections. Both of these strategies were based on the topic modelling algorithm called Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan, 2003) to retrieve the main topics. The input of the LDA model consists in a balanced set of positive (annotated as mental rotation) and negative (randomly selected) articles examples.

5.6.4.1 Sub-Topics on Seed Articles

In order to extract a list of key-phrases that can describe the mental rotation topic as well as possible subtopics, we applied a topic modeling approach that is able to define and categorize a given set of documents into a number of topics. In our experiments we generated 400 positive examples out of the seed articles as well as the articles we found containing the key-phrase *mental rotation* in their metadata. Another 400 negative examples uniformly randomly selected from the corpus, more than 4 million documents, compose the negative set. Since we do not know in advance the number of topics that the topic modeling will best partition this 800 articles dataset, we iterated from 2 to 50 topics. We know that at least one topic must be a mental-rotation related topic as we know that half of the dataset consists of such documents. Indeed, the results showed that there is at least 1 topic in which its top featuring tokens are mental rotation related topics. We had this a-priori knowledge of such key-phrases from the experts who firstly provided the list of 182 seed documents. Using 8 topics as an input of LDA model, we found 2 topics with sufficient amount of articles related to mental rotation according to the features:

- Topic 1 (Mental Rotation Methods): Mental rotation, motor, task, orientation, stimuli ..
- Topic 2 (Spatial Ability): Spatial ability, visual, mental rotation, performance, sex/age/profession differences (demographics differences)...

These results are aligned with the initial analysis of the experts who provided the 182 documents about mental rotation. This analysis provides us with extracted key-phrases for the main topic "mental rotation," as well as the 2 sub-topics, i.e., mental rotation tasks and spatial ability studies on demographics differences. An expanded ordered list of terms defines these 2 sub-topics are as follows:

Topic 1:

rotation | mental | mental rotation | task | subjects | motor | object | objects | right | tasks | different | visual | stimuli | body | time | left | systems | processing | orientation | stimulus | performance | activation | participants | experiments | perspective | presented | imagery | effects | rotation task | hand | results | spatial | cognitive | information | response | parietal | cortex | support | figures | imagine | brain | activity | rotated | used | growth | experiment | children | explanation | solution | group | processes | mirror | related | dimensional | orientations | present | recognition | studies | images | reaction | process | showed | study | memory | increased | number | shown | image | effect | form | complexity | based | central | condition | 180 | tissue | times | suggest | suggests | turn | simple | hands | transformation | performed | normal | greater | human | neural | differences | transformations | affect | shepard | functional | cortical | areas | test | imagined | article | research | making

Topic 2:

spatial | differences | sex | patients | high | ability | using | risk | mental | study | sex differences | performance | data | group | results | groups | analysis | model | test | rotation | gender | abilities | spatial ability | mental rotation | use | children | women | higher | males | students | low | men | cognitive | based | structure | used | control | effects | method | methods | reasoning | levels | females | present | tests | significant | related | paper | mean | 10 | observed | important | phase | period | 50 | treatment | non | development | showed | models | different | training | gene | cells | age | new | studied | production | studies | temperature | education | effect | time | theory | water | evidence | obtained | findings | human | cases | years | species | field | health | factors | range | acid | verbal | cell | experimental | provide | scores | learning | difference | numerical | 12 | changes | family | male | problems

One question that was interesting for the researchers in the mental rotation domain to be explored is the key-phrase usage variation over time in our experimented scientific text corpus, i.e.,, ISTEX. Accordingly, we have provided the following trending charts of key-phrases on the obtained expanded corpus of the sub-topic 'mental rotation method' in Figure 5.11 for the sub-topic 'spatial ability' in Figure 5.12.

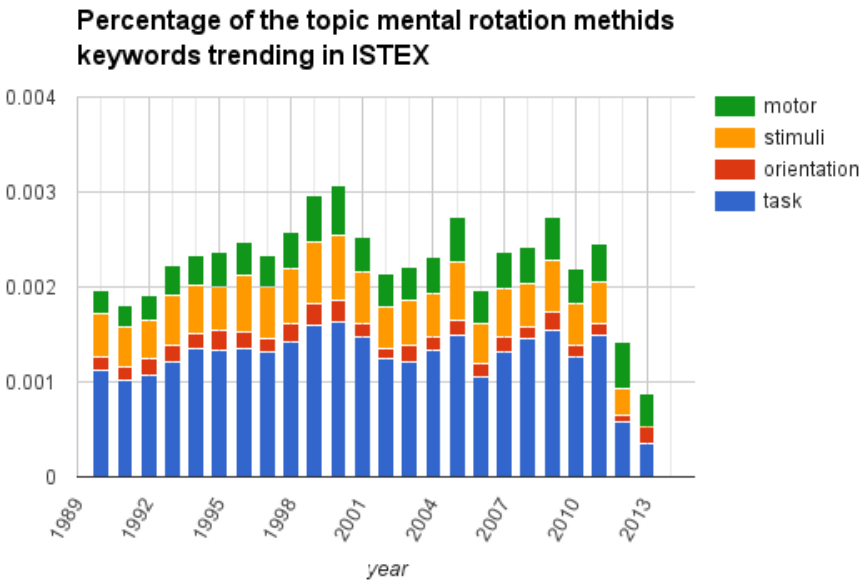


FIGURE 5.11: Emerging concepts in the expanded corpus articles.

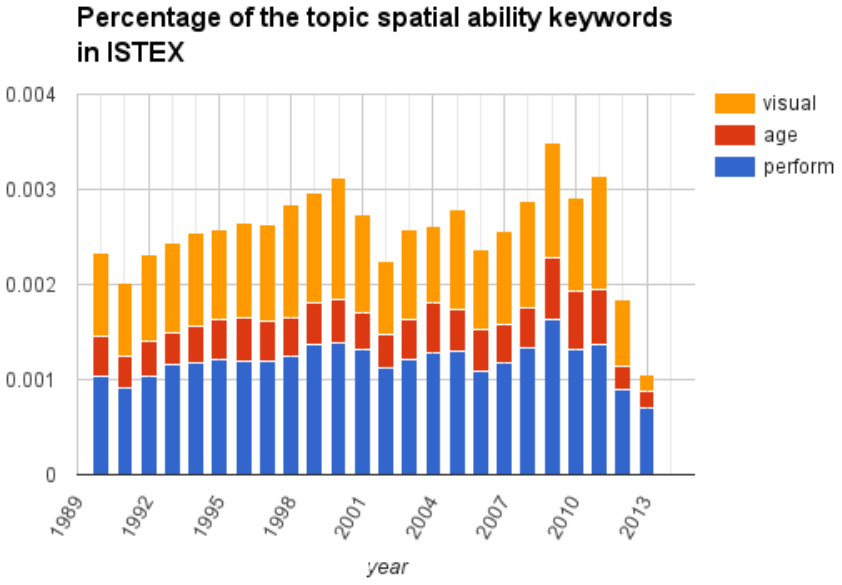


FIGURE 5.12: Ratio of the number of expanded corpus articles to the global number of ISTEX articles for the emerging concepts in the expanded corpus articles.

5.6.4.2 Emerging Topics on Results

We address the same question of the topics representation in the results and top list of documents extracted by $SSbE_p$, first, to have a way to evaluate the quality of our results and to underline the diversity of the fields that are concerned by the scientific aspects risen by the inter-disciplinary "mental rotation" research domain. We track the emerging key-phrases related to mental rotation by analyzing the top 10K results of the $SSbE_p$ model with LDA topic modeling. We could find the following main additional cognitive science related concepts that seems to overlap with mental rotation:

- Event related potentials (ERPs)
- Mismatch negativity (MMN)
- Attention deficit hyperactivity disorder (ADHD)
- Lingual gyrus
- Perirhinal cortex
- Transcranial magnetic stimulation (TMS)

Figures 5.13 and 5.15 shows the trending of the emerging concepts. Figure 5.13 shows the numbers of the seed articles, while Figure 5.15 shows it for the articles of the expanded corpus. In order to capture the trending line of these concepts, we plot the ratio of the number of expanded corpus articles to the number of ISTEX articles in Figure 5.14.

We can see from Figure 15 that only the concept 'Transactional Magnetic Stimulation' (TMS) was mentioned in the seed articles and it was only 3 times; once in the year 200, once in the year 2007 and also once in the year 2008. All other emerging concepts were not mentioned at all. This somehow shows how the mental rotation seeded sub-topics overlap with other scientific sub-domains we when we semantically expand the corpus.

5.6.5 Examples of some Surprising Articles

By identifying how accidental discovery processes occur, (Langley et al., 1987) resumed the words of Louis Pasteur who said "accidents favor the prepared mind," adding that "it is well known that attention is often attracted to phenomena that are familiar to

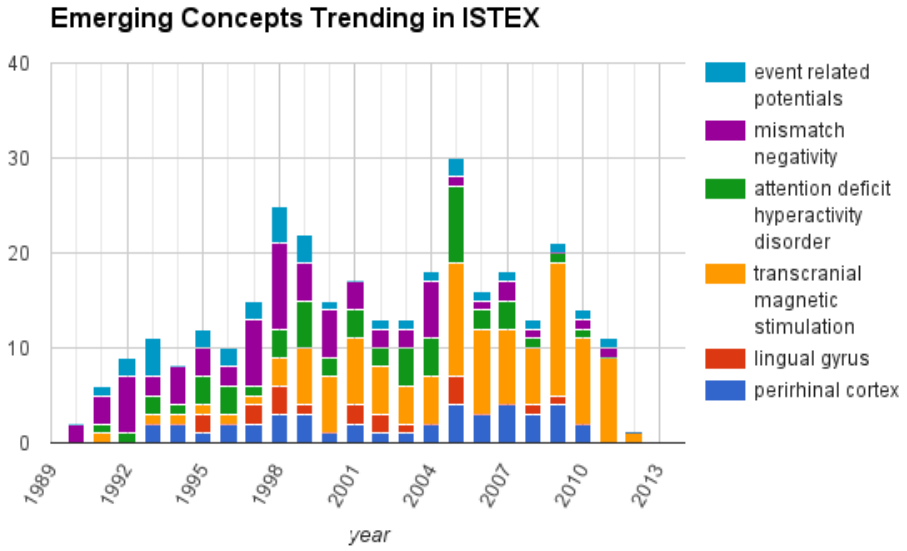


FIGURE 5.13: Emerging concepts in the expanded corpus articles.

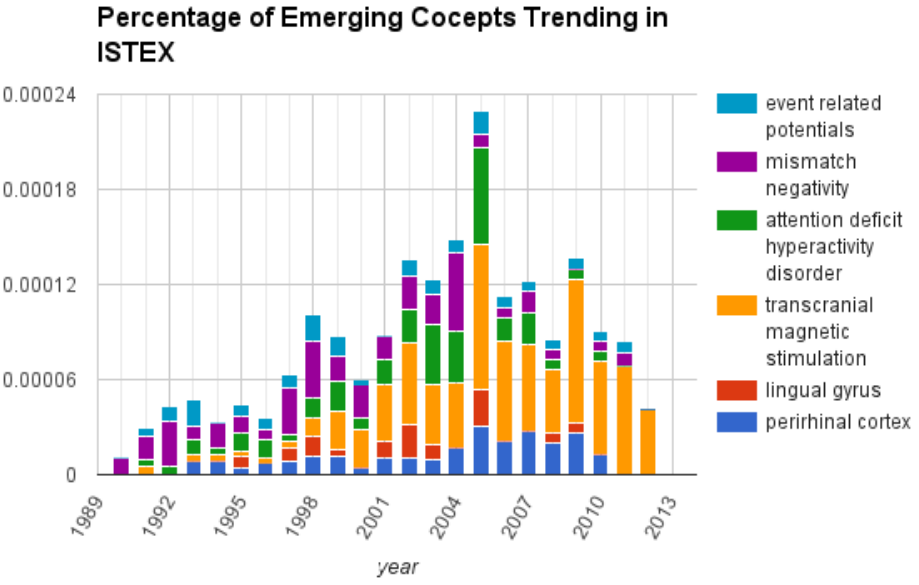


FIGURE 5.14: Ratio of the number of expanded corpus articles to the global number of ISTEX articles for the emerging concepts in the expanded corpus articles.

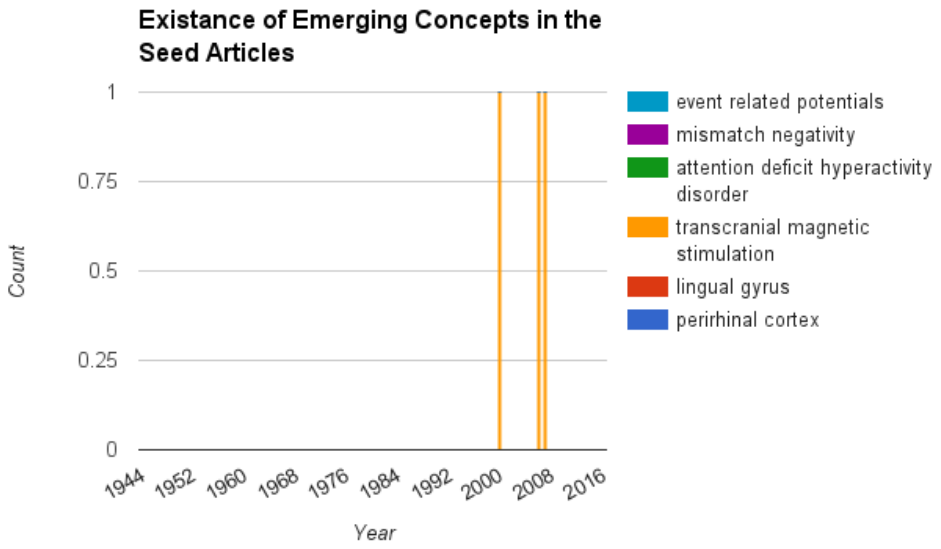


FIGURE 5.15: Emerging concepts in the seed articles.

the observer but that turn up in an unusual environment, or to new phenomena in a familiar environment, provided that the phenomena are relevant to the viewer's usual range of interests." We consider that our approach with *SSbE* model will favor such accidental discoveries by connecting scientific papers describing relevant similarities seen on a higher level than the topic targeted by a given discipline.

The articles found by $SSbE_p$ model and recommended to the researchers are not always considered to be relevant. However, since these proposed articles contain semantic similarities to those used as input (i.e., the initial corpus), the recommended papers share some topic connections with the input papers and open the research on new thematics. In our study, some recommended articles surprised the experts who evaluated these documents and gave them ideas for further research in new directions. For information purposes, the sports science experts came across an article which, without mentioning the mental rotation task, evoked a near theme concerning the studies on abilities to read a map in different orientations (Tlauka, 2006). This discovery has led the sports scientists working on mental rotation to see extensions of their work to the field of *orienteering*, a sport that requires navigational skills using map and compass to run in an unfamiliar terrain.

Another example of transdisciplinary discovery made by the mental rotation experts is the following: through a similarity of activation of brain areas, they find that there are some connections between the mental rotation and the sign language (Sadato et al.,

2005). Indeed, sign language and lip-reading used by deaf signers are actions that require some mental rotation abilities for reading the manual communication. Scientific bridges had not been made between such fields of study until now.

5.7 Conclusion

We proposed a novel model to expand a given set of scientific article examples into a corpus of semantically relevant articles of the scientific topic. Beyond keyword matching, these explored articles might belong to variant disciplines that tend to use different terminologies. We call this model Semantic Search-by-Examples *SSbE*. We conducted an experiment of our proposed model over *ISTEX*, a big digital library corpus, on a use-case of a multi-disciplinary scientific domain, i.e., *Mental Rotation*. The experiment showed the superiority of our model against an existing method, i.e., More-Like-This query which exists in a widely used open-source search engine for digital libraries. The comparative evaluation was possible by having a feedback annotation of two domain experts. We also showed the applicability and the importance of active learning process in the model pipeline.

Additionally in this chapter, we introduced a new semantic relatedness evaluation measure to avoid the need of human annotators for result evaluation. The measure we introduced is based on a pre-trained sentence semantic relatedness estimator. We finally presented a further result analysis of the topic extraction and topic diversity. Our proposed approach produced more diversity on a set of related topic categories rather than the compared method. The code used in our experiment in addition to the script for downloading the dataset is made available for other researchers for repeatability and further comparative studies in this open research problem. This model could be applied not only for scientific corpus expansion but also for enriching the metadata of the digital library in off-line fashion. Once the articles are annotated with this semantically related scientific topic, it would be much easier for researchers to query such articles using any semantic variation of the topic key-phrase.

As a future work, we would like to study the usability of the sentence semantic relatedness measure inside the model pipeline to boost-up articles with high semantically related sentences. We also want to examine a topic modeling approach on the top ranked results

trying to identify the clusters that are mostly relevant to the initial corpus. Finally, based on an enhanced semantic sentence relatedness model, we can also introduce a semantic sentence highlighter that will identify interesting part in the text of the recommended articles. This will make it easier for the user to provide her/his annotation to the system and thus to feed in the active learning process.

Chapter 6

Semantic Metadata Enrichment

6.1 Chapter Overview

In this chapter, we will talk about the semantic metadata enrichment using semantic based scientific categorization in digital libraries and the importance of that in trans-disciplinary research. In addition, we will mainly present our published approaches (Martinet et al., 2018; Al-Natsheh et al., 2018) comparing to other related work methods like supervised LDA as a multi-label classification problem.

6.2 Introduction

With the digital revolution and the presence of (open-access) digital libraries, the activity of the researchers has completely changed. First of all, researchers are moving less and less into real libraries to do their work. Moreover, because of the dematerialization of the physical objects that are the books or the scientific journals, these documents in digital form take up a lot less space, so it became possible to have a fabulous amount of scientific knowledge. As a result, it has become necessary to have an effective automatic filtering system in order to make access to this mass of information possible, that is why research-paper recommender-system are becoming increasingly popular (Beel et al., 2016). Unfortunately, although they use different strategies to search for articles that might be of interest to their users, these filtering tools do not work so well because they

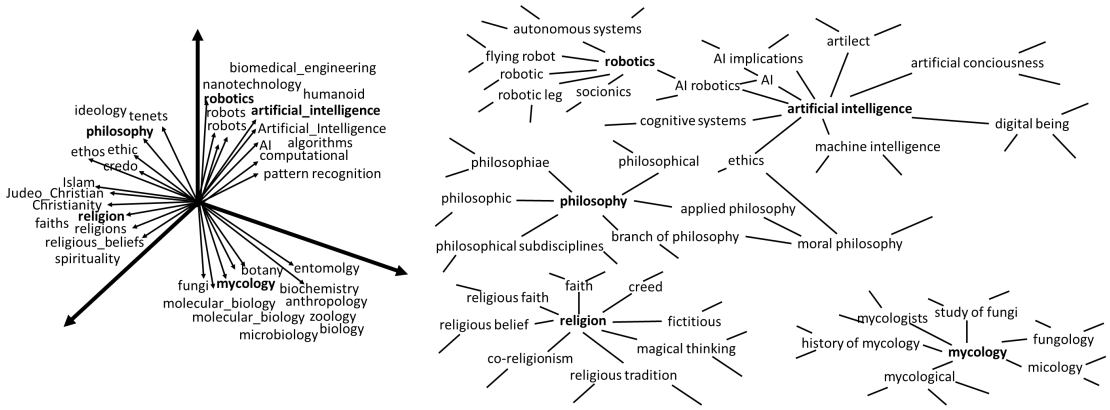


FIGURE 6.1: Representation of the hybrid semantic-based approach. The hybrid character of the method is associated to the combination of a semantic vector representation (left part of the picture) and a synonym set (right part), as shown for five example topics used in the experiments: **artificial intelligence**, **robotics**, **philosophy**, **religion**, and **mycology**. On the right, by querying the synonym set (e.g., obtained with *BabelNet* and *Elasticsearch*) using a text-based search engine, we generate a first ranked list of articles. On the left, the semantic vector representation (e.g., with *Word2vec*) is used in a semantic feature-based topic classifier phase to generate a second *Top N* article list with articles ranked by the probability of topic belonging. A per-topic fusion is made by combining the two ranked lists. Note that when more elements are added from the semantic vector representation, the associated words or concepts are less close from the initial keywords and bring more diversity and unexpectedness to the system.

are mainly focusing on accuracy of their recommendations, rarely on criteria of novelty or diversity (Castells, Hurley, and Vargas, 2015).

In the past, real libraries were places where happy accidents could occur. The attention is often attracted to phenomena that are familiar to the observer but that turn up in an unusual environment (Langley et al., 1987). For example while we search for a particular book, we can accidentally find another one with an appealing title. With digital libraries, unfortunately, such accidental or "serendipitous" discovery processes no longer appear. Another problem with digital libraries is the disciplinary compartmentalization. Even for the multi-disciplinary digital libraries, it is often difficult to get beyond the boundaries that scientific disciplines draw with each other. Indeed, a given term or expression may have different meanings in two different disciplines (e.g., "neural networks" in neurobiology and in artificial intelligence). But there is also the opposite case where a given concept is not expressed in the same way in two different scientific fields (e.g., what physicists refer to as "multivariate analysis" means "machine learning" for computer scientists).

With this unprecedented access to knowledge in digital form, we would like to obtain

from the interrogation of digital libraries, some associations between ideas from different disciplines that can produce a genius discovery. The filtering and recommendation steps then become key steps in how we access information and how we view the world. The way standard recommendation algorithms are used has important social consequences: they enclose individuals in the "bubble" of their own choices (Pariser, 2011). Recommending popular research papers or papers similar to those that have already been read will not help to cross the barriers of scientific disciplines.

In the following, we want to promote these transdisciplinary approaches through digital libraries covering all fields of knowledge of science, and even the arts and humanities. On the basis of semantic similarity we consider that we can propose diversity that does not exist in competing approaches, especially those based on keywords. We consider that the problems presented here with the scientific digital libraries can be related to a bad attribution of keywords to articles. Because of domain-specific jargon, keywords that make sense for a given scientific community may mean nothing to another one even if this topic is known by this other community but with other terms and keywords. Typically, when a researcher enters a query for finding interesting papers into the search engine of such a digital library, it is done with a few keywords. The match between the keywords entered and those used to describe the relevant scientific documents in these digital libraries may be limited if the terms used are not the same. Every researcher belongs to a community with whom she or he shares common knowledge and vocabulary. However, when the latter wishes to extend the bibliographic exploration beyond her/his community in order to gather information that leads him/her to new knowledge, it is necessary to remove several scientific and technical obstacles like the size of digital libraries, the heterogeneity of data and the complexity of natural language.

Another strategy is to make a manual enrichment of the digital libraries with metadata in order to facilitate the access to the semantic content of the documents. Such metadata can be other keywords, tags, topic names but there is a lack of a standard taxonomy and they are penalized by the subjectivity of the people involved in this manual annotation process (Abrizah et al., 2013).

In this paper we present a hybrid semantic-based approach for automatically tagging the papers of a multidisciplinary digital library by combining two different semantic information sources. The first information source is provided by the synonym set of a

semantic network and the second one from the semantic representation of a vectorial projection of the research articles of the scientific digital library.

6.3 Trans-disciplinary Research

In this paper, we focus on the "trans-disciplinary," a way of conceiving research that transcends divisions between disciplines. "Multidisciplinary," "interdisciplinary" and "transdisciplinary" are terms used to characterize different approaches in relation to the academic disciplines. The definitions given by dictionaries, Wikipedia or some authors are nevertheless quite confusing (e.g., in Oxford English Dictionary, for having the definition of the noun "trans-disciplinary", we need to see the definition of the adjective "transdisciplinary" where it is indicated "Relating to more than one branch of knowledge; *sic* interdisciplinary"). Often the definitions proposed by some authors are related to their disciplines of belonging, e.g., in medicine (Choi and Pak, 2006), in geography (Craciun, 2014), in economics (Max-Neef, 2005), in psychology (Stokols, 2006), or in metapsychology (Nicolescu, 2010). Based on the etymology, the Latin prefixes mean respectively "many" for *multi-*, "several" for *pluri-*, "between" for *inter-*, and "through" for *trans-*. Thus, the scholars of the past were *multidisciplinary* researchers in the sense that they were experts in different fields, not limited to a given discipline.

From Ancient history to the Renaissance until the Age of Enlightenment, as mentioned in the introductory chapter, most of the scholars were polymaths, experts in science and engineering as well as in arts and humanities, e.g., the Italian Leonardo da Vinci (1452–1519), the French Blaise Pascal (1623–1662), or the German Gottfried Wilhelm von Leibniz (1646–1716). *Interdisciplinarity* involves the combining of two or more academic disciplines into one new activity by crossing boundaries (e.g., bioinformatics is an interdisciplinary field that combines computer science, statistics, mathematics, and engineering to analyze and interpret biological data). *Transdisciplinarity* concerns the transfer of methods from one discipline to another through a unity of knowledge beyond discipline boundaries, for example Herbert A. Simon –mentioned in the Introduction too (Section 1.2.2)– was considered to be a very special transdisciplinary researcher (Simon, 1996).

In the following, we want to promote these trans-disciplinary approaches through digital libraries covering all fields of knowledge of science, and even the arts and humanities. On the basis of semantic similarity, we consider that we can propose diversity that does not exist in competing approaches, especially those based on keywords.

Beyond promoting diversity in research, we believe that our approach can also fix the problem of the rediscovery, years later, of work already carried out in another discipline, and the issues related to claimed novelty of a new introduced concept in a certain discipline. In some cases, scientists found after a while that such claimed novelty was actually proposed a few years ago but with different wordings and in a different discipline. Such a problem also raises the need for a citation recommendation system that considers the semantics of a text description of a claimed new work by extending the outreach to similar previous works from other disciplines.

6.4 State of the Art

For accessing efficiently to the knowledge of scientific digital libraries, the users –mostly the researchers– can use various tools made available to them by the owners of the digital libraries. One of the approaches of these tools is the computation of a relevance measure between the search queries and the research papers. Research-paper recommender systems are other kind of responses, with different strategies; e.g., the collaborative filtering approach when the assessments from researchers about research papers can be collected, or the content-based filtering approach when specific models of the researchers can be deduced from the research papers that researchers interacted with (Beel et al., 2016). However, if this explicit or implicit knowledge on the interest of users for a certain type of research article can not be collected for personalized results; a special attention should be paid to the keywords associated with the articles since the returns of the search queries are based on these keywords.

To overcome the limitations of keyword matching, semantic networks (Borgida and Sowa, 1991) are most often a good answer to the problems of linguistic variations in non-thematic digital libraries by finding synonyms or common lexical fields. The search query can be enriched by semantically relevant keywords extracted from lexical databases (e.g., *WordNet* (Miller, 1995)) or knowledge bases (e.g., *BabelNet* (Navigli and Ponzetto,

2012), *DBpedia* (Lehmann et al., 2015), or *YAGO* (Mahdisoltani, Biega, and Suchanek, 2015)). In the scientific field, this approach is however not sufficient because the technical terminology is unique to a particular subject, and this jargon has the particularity to evolve very quickly, which requires very frequent updates of the semantic networks.

The word embedding approach is another solution for finding semantically similar terminologies (Mikolov et al., 2013a; Bojanowski et al., 2017). Nevertheless, it is difficult, in this approach, to identify precisely the closeness of the terms in the projection and then if two terms have still close meanings. Moreover, word embedding techniques work well for finding similar concepts when they are described by single words but they work less well when the concepts are described by expressions of several terms.

Finally, generative statistical models like *latent Dirichlet allocation* (LDA) (Blei, Ng, and Jordan, 2003) are also interesting for finding association between documents sharing similar topics. For a scientific digital library, LDA is helpful by integrating the semantics of topic-specific entities (Pinto and Balke, 2015), even if it is difficult to implement an efficient solution on real digital library applications with millions of scientific papers.

When the set of terms is hierarchically organized, it composes a taxonomy. A *faceted* or *dynamic taxonomy* is a set of taxonomies, each one describing the domain of interest from a different point of view (Sacco and Tzitzikas, 2009). It has been shown that it is possible to expand an existing thesaurus using the abstracts of articles from state-of-the-art technological domains with limited structured information with word embedding techniques (Kawamura et al., 2016).

The use of *Latent Dirichlet Allocation* (LDA) (Blei, Ng, and Jordan, 2003) for assigning documents to topics is an interesting strategy in this problem and it has shown that it helps the search process in scientific digital libraries by integrating the semantics of topic-specific entities (Pinto and Balke, 2015). For prediction problems, the unsupervised approach of LDA has been adapted to a supervised one by adding an approximate maximum-likelihood procedure to the process (Blei and McAuliffe, 2007). Moreover, LDA technique has been declined in various ways for finding a solution to the original method drawbacks, by example for the document tagging problem: how is it possible to define a one-to-one correspondence between LDA's latent topics and user tags? An example of answer to this problem has been proposed with the *Labeled LDA* technique (Ramage et al., 2009). Semi-supervised LDA approaches are interesting solutions for

being able to discover new classes in unlabeled data in addition to assigning appropriate unlabeled data instances to existing categories. In particular, we can mention the use of weights of word distribution in *WWDLDA* (Zhou, Wei, and Qin, 2013), or an interval semi-supervised approach (Bodrunova et al., 2013).

However, in the case of a real application to millions of documents, such as a digital library with collections of scientific articles covering many disciplines, over a large number of years, even recent evolutionary approaches of LDA would require the use of computationally powerful systems, like the use of a computer cluster (Liang, Yang, and Bradley, 2015), which is a complex and costly solution.

6.5 Hybrid Model Description

Our model proposal is based on the hybrid combination of two different semantic-based approaches: a semantic vector representation and a synonym set (see Figure 6.1).

In more detail, as shown on Figure 6.2, the process starts by taking a corpus of millions of scientific articles from a digital library to extract the semantic features. The metadata of the research papers (most often available in open access, i.e., title and abstract) are considered to be the textual representation of each paper. The article textual representations is transformed into a sparse vector space with a TF–IDF weighted bag-of-word vectorization. The sparse vectorized representation is then semantically transformed into a dense vector representation of 100 to 600 vectors, e.g., with a SVD decomposition or a word-embedding approach like *Word2vec* (Mikolov et al., 2013a).

The second important step of the pipeline is the design of a topic classifier of the research papers of the digital library. With a search engine based on text (e.g., *Elasticsearch* of Apache Lucene), we can define positive and negative examples of a given topic name (i.e., a scientific category). Here, the negative examples are simply research papers randomly selected from the library without any match with the topic name in the metadata. Following the same vector representation process described in the previous step for having a semantic features, we build a "one-vs.-all" topic classifier to predict the probability value of belonging to the topic, e.g., with a random forest model which is intrinsically suited for multi-class problems.

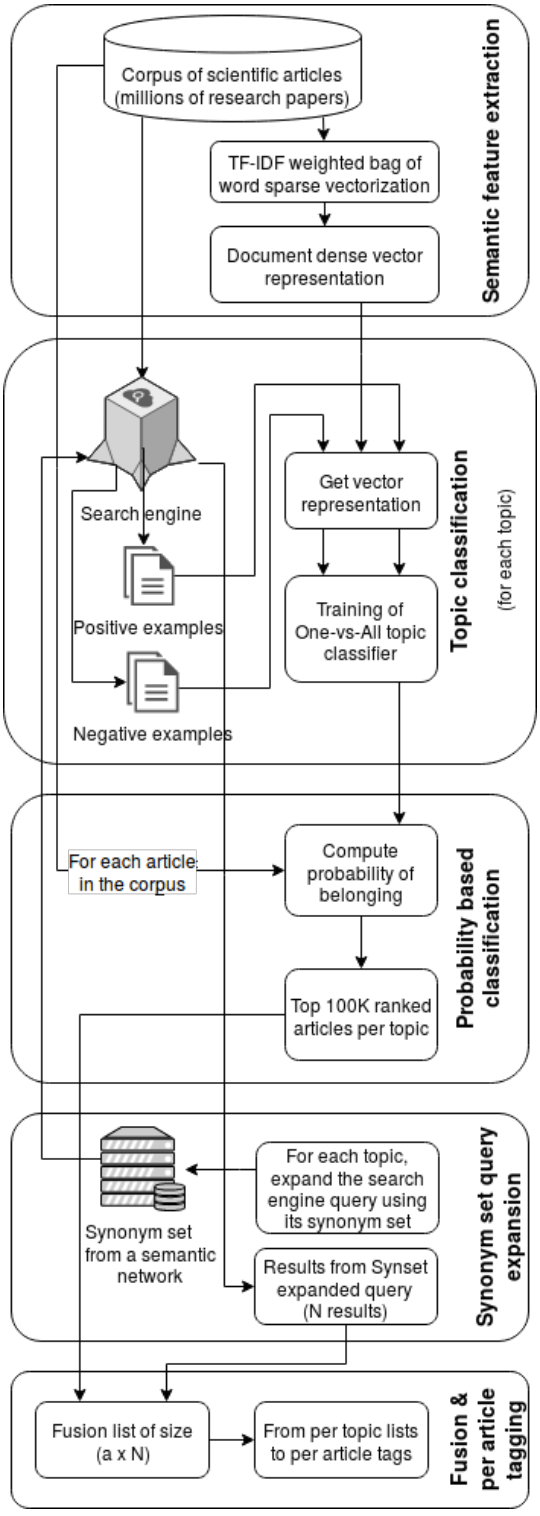


FIGURE 6.2: *Semantic Feature-based Topic Classifier*. After transforming all the articles into their semantic feature representation, a supervised learning classifier is trained on a balanced set of positive (initial corpus) and negative (randomly selected) article examples. The results are then ranked by the probability value that the trained binary classifier predicted each article in the digital library as the positive class.

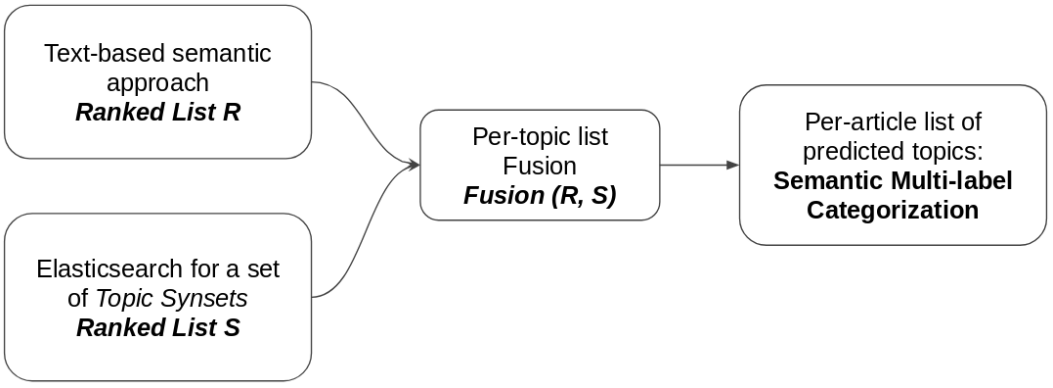


FIGURE 6.3: Fusion in Semantic-based Multi-labelling

On a third step, all the topic-model classifiers build on a specific topic are used on all research papers of the digital library. The result is a probability of belonging to a given topic for all the scientific articles of the library. By doing this, which is similar to a supervised topic modelling approach, we can obtain for each topic the top 100K ranked list of articles with the probability value as the ranking score.

The fourth step concerns the expansion of a synonym set query. A semantic network or a lexicon database, e.g., WordNet or BabelNet, is used to obtain a set of synonyms (or a "synset") of a giving concept name. A set of topic name synonyms will then be used for all the topics to expand the search query in a text-based search engine (e.g., *Elasticsearch*) in order to generate a ranked list of articles that have matches in their metadata with any of the synset of the topic. The ranked list of articles per topic can also be considered as a multi-label classification output, just like the result obtained in the previous step.

On the fifth and final step, a fusion is made with the two ranked article lists for having a hybrid character (check Figure 6.3). By denoting N the ranked list of research articles obtained by the synonym set approach (step four) and a the parameter used to select a given number of times the initial size of the N ranked research articles obtained by the list provided by the semantic vector representation approach (step three) merged with the list provided by the synset approach, we can obtain for each topic: N articles (obtained by having in the fusion list as much as the number of articles obtained with the synonym set approach only, i.e., $a = 1$); $2 \times N$ articles (obtained by having in the fusion list as many articles as the number of articles obtained with the synonym set approach alone, i.e., $a = 2$); $3 \times N$ articles ($a = 3$); $4 \times N$ articles ($a = 4$); and so on. For merging

the two lists, we introduce a new ranked list fusion criteria by taking into account, for a given research paper A, s_A which is the rank of an article A in the synonym set list, and r_A which is the rank of an article A in the semantic vector representation list with the following condition:

- if an article is present both in the two lists, the rank t_A is given by $t_A = \frac{s_A + r_A}{2}$;
- if an article is only present in the semantic vector representation list, the rank t_A of the article is given by $t_A = r_A \times |S|$ where $|S|$ is the size of the list of articles that is retrieved using the synonym set query expansion.

The fifth step ends with a transformation stage from the list of articles per topic to a list of keywords or tags by article. The fusion list, obtained by ranking the two lists with the score t_A, is a set of articles associated to a specific topic. For all the topics, we can use these fusion lists by applying a list inversion process that generates a per article list of topics for all articles presented in any of the fusion lists.

6.6 Methodology and Experiments

6.6.1 Datasets

Our methodology requires the use of the following datasets:

1. A corpus of millions of multidisciplinary research papers stored in a scientific digital library. In our experiments, the digital library used is *ISTEX* (EXcellence Initiative of Scientific and Technical Information, a French open-access metadata scientific digital library (Scientific and Technical Information Department – CNRS, 2016)) with 21 million documents from 21 scientific literature corpora in all disciplines, more than 9 thousands journals and 300 thousands ebooks published between the years 1473 and 2015 (in June 2018). Note that the titles, names of the authors, full references of the publications and other metadata can be accessed by anybody from ISTEX platform but the global access to the documents in full text is limited to the French universities or public research centres. The subpart of ISTEX corpus used in the experiments are research papers (from journals or conference proceedings) published during the last twenty years, written in English;

2. A set of topics covering all fields of scientific research. To have a maximum diversity of research themes, we do not want to restrict ourselves to the limited vocabulary of the computer science or even the science and technology domains, that is why we did not select too specialized taxonomies, e.g., IEEE Taxonomy Version 1.0 (2017) or ACM Computing Classification System (2012). The list of tags used in our experiments is extracted from *Web of Science*[1] collection which contains more than 250 topics (e.g., [computer science, artificial intelligence] or [computer science, network]). The selected 33 topics are listed in Table 6.1;

3. some synonym sets (synsets) extracted from a semantic network. The semantic enrichment of the 33 topics defined previously is made by using a list of synonyms for each topic with a semantic network. In our experiments, we chose *BalbelNet* (Navigli and Ponzetto, 2012) after testing on several semantic networks because it was the one which gave better results. Examples of synsets extracted from *BabelNet* are given on the right part of the Figure 6.1;

4. a set of texts to train the semantic vector representation. For having a semantic space of word dense vector representation with the word embedding technique (Mikolov et al., 2013b), we use a pre-trained word embedding model called *GoogleNews-vectors-300negative*[2] in our experiments. From a hundred billion terms from a Google News dataset, with a vocabulary of 3 million words and phrases, it is possible to extract a semantic vector representation with a vector size of 300. Examples of close terms in the semantic vector representation space obtained with *Word2vec* on the *GoogleNews* dataset are given on the left part of the Figure 6.1.

6.6.2 Evaluation Criteria and Accuracy Measuring

For testing the accuracy properties of the proposed model, we defined a parameter based on the count of tagged articles with a list of prediction topics that has at least one label intersection with ground truth. We call this measure "at least one common label" metric.

The other evaluation criteria are classic measures of the information retrieval or classification literature. Let P denotes the prediction label set, T the ground truth set for each article, L the list of all labels, N is the number of samples, $p_{i,j}$ the predicted list

[1] `https://images.webofknowledge.com/images/help/WOS/hp_subject_category_terms_tasca.html`
[2] `https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM`

of labels, and $t_{i,j}$ the ground truth set of labels. Then, the statistical and multi-label classification evaluation metrics used are:

$$\text{Jaccard index} = \frac{|P \cap T|}{|P \cup T|}$$

$$\text{Precision} = \frac{|P \cap T|}{|P|}$$

$$\text{Recall} = \frac{|P \cap T|}{|T|}$$

$$\text{F1-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{Label cardinality} = \frac{1}{N} \sum_{i=1}^{N} |L_i|$$

$$\text{Hamming loss} = \frac{1}{|N| \, . \, |L|} \sum_{i=1}^{|N|} \sum_{j=1}^{|L|} xor(p_{i,j}, t_{i,j})$$

It is important to note that the smaller the Hamming loss value is the better.

6.6.3 Experimental Process

The first experiment we conducted was to compare the results *Per-topic Fusion List* versus both the *Semantic Feature-based Topic Classifier* and the *Synset Expanded Query* method. This experiment will justify the usage of the fusion step in our experimentally designed pipeline shown in Figure 6.2.

In order to build an experiment of our proposed pipeline, we need to experimentally determine some of its hyper-parameters as follows:

6.6.3.1 Semantic Feature-base Topic Classifier

We limit our text representation of the article to its title and abstract, which are available metadata. The TF–IDF weighted sparse bag of word vectorization was applied on a word n-gram range of (1, 2). Comparing *Paragraph vector* and *Randomized truncated SVD* (Halko, Martinsson, and Tropp, 2011) based on a metric that maximizes the inner cosine similarity of articles from the same topics and minimizes it for a randomly selected articles, we choose SVD decomposition of the TF–IDF weighted bag of words in 150 features for more than 4 millions articles. As for the topic classifier, also by comparative evaluation, we select *Random Forest Classifier*, tuning certain design parameters, and use it to rank the scientific corpus. We consider the top 100K articles of each topic classifier to be used in the fusion step.

6.6.3.2 Synset Expanded Query

Reviewing many available semantic networks, we found that BabelNet was the most comprehensive one combining many other networks (Navigli and Ponzetto, 2012). So, we use it to extract a set of synonyms, i.e., a *synset* for each topic. This synset is then used to query the search engine of ISTEX which is built on Elasticsearch server. This technique will be used as the experiment baseline.

6.6.3.3 Fusion and per Multi-Label Categorization

The main design parameter of this phase is the size of the ranked list that is achieved by setting it to the double size of the *Synset Expanded Query* list.

6.7 Results and Discussion

6.7.1 Accuracy Results

As introduced in Section 6.6.3, the comparative evaluation results of the *Per-topic Fusion List* versus both the *Semantic Feature-based Topic Classifier* and the *Synset Expanded Query* methods obtained for all the experimented topics are listed in Table 6.1. This

table presents the recall obtained for each topic, according to the model we use. The values in bold underline the best results obtained over the three tested methods. Most of the recalls are obviously higher when using the hybrid method we call *Fusion*, that confirms it is an accurate method we can use further.

TABLE 6.1: Recall of the "Per-topic Fusion List" method versus both the "Semantic Feature-based Topic Classifier" and the "Synset Expanded Query" methods

Synset	**Topic Classifier**	**Fusion**	**Topic**
6.54%	12.18%	**19.18%**	Artificial Intelligence
14.16%	8.58%	**28.33%**	Substance Abuse
22.70%	5.41%	**24.98%**	Information Systems
14.37%	5.39%	**18.12%**	Thermodynamics
0.00%	5.69%	**10.35%**	Rehabilitation
7.16%	3.69%	5.83%	Psychology
16.25%	13.29%	**20.68%**	Philosophy
5.45%	3.05%	**5.45%**	Ophthalmology
3.71%	6.39%	**11.17%**	Microscopy
0.00%	7.92%	**12.46%**	Ceramics
3.64%	0.22%	**9.89%**	Infectious Diseases
9.41%	7.93%	**12.69%**	Toxicology
9.52%	2.12%	**16.40%**	Respiratory System
9.96%	**32.95%**	28.54%	Neuroimaging
12.44%	5.12%	**12.44%**	Literature
7.59%	4.44%	7.31%	Sociology
32.40%	14.46%	**35.48%**	Robotics
14.71%	5.53%	**18.76%**	Psychiatry
29.85%	7.10%	22.76%	Pediatrics
3.64%	**10.62%**	9.74%	Oncology
0.02%	4.40%	**5.63%**	Mechanics
5.88%	**8.98%**	8.05%	Biophysics
0.07%	**5.81%**	1.19%	Condensed Matter
4.91%	4.21%	**7.72%**	Emergency Medicine
18.21%	14.07%	**35.98%**	Transplantation
8.81%	10.73%	**15.47%**	Surgery
16.70%	18.06%	**19.93%**	Religion
4.45%	0.11%	3.40%	Physiology
6.38%	2.82%	**7.74%**	Pathology
0.57%	**1.89%**	0.57%	Mycology
4.26%	2.85%	**8.45%**	Immunology
9.02%	12.65%	**16.37%**	Biomaterials
8.41%	28.31%	**37.45%**	Nursing
9.75%	8.81%	**15.82%**	**Mean**
6 of 33	5 of 33	**24 of 33**	**Better count**

Looking for a good baseline to evaluate the accuracy of our proposed pipeline (*Fusion*), we

first tried to compare it against a comparative topic modelling approach which is a version of supervised LDA (sLDA) using a priori knowledge of predefined number of topics, i.e., 33 topics, where the synonym set of each topic will be boosted in the documents. Accordingly, we will have a version of multi-label document classification model. We took care of finding parameters giving the best results for sLDA (a boosting of 30) that we could find for our purpose. The results we obtained for sLDA, using four different measures (*F1-measure*, *At-least-one-common-label*, *Jaccard index*, *Hamming loss*), were absurdly low, compared to the ones obtained for our model *Fusion*, with $a = 2$. For example, the F1-measure of sLDA was less that 0.03 comparing to the F1-measure of Fusion2 exceeded 0.6.

After dropping *sLDA* from further experiments due to the very low evaluation results, we have added 2 more topics to the set of the 33 topics totaling to 35 topics. The 2 additional topics were [International Relations; Biodiversity Conservation]. We have also added more examples to the test set counting for an additional ISTEX metadata field called *categories:wos* that actually does not exist in all the articles but was still considered as a good source for increasing the test examples in our published benchmark.

Accordingly, the chosen baseline is the method of synonym set expanded query that will be denoted as *Synset*. We will compare it against four versions of our hybrid approach with four different values of the design parameter $a = 1, 2, 3, 4$, resulting in these corresponding method names *Fusion1*, *Fusion2*, *Fusion3* and *Fusion4*. Table 6.2 summarizes the results of the multi-label classification evaluation metrics, described in section 6.6.3.

Based on the metrics presented in Table 6.2, we recommend to use parameter $a = 2$ since it gives the best evaluation results across all the metrics. The precision metrics tracks the common categories assigned to articles by our system with the ones assigned in the test set. A low value has two possible meanings: the predicted tags might be either wrong, mismatching the test set or it could provide additional discovered tags that are actually semantically relevant (which is actually something we would like to achieve). The latter case would be the aimed added knowledge to the metadata of the digital library where the users can find relevant articles from different topic category names. Such category names would otherwise not be retrieved without applying the proposed tagging approach.

TABLE 6.2: Evaluation results based on the evaluation metrics: label cardinality difference from the test set, Hamming loss, Jaccard index, precision and F1-measure. Best values are formated as bold. Precision is equivalent to Jaccard index in this case of label cardinality of the test set = 1.

Method	Label cardinality difference	Hamming loss×10	Jaccard index	Precision	F1 Measure
Synset	**0.0741**	0.2906	0.5011	0.5011	0.5101
Fusion1	0.0833	0.2674	0.5431	0.5431	0.5529
Fusion2	0.1572	**0.2521**	**0.5825**	**0.5825**	**0.5998**
Fusion3	0.2223	0.2652	0.5720	0.5720	0.5954
Fusion4	0.2858	0.2833	0.5546	0.5546	0.5825

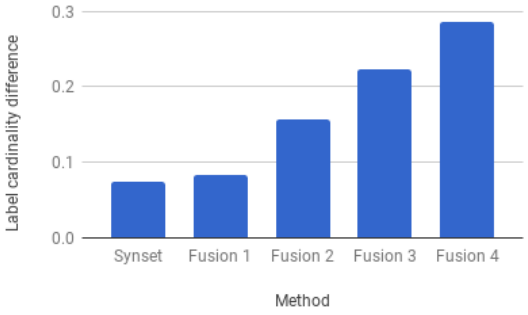


FIGURE 6.4: Label cardinality difference with the label cardinality of the compared test set of each of the methods.

Due to the fact that the test set was not generated manually but by filtering on a set of scientific category terms in relevant metadata fields, we believe that it is an incomplete ground truth. However, we think it is very suitable to compare models as a guidance for designing an efficient one because the test labels are correct even incomplete. Accordingly, we tried to perform some error analysis where we found that in most cases, the extra suggested category names are either actual correct topic (the article is a multi-disciplinary one) or they are topics from very similar and related ones.

Let us remember that the test set was generated by filtering a set of scientific category terms present in relevant metadata fields. This can lead to an incomplete ground truth. We estimate, however, that this kind of set can be very suitable to compare models, and can be used as a guide to design new models since the labels, even if incomplete, are true. From this set, we performed error analysis which lead us to evaluate the accuracy of the extra suggested category proposed by our model. In most cases, this extra categories are actually relevant for multi-disciplinary topics or belong to very similar and related topics. Actually, in many cases, our system retrieves at least one tag provided by ISTEX and suggests some others that turn out to be the aimed discovered knowledge rather than false prediction.

For example, a medical article from ISTEX[3] is tagged with the category name ['Transplantation'] in the test set. The predicted topics by our method was ['Mycology', 'Transplantation'] resulting into 0.5 precision value. However, when we read the abstract of that article, we find that it talks about *dematiaceous fungi* which is actually a *Mycology* topic. As a short example of this kind of results, we can take the case of the medical article whose title is: "*Sexuality of people living with a mental illness: A collaborative challenge for mental health nurses*," from Quinn and Browne in 2009[4]. This article is tagged with ["Nursing"] category name in ISTEX while our system assign to it two category names: ["Nursing," "Psychiatry"], which seems to be even more coherent than the set provided by ISTEX, since "Mental Health" is very related to "Psychiatry." Another article by Weiss and Bynoe, published in 2001, called "*Injection of tissue plasminogen activator into a branch retinal vein in eyes with central retinal vein occlusion*"[5], is presented with the category name ["Ophtalmology"] in ISTEX while our system assigns to it ["Surgery," "Ophtalmology"]. Once again, the text of the title indicates that the paper is related to the eyes surgery, as our model proposes.

So, in many cases where there is at least one common tag, the other tags are actually the aimed discovered knowledge rather than a false prediction. The complete list of results –where these cases could be verified– are published as well as all the experimental data and code for the reproduction of the experimental results [6].

In the end, the precision is a good indicator to quantify the accuracy of the results. However, we need to additionally consider other criteria, e.g., diversity and unexpectedness, to judge the overall quality of the results. Indeed, in addition to retrieving relevant articles, the model should also discover knowledge (i.e., predicting unexpected tags) that we can enrich the metadata of the article with.

6.8 Conclusion

In this paper, we proposed an efficient and practical pipeline that solves the challenge of the community-dependent tags or keywords and the issue caused by aggregating articles

[3] `https://api.istex.fr/document/23A2BC6E23BE8DE9971290A5E869F1FA4A5E49E4`
[4] `https://api.istex.fr/document/1CAAD07F9C1402C04CA28C96CCCB12CA45F6873B`
[5] `https://api.istex.fr/document/BB3D8F1D6F402FA93B869619B656439D6BFB58B6`
[6] `https://github.com/ERICUdL/stst`

from heterogeneous scientific topic ontologies and category names used by different publishers. We demonstrated that combining two main semantic information sources –the semantic networks and the semantic features of the text of the article metadata– was a successful approach for semantic-based multi-label categorization. This study aims to facilitate trans-disciplinary research by a semantic-based metadata enrichment with relevant scientific topic categorization tags.

Looking into the proposed pipeline as a paper tagging recommender system, recommending "very similar" tags (i.e., topics) would not bring much added knowledge to the user. That is why measures like unexpectedness and diversity play a key role in defining the quality of the recommender system beside the accuracy and the high relevancy of the recommendations. We applied these two measures to our proposed approach where we found that the best performing designed model provided a good trade-off between these three factors, i.e., accuracy, unexpectedness, and diversity. We have chosen a challenging test set by design, however, even for low precision evaluation values of the article tagging, we actually found that the additional predicted tags were in many cases relevant and provide the aimed knowledge discovery of the pipeline that will also provide sufficient diverse and unexpected tags serving the purpose of trans-disciplinary research.

In order to go further, we are planning to study the impact of using extra information from *BabelNet* semantic network other than only the synonym set. In particular, we want to include the neighbouring concept names as well as the category names of the concept. We expect that such term semantic expansion will improve the performance of the method.

In terms of future work, we aim to improve the pipeline by enhancing its main components. For instance, we are planning to enhance the process of generating synonyms from *BabelNet*. This includes, for example, taking into account the category name of the synonyms using a concept disambiguation technique that is based on common keywords among articles (Latard et al., 2017). We have recently initiated a collaboration with the authors of that technique in order to obtain a better system that could be implemented in the open-access digital library they are working on. We are also planning to measure the correlation among the co-occurred predicted topic tags. Detecting such correlation would lead to semantic-based linking between articles and thus be used to develop research-paper recommender systems.

Chapter 7

From Meaningless Words to Corpus Semantics

7.1 Chapter Overview

In this chapter, we will discuss a multi-level overview of dealing with a text mining problem starting from meaningless words, i.e., people names, to a semantic level of exploration in a large text corpus. We will also talk about the text granularity level from words to documents as well as the different text mining approaches. For instance, some text mining approaches are solely designed using machine learning methods, others are only developed on the base of lexical databases while a few are built utilizing both methods as a hybrid approach. Nevertheless, there is no such perfect solution that can work for all the problems. It mainly depends on the problem requirement and the data availability. We will pair each approach with at least one example application mainly from our previous work.

7.2 Text Granularity and Text Mining Hyprid Approach

In this section, we will discuss the various text granularity multi-level consideration. We will also talk about the two main techniques, which are machine learning and computational linguistics, in dealing with a text mining problem. Finally, we will discuss how

combining all of these text granularity levels in a hybrid approach of both the statistical and the linguistical techniques could enhance the text mining solution in semantic similarity problems. We summarize this synthetic point of view in Table 7.1.

Text Granularity Level	**Text Mining Approach**	**Applications**
string of characters (without meaning)	machine learning	author name disambiguation
word and sentence semantics	computational linguistics, machine learning	semantic text similarity
documents and knowledge corpus	computational linguistics, machine learning	paper recommender system and metadata semantic enrichment

TABLE 7.1: Text granularity levels and text mining techniques and its usage in our contribution use cases

There are many ways for extracting features from text for difficult tasks like semantic similarity. As discussed in Part I, we can call that task as text vectorization problem in which we extract a numerical representation of the text. However, there is a difference between having string features versus semantic features. It is much simpler to represent a word as a string of characters without counting for the meaning. One-hot-encoder and bag-of-words vectorizations are examples of such vectorizations. These vectorization methods are useful for many tasks as in traditional search engines as well as some machine learning based solutions of specific types of problems, i.e., the author name disambiguation (Section 7.2.1). The machine learning approach that is used to solve such issue does not require a semantic level approach since the people names are meaningless words.

7.2.1 Solving Entity Name Ambiguity: A Case of a Meaningless Words

Another recently raised issue in digital libraries is the ambiguity of some key metadata values like the authors, the affiliations and the references. It is important to have a unique identification for such concepts that sometimes have several used text representation. For example, if we want to list and distinguish the publication list of an author, it appeared that we cannot rely on the text representation used in the metadata. In some cases, the same author has many ways of writing his name and there are other authors who used similar names. Another case is when the digital library wants to construct a citation graph, or a co-authorship graph, in which we need to have a unique identification of

the nodes. Additional need for disambiguation is listing the articles per affiliation or institution name where such names are sometimes not unified in all publications.

The heterogeneous source of publications from different publishers is not the sole cause of this name disambiguation issue. In the same publisher, neither the author name nor the affiliation name are necessarily uniquely identified over years due to lacking of using an unique identifier. Sometimes, the author's email is used as an identifier, but when the author changes his affiliation, the email address is not unique. Other causes could be that some female authors changes their name after marriage, affiliations got merged, or renamed or even facing spin-offs. Moreover, citations are not necessary well mapped to a unique identifier especially for old references.

The problem of author name disambiguation is a special case of bibliographical entity name disambiguation. Working in this problem was my first experience on using text mining to solve metadata issue in digital libraries. I studied this problem as part of my Master's studies internship at CERN[1]. During the beginning on my PhD studies, we published our contribution on that problem with my ex-colleagues at CERN (Louppe et al., 2016). We showed the importance of using a set of author name ethnicity features in designing a scalable semi-supervised learning model. The proposed model, see Figure 7.1, consists in two main phases, the first one is to group a set of *Signatures* (metadata occurrence of an author-name and the publication-id he co-authored) in a *Block* (a potential list of publications for a single real author with several variations of how the name is written in the metadata).

The blocking could be done using a phonetic similarity of the author name or simply by using the last name and the initial letter of the first name. The second phase is to apply a threshold guided *Hierarchical Clustering* in order to distinguish between the different real authors. The numeric similarity values are estimated by learning a *Linkage Function* based on the features extracted from the publication metadata details (e.g., list of co-authors, topic, affiliation, title, references, *etc.*). The guided threshold to flatten the clusters is experimentally designed based on a human curated information of knowing a few number of signatures that belong to a single real author. The proposed ethnicity sensitive semi-supervised approach achieved more than 98% accuracy on a dataset of more than one million signatures. The proposed author name disambiguation system

[1] `https://home.cern/`

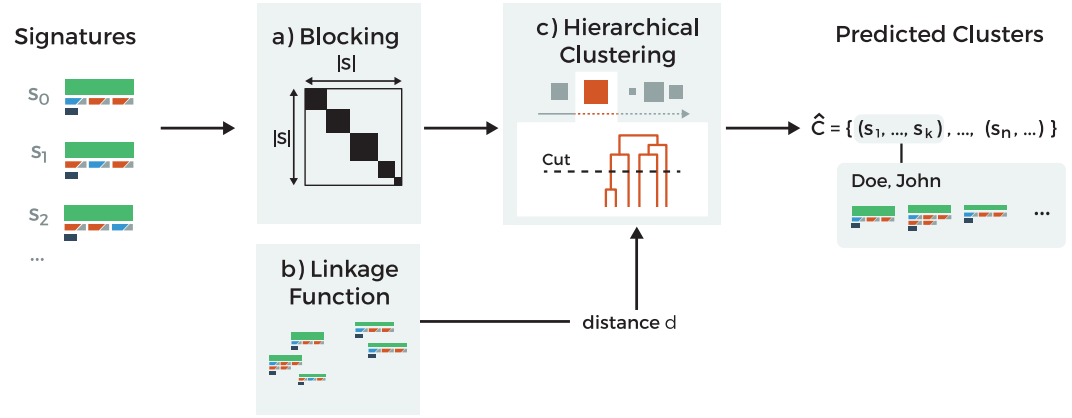


FIGURE 7.1: Pipeline for author disambiguation: (a) signatures are *blocked* to reduce computational complexity, (b) a linkage function is built with supervised learning, (c) independently within each block, signatures are grouped using hierarchical agglomerative clustering. (Figure as in my co-published work with my ex-colleagues at CERN (Louppe et al., 2016))

has been published as open-source project[2] as part of the information retrieval system used at CERN digital library (INSPIRE-HEP)[3].

7.2.2 Various Semantic Level Approaches

A higher level of vectorization or text representation is the semantic one. Word embedding and sentence semantic dense representation are examples of such techniques. although we could extract such semantic features using machine learning techniques over big text corpora, computational linguistics techniques and lexical databases could also enhance such representation. For instance, knowing if a word is a noun or verb, if a term is a name of a place or a person, and knowing the synonym set of a term would enhance the machine capability in semantic similarity tasks. We showed the performance of using such hybrid approach in a couple of use cases in Chapter 4.

A higher text granularity level is the documents and text corpus. Finding a document semantic representation in respect to a corpus of documents is another level of semantic feature extraction. This could be used for some applications like text corpus semantic-based expansion as well as paper recommender systems as we presented in Chapter 5. As in word and sentence cases, we could also use computational linguistics approaches and lexical databases in addition to machine learning techniques in order to enhance

[2] `https://github.com/inspirehep/beard`
[3] `http://inspirehep.net`

the semantic similarity accuracy of such applications. We also presented how a sentence semantic similarity estimator could be used on a paper title level for semantic evaluation of the recommended papers.

Metadata semantic-based enrichment with scientific category names learned on the context of a scientific text corpus is another practical example of using such hybrid text mining approach, as we showed in Chapter 6. Therefore, we believe that utilizing all levels of text granularity and combining several text mining techniques is necessary for better semantic exploration of a text corpus. For instance, the semantic information extracted from a linguistical knowledge graph like in *BabelNet*, contributed to the accuracy of the machine learning approach for semantic-based multi-labelling task. This work is expected to be enhanced further when we consider word embedding and sentence embedding in the pipeline. We believe that such multi-level hybrid approach is not only useful in the use cases we experimented but also for many other text mining applications.

In the following sections (Section 7.2.3.1 and Section 7.2.3.2), we will provide a study of a high semantic level exploration result (i.e., diversity and unexpectedness) of using our proposed approach in Chapter 6. Using the semantic tags in the enriched metadata, the search engine of the digital library can now recommend new papers, in which we have measured the accuracy based on relevancy. However, the relevancy is not the only "relevant" metric for evaluating a recommender system. The diversity of the results and the unexpectedness are other measures that determine the quality of using the system. These two measures will be considered in the model comparison and the pipeline experimental design. In our context, promoting trans-disciplinary research, detecting multidisciplinary articles and tagging them with diverse scientific topic tags would provide better value to the user in addition to discovering unexpected articles that are poorly tagged without such proposed pipeline.

7.2.3 Introduction of Diversity and Unexpectedness

Novelty, *serendipity*, *diversity* and *unexpectedness* are criteria that have come to modulate the field of recommender systems which focused, until recently, only on the improvement of *accuracy*. It has been shown that user satisfaction is negatively dependent on novelty and positively dependent on diversity (Ekstrand et al., 2014), even if a diversity that is badly used can lead users to mistrust the recommender system (Castagnos, Brun, and

Boyer, 2013). Recommending to users items that defer from what they expect from the system is another interesting strategy (Adamopoulos and Tuzhilin, 2014), that is why unexpectedness must also be taken into consideration.

In our approach, we are not recommending items directly, whether these items are research papers, tags or keywords associated to these research papers, but we propose an approach for extending the initial keywords characterizing research articles for a better finding, and trying to favour cross-disciplinary exchanges with the introduction of diversity and unexpectedness.

We can summarize our hypothesis in the following way:

- the addition of diversity and unexpectedness is possible by introducing elements from the semantic vector representation: the more we add information from a word-embedding approach, the more the terms will be semantically far from the initial concepts, and the more we will get diversity and unexpectedness;
- in our results, the more diversity and unexpectedness, the less accuracy will be found;
- the introduction of diversity and unexpectedness will favour discoveries between different scientific disciplines.

Referring to our experiments introduced in Chapter 6, we will study the gradual introduction of diversity by changing the value of the hyper-parameter a for having more or less information from the semantic vector representation: Synset list only, Fusion with $a = 1$ (i.e., as many items in the fusion list as in synset list), Fusion with $a = 2$, Fusion with $a = 3$, and Fusion with $a = 4$.

7.2.3.1 Measuring Diversity and Unexpectedness

A recommender system that provides accurate recommendations but only expected results does not actually provide a great value to the user. Recommending unexpected items does the contrary. This is one of the key principle in data mining and knowledge discovery where recommender systems are not an exception. Unexpectedness of a recommended item could be defined as the distance of the item from the set of expected items

(Adamopoulos and Tuzhilin, 2014). The distance between tags would be determine in the semantic space of word dense vector representation, i.e., word embedding (Mikolov et al., 2013b). Using this word embedding model, the similarity between two word vectors could be estimated by the cosine similarity between their vector representations:

$$D(word_1, word_2) = 1.0 - \cos(\text{Vect}(word_1), \text{Vect}(word_2)) \tag{7.1}$$

Thus the distance in the experiments will be the estimated dissimilarity D as in Equation 7.1. This dissimilarity could be measured between the item and the centroid of the set of expected items or aggregated for all pairs between the recommendation and the expected set. The aggregation could be taking the mean, the minimum or the maximum. In our case, we are recommending tags, we will then consider the set of expected items as the test set while considering the recommendations as the predicted tags. We will apply mean aggregation for all pairs between the recommendation and the expected set. In order to provide an overall unexpectedness measure, we will take the mean of the unexpectedness of all articles tagged by the model as in Equation 7.2 where D is the dissimilarity function as in Equation 7.1, R is the list of recommended tags and E is the list of expected tags.

$$\text{Unexpectedness} = \frac{1}{|R| + |E|} \sum_{i=1}^{|R|} \sum_{j=1}^{|E|} D(R_i, E_j) \tag{7.2}$$

Diversity is another important measure for the recommender system. It could be defined as the average of dissimilarity between all pairs of the recommended list (Adomavicius and Kwon, 2012). A good diversity measure value would avoid recommending "very similar" items which the user would not necessary find useful. Similar to what we described for unexpectedness, we will use word embedding as the semantic dense vector representation to compute the dissimilarity among the recommended items. We will also use mean aggregation of the diversity measure of all articles to come up with the overall model diversity as in Equation 7.3, where P is the set of all tag pairs, D is the dissimilarity function as in Equation 7.1 and R is the recommended list of tags.

$$\text{Diversity} = \frac{1}{|P|} \sum_{i=1}^{|R|} \sum_{j=i}^{|R|} D(R_i, R_j) \tag{7.3}$$

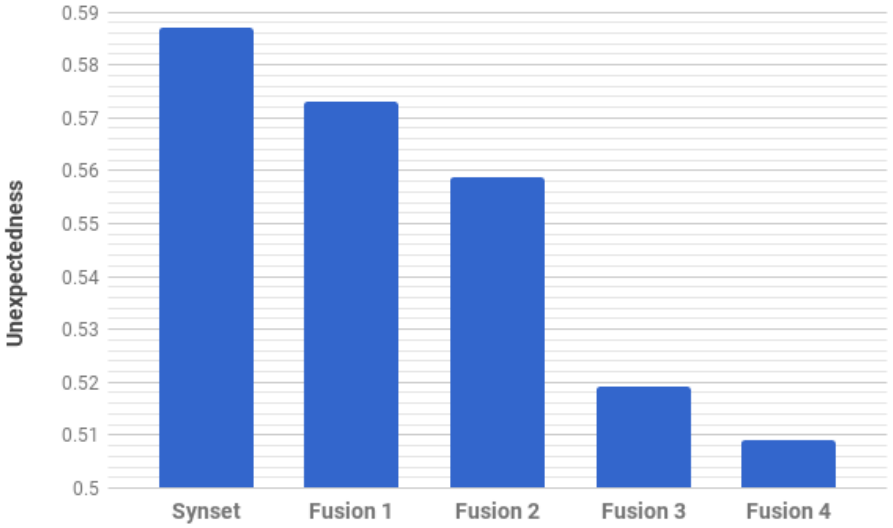


FIGURE 7.2: Unexpectedness measure comparison of the five models

7.2.3.2 Diversity and Unexpectedness Results

TABLE 7.2: Dissimilarity matrix of a sample of scientific topic names computed based on Equation 7.1 using a pre-trained word embedding model "GoogleNews-vector-Negative300." The below-diagonal part of the matrix is left blank as it equals to the above-diagonal part of this symmetric matrix.

Dissimilarity Matrix	Robotics	Artificial-Intelligence	Religion	Philosophy	Surgery	Pathology	Ceramics
Robotics	0.0000	**0.5773**	0.8918	0.8194	0.7650	0.7182	0.7505
Artificial-Intelligence		0.0000	0.7981	*0.6839*	0.8890	0.7831	0.8880
Religion			0.0000	**0.5414**	0.8256	0.8434	0.8891
Philosophy				0.0000	0.7523	0.7009	0.7721
Surgery					0.0000	**0.5822**	0.8863
Pathology						0.0000	0.8211
Ceramics							0.0000

We started this experiment by validating our proposed dissimilarity function D of Equation 7.1. A sample of topic dissimilarity values is presented in Table 7.2 as a dissimilarity matrix. We can see a few bold-font highlighted values in Table 7.2 of pairs that have relatively low dissimilarity which are (Artificial-Intelligence, Robotics), (Religion, Philosophy) and (Surgery, Pathology). These values make sense as they are semantically very related. An interesting italic-font highlighted value is the one for the pair (Artificial-Intelligence, Philosophy) which has a relatively average value as they are slightly related unlike for example (Ceramics, Religion) which are far to be any related. AI and Philosophy share many concepts like action, consciousness, epistemology, and even free will (McCarthy, 2008).

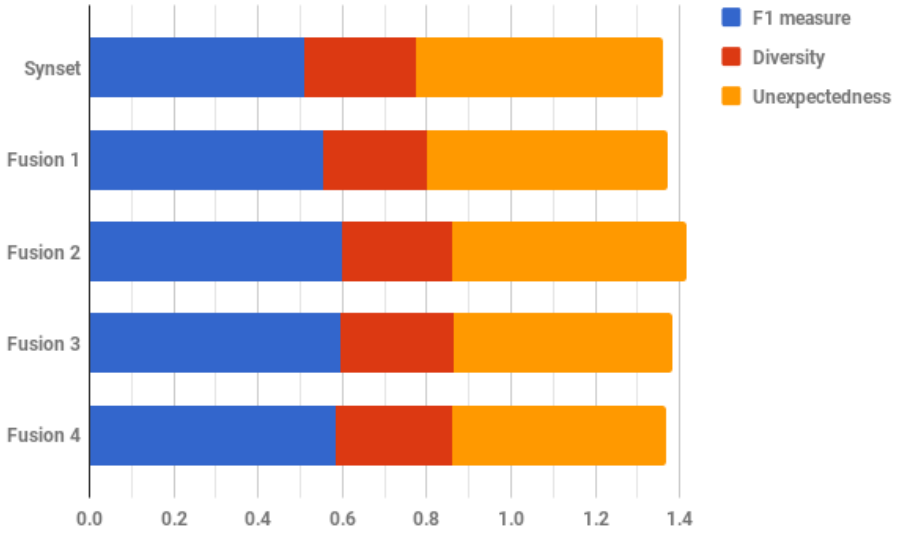


FIGURE 7.3: Stacked values of F1-Measure, Unexpectedness measure and Diversity measure comparison of the five models

The results of unexpectedness and diversity are illustrated in Figure 7.2 and Figure 7.3. These results were computed using a published code[4].

As could be observed from Figure 7.2, when the size of tagged list of articles increases, the unexpectedness decreases. *Synset* and *Fusion1*, however, have the same list size but the Synset model has a higher unexpectedness. We can also note that there is a drop of the unexpectedness measure value after *Fusion2*. Since a good recommender system quality is estimated as a trade-off between accuracy and unexpectedness, the reported results propose picking *Fusion2* model as the best one having the unexpectedness being dropped afterwards when we increase the list of tagged articles. As for the diversity, the results show that as we increase the size of the article list in the fusion method pipeline, the diversity increases, which is kind of trivial since increasing the number of recommended tags eventually leads to increasing the calculated diversity. *Synset* model however, came in the median position among the diversity measure values of the five models. So, we might not be able to say much nor decide on the best model if only considering the diversity measure. Accordingly, we have constructed a stacked chart as in Figure 7.3 where we can also consider both the accuracy and the unexpectedness. As a result of this stacked chart, we can indicate that *Fusion2* model is the best overall model.

[4] `https://github.com/ERICUdL/diversity_measure`

7.3 Conclusion

One of the common principles among text mining practitioners is that there is no such single solution for all problems. As we saw in Section 7.2.1, it was very sufficient to provide statistical learning approach dealing with text as meaningless string of characters. In the other side, some problems require capturing textual semantics in which extra techniques are usually used. Despite having many and various semantic text similarity methods, as discussed in Chapter 3, that are generally categorized into two main methods: One is based on statistical learning and the other is based on semantic networks.

Providing a hybrid approach of these two categories performed better in many text mining solutions. Hybrid approaches did not only provide higher accuracy evaluation, but also showed better level of diversity and unexpectedness of the results.

Chapter 8

General Conclusion and Perspectives

8.1 Conclusion: AI for Digital Libraries

In this thesis, we studied the problem of the search engine limitation of retrieving semantically relevant document beyond keywords matching. A semantic-based approach to explore such documents is needed to wider the knowledge access of researchers who use digital libraries. The need of such solution is required mainly for interdisciplinary research where various scientific domains tend to use different terminologies to describe the interdisciplinary topic. Towards solving this issue we provided the following contributions;

Firstly, a sentence semantic representation model based on sentence pairwise features. The model is able to utilize both linguistic features as well as unsupervised word and sentence embedding features. The model performed well not only in a semantic text similarity benchmark, but also in a couple of use cases. The first use case was the ability to identify similar sentences and documents written in different styles, while the other use case was to highlight (to pair) sentences in the paper abstract to their semantically similar sentences in the paper content.

Secondly, we proposed a novel pipeline for expanding a corpus of an interdisciplinary research topic. The pipeline recommends semantically relevant articles that does not

necessary use the same terminologies of the topic name across related scientific disciplines. We have presented a use case on a multidisciplinary digital library that contains millions of articles for an interdisciplinary topic. The topic domain experts have manually evaluated the recommended articles against another recommendation system where our pipeline performed much better. The model also shows correlated evaluation results using title-to-title semantic similarity estimation model we developed for sentence semantic similarity. Sub-topics diversity analysis was conducted to the recommended results of both models where our approach also provided better recommendation diversity.

Thirdly, we applied a hybrid approach using the pipeline of the previous corpus expansion model and the semantic query expansion using lexical databases to enrich the metadata of a multi-disciplinary digital library. We provided a case study of 33 scientific topic categories taken from "Web of Science" and conducted an evaluation experiment against topic modelling technique. We also experimentally determined the best performing hybrid system based on the amount of fusion between the two semantic approaches. Results also showed good diversity and unexpectedness of the results using the enriched metadata.

Lastly, we provided an overview of the text granularity levels and the various text mining approaches. We concluded that hybrid approaches based on both statistical machine learning semantic features and utilizing semantic networks are usually better for solving semantic text similarity problems. Not only better in accuracy, the hybrid approach is also useful for other aimed evaluation metrics like diversity and unexpectedness.

8.2 Perspectives: Fostering Trans-disciplinary Research

There are some points that we would like to improve in our proposed approaches. First of all, we would like to enhance the hybrid pipeline of metadata enrichment with two additional components: the sentence embedding of paper titles as well as the use of word embedding for expanding the synonym sets in the process of the query expansion. We could also extract more semantically related terms for the query expansion, including, for example, the category names of the scientific topic name found in the semantic network. Also, for our proposed sentence semantic similarity estimation model (SenSim), we are planning to conduct further experiments on all other available benchmarks. We would

also like to combine other features to our pairwise feature sets (i.e., sentence embedding and word embedding of synonyms).

Secondly, we would like to explore and study the benefit of adopting the recent advancements in deep learning based methods for text embedding (published after our contributions). For instance, there are two very interesting approaches, one called ELMo (Peters et al., 2018) and the other called BERT (Devlin et al., 2018).

Thirdly, we would like to provide more case studies and applications of our proposed method of "semantic-based metadata enrichment" in domains other than scientific digital libraries. For example, the same approach could be used in other information retrieval systems of news articles, electronic encyclopedias (like Wikipedia) and semantic-based recommender systems of content relevant advertisements.

Lastly, we think that the natural language is one of the most complex tasks in Artificial Intelligence (AI). Based on our study on the field, we believe that we are still far from solving this problem. However, applying recent AI advancements in text mining for solving information retrieval issues in the digital libraries, including our contributions, provided some encouraging results. We think that the research on that direction should continue. We hope to have human usability validation of our proposed approaches. We would like to take it into an implemented solution for the information retrieval system of a widely used digital library and measure its impact. During this thesis, we only had the chance to experiment one interdisciplinary topic with a human usability validation of two experts (Al-Natsheh et al., 2017a); however, we could not yet validate the recent proposed approach for a larger set of topics (Al-Natsheh et al., 2018). Our study aims at helping researchers to push their discipline boundaries, and at motivating them to work in more trans-disciplinary research towards breakthrough discoveries. Showing measured enhancements in practice will motivate further research on this domain and more "AI-powered" solutions for our humankind treasure of knowledge, the digital libraries.

Appendix A

List of paper publications during the PhD studies

1. Hussein T. Al-Natsheh, Lucie Martinet, Fabrice Muhlenbach, Fabien Rico, and Djamel Abdelkader Zighed. Metadata enrichment of multi-disciplinary digital library: A semantic-based approach. In Eva Méndez, Fabio Crestani, Cristina Ribeiro, Gabriel David, and João Correia Lopes, editors, Digital Libraries for Open Knowledge, pages 32–43, Cham, 2018. Springer International Publishing. ISBN 978-3-030-00066-0.

2. Lucie Martinet, Hussein T. Al-Natsheh, Fabien Rico, Fabrice Muhlenbach, and Djamel A. Zighed. Étiquetage thématique automatisé de corpus par représentation sémantique. volume Extraction et Gestion des Connaissances, RNTI-E-34, pages 323–328, 2018.

3. H. T. Al-Natsheh, L. Martinet, F. Muhlenbach, F. Rico and D. A. Zighed, "Semantic Search-by-Examples for Scientific Topic Corpus Expansion in Digital Libraries," 2017 IEEE International Conference on Data Mining Workshops (ICDMW), New Orleans, LA, 2017, pp. 747-756. doi: 10.1109/ICDMW.2017.103

4. Al-Natsheh, H. T., Martinet, L., Muhlenbach, F., and Zighed, D. A. (2017). Udl at semeval-2017 task 1: Semantic textual similarity estimation of English sentence pairs using regression model over pairwise features. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017) (pp. 115-119). doi: 10.18653/v1/S17-2013

5. Louppe G., Al-Natsheh H.T., Susik M., Maguire E.J. (2016) Ethnicity Sensitive Author Disambiguation Using Semi-supervised Learning. In: Ngonga Ngomo AC.,

Křemen P. (eds) Knowledge Engineering and Semantic Web. KESW 2016. Communications in Computer and Information Science, vol 649. Springer, Cham. doi: 10.1007/978-3-319-45880-9_21

Other conference papers before the PhD studies

1. H. T. Al-Natsheh and A. M. S. Zalzala, "Commercializing computational intelligence techniques in a business intelligence application," IEEE Congress on Evolutionary Computation, Barcelona, 2010, pp. 1-7. doi: 10.1109/CEC.2010.5586249

2. H. T. Al-Natsheh and T. M. Eldos, "Performance Optimization of Adaptive Resonance Neural Networks Using Genetic Algorithms," 2007 IEEE Symposium on Foundations of Computational Intelligence, Honolulu, HI, 2007, pp. 143-148. doi: 10.1109/FOCI.2007.372160

National posters

1. H. T. Al-Natsheh, F. Muhlenbach, L. Martinet, F. Rico and D. A. Zighed, "3SH: “Semantic-Similarity Shadow Hunter” for Scientific Articles", StatLearn 2017, April 6-7, 2017, Lyon, France

2. H. T. Al-Natsheh, F. Muhlenbach, and D. A. Zighed, "Investigation et Exploitation de Corpus Textuels Scientifiques", Journée Scientifique de l’ARC 6 à Lyon à l’Université Lyon 1 le Jeudi 24 Novembre 2016, Lyon, France

3. H. T. Al-Natsheh, F. Muhlenbach, and D. A. Zighed, "Sentence Semantic Similarity", StatLearn 2016, April 7-8 2016, Vannes, France

References

Abrizah, Abdullah, A. N. Zainab, Kiran Kaur, and Ram Gopal Raj (2013). "LIS journals scientific impact and subject categorization: a comparison between Web of Science and Scopus". In: *Scientometrics* 94.2, pp. 721–740.

Adamopoulos, Panagiotis and Alexander Tuzhilin (2014). "On Unexpectedness in Recommender Systems: Or How to Better Expect the Unexpected". In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 5.4, 54:1–54:32. URL: http://doi.acm.org/10.1145/2559952.

Adomavicius, Gediminas and YoungOk Kwon (2012). "Improving Aggregate Recommendation Diversity Using Ranking-Based Techniques". In: *IEEE Transactions on Knowledge and Data Engineering (TKDE)* 24.5, pp. 896–911. DOI: 10.1109/TKDE.2011.15.

Agirre, Eneko, Enrique Alfonseca, Keith B. Hall, Jana Kravalova, Marius Pasca, and Aitor Soroa (2009). "A Study on Similarity and Relatedness Using Distributional and WordNet-based Approaches". In: *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, May 31 - June 5, 2009, Boulder, Colorado, USA*. The Association for Computational Linguistics (ACL), pp. 19–27. URL: http://www.aclweb.org/anthology/N09-1003.

Agirre, Eneko, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe (2016). "SemEval-2016 Task 1: Semantic Textual Similarity, Monolingual and Cross-Lingual Evaluation". In: *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. Ed. by Steven Bethard, Daniel Cer, Marine Carpuat, David Jurgens, Preslav Nakov, and Torsten Zesch. San Diego, California: Association for Computational Linguistics (ACL), pp. 497–511.

Al-Natsheh, Hussein T., Lucie Martinet, Fabrice Muhlenbach, Fabien Rico, and Djamel A. Zighed (2017a). "Semantic Search-by-Examples for Scientific Topic Corpus Expansion in Digital Libraries". In: *2017 IEEE International Conference on Data Mining Workshops, ICDM Workshops 2017, New Orleans, LA, USA, November 18-21, 2017*. Ed. by Raju Gottumukkala, Xia Ning, Guozhu Dong, Vijay Raghavan, Srinivas Aluru, George Karypis, Lucio Miele, and Xindong Wu. IEEE Computer Society, pp. 747–756. URL: https://doi.org/10.1109/ICDMW.2017.103.

Al-Natsheh, Hussein T., Lucie Martinet, Fabrice Muhlenbach, and Djamel Abdelkader Zighed (2017b). "UdL at SemEval-2017 Task 1: Semantic Textual Similarity Estimation

of English Sentence Pairs Using Regression Model over Pairwise Features". In: *Proceedings of the 11th International Workshop on Semantic Evaluation, SemEval@ACL 2017, Vancouver, Canada, August 3-4, 2017*. Ed. by Steven Bethard, Marine Carpuat, Marianna Apidianaki, Saif M. Mohammad, Daniel M. Cer, and David Jurgens. Vancouver, Canada: Association for Computational Linguistics (ACL), pp. 115–119. URL: https://doi.org/10.18653/v1/S17-2013.

Al-Natsheh, Hussein T., Lucie Martinet, Fabrice Muhlenbach, Fabien Rico, and Djamel Abdelkader Zighed (2018). "Metadata Enrichment of Multi-disciplinary Digital Library: A Semantic-Based Approach". In: *Digital Libraries for Open Knowledge, 22nd International Conference on Theory and Practice of Digital Libraries, TPDL 2018, Porto, Portugal, September 10-13, 2018, Proceedings*. Ed. by Eva Méndez, Fabio Crestani, Cristina Ribeiro, Gabriel David, and João Correia Lopes. Vol. 11057. Lecture Notes in Computer Science. Cham: Springer, pp. 32–43. ISBN: 978-3-030-00066-0. URL: https://doi.org/10.1007/978-3-030-00066-0_3.

Alexander, Eric C., Joe Kohlmann, Robin Valenza, Michael Witmore, and Michael Gleicher (2014). "Serendip: Topic model-driven visual exploration of text corpora". In: *2014 IEEE Conference on Visual Analytics Science and Technology, VAST 2014, Paris, France, October 25-31, 2014*. Ed. by Min Chen, David S. Ebert, and Chris North. IEEE Computer Society, pp. 173–182.

Altshuller, Genrich (2002). *40 principles: TRIZ keys to innovation. Translated by Lev Shulyak and Steven Rodman*. Vol. 1. Technical Innovation Center, Inc. ISBN: 0-9640740-3-6.

Amir, Samir, Adrian Tanasescu, and Djamel A. Zighed (2017). "Sentence similarity based on semantic kernels for intelligent text retrieval". In: *Journal of Intelligent Information Systems (JIIS)* 48.3, pp. 675–689. URL: https://doi.org/10.1007/s10844-016-0434-3.

Beel, Joeran, Stefan Langer, Marcel Genzmehr, Bela Gipp, Corinna Breitinger, and Andreas Nürnberger. "Research Paper Recommender System Evaluation: A Quantitative Literature Survey". In: *RepSys'13, Proceedings of the International Workshop on Reproducibility and Replication in Recommender Systems Evaluation, Hong Kong, China, October 12, 2013*. New York, NY, USA: ACM, pp. 15–22.

Beel, Joeran, Bela Gipp, Stefan Langer, and Corinna Breitinger (2016). "Research-paper recommender systems: a literature survey". In: *International Journal on Digital Libraries* 17.4, pp. 305–338.

Berners-Lee, Tim, James Hendler, and Ora Lassila (2001). "The Semantic Web". In: *Scientific American* 284.5, pp. 34–43.

Blei, David M (2012). "Probabilistic Topic Models". In: *Communications of the ACM* 55.4, pp. 77–84.

Blei, David M. and Jon D. McAuliffe (2007). "Supervised Topic Models". In: *Advances in Neural Information Processing Systems 20 (NIPS 2007)*. Ed. by J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis. Curran Associates, Inc., pp. 121–128. URL: `http://papers.nips.cc/paper/3328-supervised-topic-models.pdf`.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan (2003). "Latent Dirichlet Allocation". In: *Journal of Machine Learning Research* 3, pp. 993–1022. URL: `http://www.jmlr.org/papers/v3/blei03a.html`.

Bodrunova, Svetlana, Sergei Koltsov, Olessia Koltsova, Sergey I. Nikolenko, and Anastasia Shimorina (2013). "Interval Semi-supervised LDA: Classifying Needles in a Haystack". In: *Advances in Artificial Intelligence and Its Applications - 12th Mexican International Conference on Artificial Intelligence (MICAI) Part I*. Ed. by Félix Castro-Espinoza, Alexander F. Gelbukh, and Miguel González-Mendoza. Vol. 8265. Lecture Notes in Computer Science. Springer, pp. 265–274. URL: `https://doi.org/10.1007/978-3-642-45114-0\_21`.

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov (2017). "Enriching Word Vectors with Subword Information". In: *Transactions of the Association for Computational Linguistics (TACL)* 5, pp. 135–146. URL: `https://transacl.org/ojs/index.php/tacl/article/view/999`.

Borgida, Alexander and John F. Sowa (1991). *Principles of semantic networks – Explorations in the representation of knowledge*. Morgan Kaufmann.

Borgman, Christine L (2003). *From Gutenberg to the global information infrastructure: access to information in the networked world*. MIT Press. ISBN: 0-262-02473-X.

Bottou, Léon (2014). "From Machine Learning to Machine Reasoning". In: *Machine Learning* 94.2, pp. 133–149.

Castagnos, Sylvain, Armelle Brun, and Anne Boyer (2013). "When Diversity Is Needed... But Not Expected!" In: *International Conference on Advances in Information Mining and Management (IMMM)*. ARIA XPS Press, pp. 44–50.

Castells, Pablo, Neil J. Hurley, and Saul Vargas (2015). "Novelty and Diversity in Recommender Systems". In: *Recommender Systems Handbook*. Ed. by Francesco Ricci,

Lior Rokach, and Bracha Shapira. Second Edition. Springer, pp. 881–918. ISBN: 978-1-4899-7636-9. DOI: 10.1007/978-1-4899-7637-6_26.

Cer, Daniel, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia (2017). "SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation". In: *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Vancouver, Canada: Association for Computational Linguistics (ACL), pp. 1–14. URL: http://www.aclweb.org/anthology/S17-2001.

Che, Wanxiang, Zhenghua Li, and Ting Liu (2010). "LTP: A Chinese Language Technology Platform". In: *COLING 2010, 23rd International Conference on Computational Linguistics, Demonstrations Volume, Beijing, China*. Demonstrations Volume, pp. 13–16. URL: http://aclweb.org/anthology/C/C10/C10-3004.pdf.

Chen, Danqi and Christopher D. Manning (2014). "A Fast and Accurate Dependency Parser using Neural Networks". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*. Ed. by Alessandro Moschitti, Bo Pang, and Walter Daelemans. ACL, pp. 740–750. URL: http://aclweb.org/anthology/D/D14/D14-1082.pdf.

Chiarello, Christine, Curt Burgess, Lorie Richards, and Alma Pollock (1990). "Semantic and associative priming in the cerebral hemispheres: Some words do, some words don't... sometimes, some places". In: *Brain and language* 38.1, pp. 75–104.

Cho, Kyunghyun, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio (2014). "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*. Ed. by Alessandro Moschitti, Bo Pang, and Walter Daelemans. ACL, pp. 1724–1734. URL: http://aclweb.org/anthology/D/D14/D14-1179.pdf.

Choi, Bernard C. K. and Anita W. P. Pak (2006). "Multidisciplinarity, interdisciplinarity and transdisciplinarity in health research, services, education and policy: 1. Definitions, objectives, and evidence of effectiveness". In: *Clinical & Investigative Medicine* 29.6, pp. 351–364.

Chung, Junyoung, Çaglar Gülçehre, KyungHyun Cho, and Yoshua Bengio (2014). "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling". In: *CoRR* abs/1412.3555. URL: http://arxiv.org/abs/1412.3555.

Cohen, Jacob (1960). "A coefficient of agreement for nominal scales". In: *Educational and psychological measurement* 20.1, pp. 37–46.

Cojan, Julien, Elena Cabrio, and Fabien Gandon (2013). "Filling the gaps among DBpedia multilingual chapters for question answering". In: *Proceedings of the 5th Annual ACM Web Science Conference*. ACM, pp. 33–42.

Collobert, Ronan, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa (2011). "Natural language processing (almost) from scratch". In: *Journal of Machine Learning Research* 12.Aug, pp. 2493–2537.

Conneau, Alexis, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes (2017). "Supervised Learning of Universal Sentence Representations from Natural Language Inference Data". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*. Ed. by Martha Palmer, Rebecca Hwa, and Sebastian Riedel. Association for Computational Linguistics (ACL), pp. 670–680. URL: https://aclanthology.info/papers/D17-1070/d17-1070.

Craciun, Cerasella (2014). "Pluridisciplinarity, Interdisciplinarity and Transdisciplinarity: Methods of Researching the Metabolism of the Urban Landscape". In: *Planning and Designing Sustainable and Resilient Landscapes*. Ed. by Cerasella Craciun and Maria Bostenaru Dan. Springer Geography. Springer Netherlands, pp. 3–14. ISBN: 978-94-017-8535-8,978-94-017-8536-5.

Croft, David, Simon Coupland, Jethro Shell, and Stephen Brown (2013). "A fast and efficient semantic short text similarity metric". In: *13th UK Workshop on Computational Intelligence, UKCI 2013, Guildford, United Kingdom, September 9-11, 2013*. IEEE, pp. 221–227. URL: https://doi.org/10.1109/UKCI.2013.6651309.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *CoRR* abs/1810.04805. arXiv: 1810.04805. URL: http://arxiv.org/abs/1810.04805.

Dixit, Bharvi (2017). "Chapter 2. The Improved Query DSL". In: *Mastering Elasticsearch 5.x*. Birmingham, UK: Packt Publishing, Limited, pp. 74–141.

Dolan, William Bill, Chris Quirk, and Chris Brockett (2004). "Unsupervised Construction of Large Paraphrase Corpora: Exploiting Massively Parallel News Sources". In: *COLING 2004, 20th International Conference on Computational Linguistics, Proceedings*

of the Conference, Geneva, Switzerland. International Conference on Computational Linguistics. URL: http://www.aclweb.org/anthology/C04-1051.

Dong, Ruihai and Barry Smyth (2016). "From More-Like-This to Better-Than-This: Hotel Recommendations from User Generated Reviews". In: *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization, UMAP 2016, Halifax, NS, Canada, July 13 - 17, 2016*. Ed. by Julita Vassileva, James Blustein, Lora Aroyo, and Sidney K. D'Mello. ACM, pp. 309–310.

Dong, Zhendong, Qiang Dong, and Changling Hao (2010). "HowNet and Its Computation of Meaning". In: *COLING 2010, 23rd International Conference on Computational Linguistics, Demonstrations Volume, Beijing, China*. Demonstrations Volume, pp. 53–56. URL: http://aclweb.org/anthology/C/C10/C10-3014.pdf.

Duma, Mirela-Stefania and Wolfgang Menzel (2017). "SEF @ UHH at SemEval-2017 Task 1: Unsupervised Knowledge-Free Semantic Textual Similarity via Paragraph Vector". In: *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pp. 170–174.

Ekstrand, Michael D., F. Maxwell Harper, Martijn C. Willemsen, and Joseph A. Konstan (2014). "User perception of differences in recommender algorithms". In: *Eighth ACM Conference on Recommender Systems, RecSys '14, Foster City, Silicon Valley, CA, USA*. Ed. by Alfred Kobsa, Michelle X. Zhou, Martin Ester, and Yehuda Koren. ACM, pp. 161–168. DOI: 10.1145/2645710.2645737.

Evans, Vyvyan (2006). "Lexical concepts, cognitive models and meaning-construction. Cognitive Linguistics. Vol 17.4. pp. 491–534." In: URL: https://doi.org/10.1515/COG.2006.016.

Firth, John R (1957). "A synopsis of linguistic theory, 1930-1955". In: *Studies in linguistic analysis*, pp. 1–32.

Flach, Peter A., José Hernández-Orallo, and Cèsar Ferri Ramirez (2011). "A Coherent Interpretation of AUC as a Measure of Aggregated Classification Performance". In: *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*. Ed. by Lise Getoor and Tobias Scheffer. Omnipress, pp. 657–664.

Ganitkevitch, Juri, Benjamin Van Durme, and Chris Callison-Burch (2013). "PPDB: The Paraphrase Database". In: *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June*

9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA. Ed. by Lucy Vanderwende, Hal Daumé III, and Katrin Kirchhoff. The Association for Computational Linguistics (ACL), pp. 758–764. URL: http://aclweb.org/anthology/N/N13/N13-1092.pdf.

Gao, Jianfeng, Patrick Pantel, Michael Gamon, Xiaodong He, and Li Deng (2014). "Modeling Interestingness with Deep Neural Networks". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*. Ed. by Alessandro Moschitti, Bo Pang, and Walter Daelemans. ACL, pp. 2–13. URL: http://aclweb.org/anthology/D/D14/D14-1002.pdf.

Gers, Felix A., Jürgen Schmidhuber, and Fred A. Cummins (2000). "Learning to Forget: Continual Prediction with LSTM". In: *Neural Computation* 12.10, pp. 2451–2471. URL: https://doi.org/10.1162/089976600300015015.

Girvan, Michelle and Mark E. J. Newman (2002). "Community structure in social and biological networks". In: *Proceedings of the National Academy of Sciences of the United States of America (PNAS)* 99.12, pp. 7821–7826.

Golub, Gene H and Christian Reinsch (1971). "Singular Value Decomposition and Least Squares Solutions". In: *Linear Algebra*. Springer, pp. 134–151.

Graves, Alex and Jürgen Schmidhuber (2005). "Framewise phoneme classification with bidirectional LSTM and other neural network architectures". In: *Neural Networks* 18.5-6, pp. 602–610.

Gretarsson, Brynjar, John O'Donovan, Svetlin Bostandjiev, Tobias Höllerer, Arthur U. Asuncion, David Newman, and Padhraic Smyth (2012). "TopicNets: Visual Analysis of Large Text Corpora with Topic Modeling". In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 3.2, 23:1–23:26.

Guha, Ramanathan V, Dan Brickley, and Steve Macbeth (2016). "Schema. org: evolution of structured data on the web". In: *Communications of the ACM* 59.2, pp. 44–51.

Hagen, Matthias and Christiane Glimm (2014). "Supporting More-Like-This Information Needs: Finding Similar Web Content in Different Scenarios". In: *Information Access Evaluation. Multilinguality, Multimodality, and Interaction - 5th International Conference of the CLEF Initiative, CLEF 2014, Sheffield, UK, September 15-18, 2014. Proceedings*. Ed. by Evangelos Kanoulas, Mihai Lupu, Paul D. Clough, Mark Sanderson, Mark M. Hall, Allan Hanbury, and Elaine G. Toms. Vol. 8685. Lecture Notes in Computer Science. Springer, pp. 50–61.

Halko, Nathan, Per-Gunnar Martinsson, and Joel A. Tropp (2011). "Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions". In: *SIAM Review* 53.2, pp. 217–288. URL: https://doi.org/10.1137/090771806.

He, Junxian, Ying Huang, Changfeng Liu, Jiaming Shen, Yuting Jia, and Xinbing Wang (2016). "Text Network Exploration via Heterogeneous Web of Topics". In: *IEEE International Conference on Data Mining Workshops, ICDM Workshops 2016, December 12-15, 2016, Barcelona, Spain*. Ed. by Carlotta Domeniconi, Francesco Gullo, Francesco Bonchi, Josep Domingo-Ferrer, Ricardo A. Baeza-Yates, Zhi-Hua Zhou, and Xindong Wu. IEEE, pp. 99–106.

He, Qi, Bi Chen, Jian Pei, Baojun Qiu, Prasenjit Mitra, and C. Lee Giles (2009). "Detecting topic evolution in scientific literature: how can citations help?" In: *Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM 2009, Hong Kong, China, November 2-6, 2009*. Ed. by David Wai-Lok Cheung, Il-Yeol Song, Wesley W. Chu, Xiaohua Hu, and Jimmy J. Lin. ACM, pp. 957–966.

Hill, Felix, Roi Reichart, and Anna Korhonen (2015). "SimLex-999: Evaluating Semantic Models With (Genuine) Similarity Estimation". In: *Computational Linguistics* 41.4, pp. 665–695. URL: https://doi.org/10.1162/COLI_a_00237.

Hofmann, Thomas (1999). "Probabilistic Latent Semantic Analysis". In: *UAI '99: Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, Stockholm, Sweden*. Ed. by Kathryn B. Laskey and Henri Prade. Morgan Kaufmann, pp. 289–296. URL: https://dslpitt.org/uai/displayArticleDetails.jsp?mmnu=1\&smnu=2\&article_id=179\&proceeding_id=15.

Huang, Po-Sen, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry P. Heck (2013). "Learning deep structured semantic models for web search using clickthrough data". In: *22nd ACM International Conference on Information and Knowledge Management, CIKM'13, San Francisco, CA, USA, October 27 - November 1, 2013*. Ed. by Qi He, Arun Iyengar, Wolfgang Nejdl, Jian Pei, and Rajeev Rastogi. ACM, pp. 2333–2338. URL: http://doi.acm.org/10.1145/2505515.2505665.

Iacobacci, Ignacio, Mohammad Taher Pilehvar, and Roberto Navigli (2015). "SensEmbed: learning sense embeddings for word and relational similarity". In: *Proceedings of ACL*, pp. 95–105.

Islam, Aminul and Diana Inkpen (2008). "Semantic text similarity using corpus-based word similarity and string similarity". In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 2.2, p. 10.

Ji, Yangfeng and Jacob Eisenstein (2013). "Discriminative Improvements to Distributional Sentence Similarity". In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*. ACL, pp. 891–896. URL: http://aclweb.org/anthology/D/D13/D13-1090.pdf.

Joulin, Armand, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hervé Jégou, and Tomas Mikolov (2016). "FastText.zip: Compressing text classification models". In: *CoRR* abs/1612.03651. URL: http://arxiv.org/abs/1612.03651.

Joyce, James M. (2011). "Kullback-Leibler Divergence". In: *International Encyclopedia of Statistical Science*. Ed. by Miodrag Lovric. Springer, pp. 720–722. URL: https://doi.org/10.1007/978-3-642-04898-2_327.

Kabary, Ihab Al, Ivan Giangreco, Heiko Schuldt, Fabrice Matulic, and Moira C. Norrie (2013). "QUEST: Towards a Multi-modal CBIR Framework Combining Query-by-Example, Query-by-Sketch, and Text Search". In: *2013 IEEE International Symposium on Multimedia, ISM 2013, Anaheim, CA, USA, December 9-11, 2013*. IEEE Computer Society, pp. 433–438.

Kawamura, Takahiro, Kouji Kozaki, Tatsuya Kushida, Katsutaro Watanabe, and Katsuji Matsumura (2016). "Expanding Science and Technology Thesauri from Bibliographic Datasets Using Word Embedding". In: *28th IEEE International Conference on Tools with Artificial Intelligence, ICTAI 2016, San Jose, CA, USA, November 6-8, 2016*. IEEE Computer Society, pp. 857–864. URL: https://doi.org/10.1109/ICTAI.2016.0133.

Kiros, Ryan, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler (2015). "Skip-Thought Vectors". In: *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems (NIPS) 2015, Montreal, Quebec, Canada*. Ed. by Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett, pp. 3294–3302. URL: http://papers.nips.cc/paper/5950-skip-thought-vectors.

Klein, Lauren F., Jacob Eisenstein, and Iris Sun (2015). "Exploratory Thematic Analysis for Digitized Archival Collections". In: *Digital Scholarship in the Humanities* 30.Suppl-1, pp. i130–i141.

Koren, Yehuda (2009). "The bellkor solution to the netflix grand prize". In: *Netflix prize documentation* 81, pp. 1–10.

Kosslyn, Stephen M., William L. Thompson, M. Wraga, and Nathaniel M. Alpert (2001). "Imagining rotation by endogenous versus exogenous forces: distinct neural mechanisms". In: *Neuroreport* 12.11, pp. 2519–2525.

Kresh, Diane (2007). *The whole digital library handbook*. Ed. by Diane Kresh. American Library Association. ISBN: 0-8389-0926-4.

Kruszewski, Germán, Angeliki Lazaridou, Marco Baroni, et al. (2015). "Jointly optimizing word representations for lexical and sentential tasks with the c-phrase model". In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (IJCNLP) (Volume 1: Long Papers)*. Vol. 1. The Association for Computational Linguistics (ACL), pp. 971–981.

Lam, Xuan Nhat, Thuc Vu, Trong Duc Le, and Anh Duc Duong (2008). "Addressing cold-start problem in recommendation systems". In: *Proceedings of the 2nd international conference on Ubiquitous information management and communication*. ACM, pp. 208–211.

Landauer, Thomas K, Peter W Foltz, and Darrell Laham (1998). "An introduction to latent semantic analysis". In: *Discourse processes* 25.2-3, pp. 259–284.

Langley, Patrick W., Herbert A. Simon, Gary Bradshaw, and Jan M. Zytkow (1987). *Scientific Discovery: Computational Explorations of the Creative Process*. Cambridge, MA: The MIT Press.

Latard, Bastien, Jonathan Weber, Germain Forestier, and Michel Hassenforder (2017). "Towards a Semantic Search Engine for Scientific Articles". In: *Research and Advanced Technology for Digital Libraries - 21st International Conference on Theory and Practice of Digital Libraries, TPDL 2017, Thessaloniki, Greece, September 18-21, 2017, Proceedings*. Ed. by Jaap Kamps, Giannis Tsakonas, Yannis Manolopoulos, Lazaros S. Iliadis, and Ioannis Karydis. Vol. 10450. Lecture Notes in Computer Science. Springer, pp. 608–611. DOI: 10.1007/978-3-319-67008-9_54.

Lau, Jey Han and Timothy Baldwin (2016). "An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation". In: *Proceedings of the 1st*

Workshop on Representation Learning for NLP, Rep4NLP@ACL 2016, Berlin, Germany, August 11, 2016. Ed. by Phil Blunsom, Kyunghyun Cho, Shay B. Cohen, Edward Grefenstette, Karl Moritz Hermann, Laura Rimell, Jason Weston, and Scott Wen-tau Yih. Association for Computational Linguistics (ACL), pp. 78–86. URL: https://doi.org/10.18653/v1/W16-1609.

Le, Quoc V. and Tomas Mikolov (2014). "Distributed Representations of Sentences and Documents". In: *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*. Vol. 32. JMLR Workshop and Conference Proceedings. JMLR.org, pp. 1188–1196.

Le, Tuan M. V. and Hady Wirawan Lauw (2016). "Semantic Visualization with Neighborhood Graph Regularization". In: *Journal of Artificial Intelligence Research (JAIR)* 55, pp. 1091–1133.

Lee, Daniel D and H Sebastian Seung (2001). "Algorithms for non-negative matrix factorization". In: *Advances in neural information processing systems*, pp. 556–562.

Lee, Pei, Laks V. S. Lakshmanan, and Jeffrey Xu Yu (2012). "On Top-k Structural Similarity Search". In: *IEEE 28th International Conference on Data Engineering (ICDE 2012), Washington, DC, USA (Arlington, Virginia), 1-5 April, 2012*. Ed. by Anastasios Kementsietsidis and Marcos Antonio Vaz Salles. IEEE Computer Society, pp. 774–785. DOI: 10.1109/ICDE.2012.109. URL: https://doi.org/10.1109/ICDE.2012.109.

Lehmann, Jens, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer (2015). "DBpedia – A large-scale, multilingual knowledge base extracted from Wikipedia". In: *Semantic Web* 6.2, pp. 167–195. DOI: 10.3233/SW-140134.

Li, Xiao and Qingsheng Li (2015). "Calculation of Sentence Semantic Similarity based on Syntactic Structure". In: *Mathematical Problems in Engineering, Vol. 2015, Article ID 203475* 2015.

Li, Yuhua, David McLean, Zuhair Bandar, James D. O'Shea, and Keeley A. Crockett (2006). "Sentence Similarity Based on Semantic Nets and Corpus Statistics". In: *IEEE Transactions on Knowledge and Data Engineering (TKDE)* 18.8, pp. 1138–1150. URL: https://doi.org/10.1109/TKDE.2006.130.

Liang, Feynman, Yuhao Yang, and Joseph Bradley (2015). *Large Scale Topic Modeling: Improvements to LDA on Apache Spark*. https://databricks.com/blog/2015/09/22/large-scale-topic-modeling-improvements-to-lda-on-apache-spark.html.

Lika, Blerina, Kostas Kolomvatsos, and Stathes Hadjiefthymiades (2014). "Facing the cold start problem in recommender systems". In: *Expert Systems with Applications* 41.4, pp. 2065–2073.

Louppe, Gilles, Hussein T. Al-Natsheh, Mateusz Susik, and Eamonn James Maguire (2016). "Ethnicity Sensitive Author Disambiguation Using Semi-supervised Learning". In: *Knowledge Engineering and Semantic Web - 7th International Conference, KESW 2016, Prague, Czech Republic, September 21-23, 2016, Proceedings*. Ed. by Axel-Cyrille Ngonga Ngomo and Petr Kremen. Vol. 649. Communications in Computer and Information Science. Springer, pp. 272–287. ISBN: 978-3-319-45880-9. DOI: 10.1007/978-3-319-45880-9_21. URL: https://doi.org/10.1007/978-3-319-45880-9_21.

Maharjan, Nabin, Rajendra Banjade, Dipesh Gautam, Lasang Jimba Tamang, and Vasile Rus (2017). "DT_Team at SemEval-2017 Task 1: Semantic Similarity Using Alignments, Sentence-Level Embeddings and Gaussian Mixture Model Output". In: *Proceedings of the 11th International Workshop on Semantic Evaluation, SemEval@ACL 2017, Vancouver, Canada, August 3-4, 2017*. Ed. by Steven Bethard, Marine Carpuat, Marianna Apidianaki, Saif M. Mohammad, Daniel M. Cer, and David Jurgens. Association for Computational Linguistics (ACL), pp. 120–124. URL: https://doi.org/10.18653/v1/S17-2014.

Mahdisoltani, Farzaneh, Joanna Biega, and Fabian M. Suchanek (2015). "YAGO3: A Knowledge Base from Multilingual Wikipedias". In: *CIDR 2015, Seventh Biennial Conference on Innovative Data Systems Research, Asilomar, CA, USA, January 4-7, 2015, Online Proceedings*. www.cidrdb.org. URL: http://cidrdb.org/cidr2015/Papers/CIDR15_Paper1.pdf.

Marneffe, Marie-Catherine de, Bill MacCartney, and Christopher D. Manning (2006). "Generating Typed Dependency Parses from Phrase Structure Parses". In: *Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC 2006, Genoa, Italy*. Ed. by Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard, Joseph Mariani, Jan Odijk, and Daniel Tapias. European Language Resources Association (ELRA), pp. 449–454. URL: http://www.lrec-conf.org/proceedings/lrec2006/pdf/440_pdf.pdf.

Martinet, Lucie, Hussein T. Al-Natsheh, Fabien Rico, Fabrice Muhlenbach, and Djamel A. Zighed (2018). "Étiquetage thématique automatisé de corpus par représentation sémantique". In: vol. Extraction et Gestion des Connaissances, RNTI-E-34, pp. 323–328.

Max-Neef, Manfred A. (2005). "Foundations of transdisciplinarity". In: *Ecological Economics* 53, pp. 5–16.

May, Robert M. (1997). "The Scientific Wealth of Nations". In: *Science* 275.5301, pp. 793–796. ISSN: 00368075, 10959203. DOI: 10.1126/science.275.5301.793. URL: http://www.jstor.org/stable/2891640.

McCarthy, John (2008). "The Philosophy of AI and the AI of Philosophy". In: *Philosophy of Information*. Elsevier, pp. 711–740.

Mikolov, Tomáš, Martin Karafiát, Lukáš Burget, Jan Černocký, and Sanjeev Khudanpur (2010). "Recurrent neural network based language model". In: *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association (ISCA)*. Ed. by Takao Kobayashi, Keikichi Hirose, and Satoshi Nakamura. ISCA, pp. 1045–1048. URL: https://www.isca-speech.org/archive/interspeech_2010/i10_1045.html.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean (2013a). "Distributed Representations of Words and Phrases and their Compositionality". In: *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems (NIPS) 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.* Ed. by Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger, pp. 3111–3119. URL: http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean (2013b). "Efficient Estimation of Word Representations in Vector Space". In: *CoRR* abs/1301.3781.

Miller, Eric (1998). "An introduction to the resource description framework". In: *Bulletin of the American Society for Information Science and Technology* 25.1, pp. 15–19.

Miller, George A. (1995). "WordNet: A Lexical Database for English". In: *Communications of the ACM (CACM)* 38.11, pp. 39–41. URL: http://doi.acm.org/10.1145/219717.219748.

Navigli, Roberto and Simone Paolo Ponzetto (2012). "BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network". In: *Artificial Intelligence* 193, pp. 217–250.

Nickel, Maximilian, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich (2016). "A review of relational machine learning for knowledge graphs". In: *Proceedings of the IEEE* 104.1, pp. 11–33.

Nicolescu, Basarab (2010). "Methodology of Transdisciplinarity – Levels of Reality, Logic of the Included Middle and Complexity". In: *Transdisciplinary Journal of Engineering & Science (TJES)* 1.1, pp. 19–38.

O'Shea, James, Zuhair Bandar, and Keeley A. Crockett (2013). "A new benchmark dataset with production methodology for short text semantic similarity algorithms". In: *ACM Transactions on Speech and Language Processing (TSLP)* 10.4, 19:1–19:63. URL: http://doi.acm.org/10.1145/2537046.

Pagliardini, Matteo, Prakhar Gupta, and Martin Jaggi (2018). "Unsupervised Learning of Sentence Embeddings Using Compositional n-Gram Features". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*. Ed. by Marilyn A. Walker, Heng Ji, and Amanda Stent. Association for Computational Linguistics (ACL), pp. 528–540. URL: https://aclanthology.info/papers/N18-1049/n18-1049.

Pariser, Eli (2011). *The Filter Bubble: What The Internet Is Hiding From You*. New York, NY, USA: Penguin Press.

Pavlick, Ellie, Pushpendre Rastogi, Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch (2015). "PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification". In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Beijing, China: Association for Computational Linguistics (ACL), pp. 425–430. URL: http://www.aclweb.org/anthology/P15-2070.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake VanderPlas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research (JMLR)* 12.Oct, pp. 2825–2830. URL: http://dl.acm.org/citation.cfm?id=2078195.

Pennington, Jeffrey, Richard Socher, and Christopher Manning (2014). "Glove: Global vectors for word representation". In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543.

Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer (2018). "Deep Contextualized Word Representations". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*. Ed. by Marilyn A. Walker, Heng Ji, and Amanda Stent. Association for Computational Linguistics (ACL), pp. 2227–2237. URL: https://aclanthology.info/papers/N18-1202/n18-1202.

Pinto, José María González and Wolf-Tilo Balke (2015). "Demystifying the Semantics of Relevant Objects in Scholarly Collections: A Probabilistic Approach". In: *Proceedings of the 15th ACM/IEEE-CE Joint Conference on Digital Libraries (JCDL), Knoxville, TN, USA*. Ed. by Paul Logasa Bogen II, Suzie Allard, Holly Mercer, Micah Beck, Sally Jo Cunningham, Dion Hoe-Lian Goh, and Geneva Henry. ACM, pp. 157–164. URL: https://doi.org/10.1145/2756406.2756923.

Queneau, Raymond (1947). *Exercices de style*. Gallimard; Édition : 1 (16 mars 1982) Collection : Folio.

– (1986). *Exercises in style (1947. Exercices de style), translated by Barbara Wright*. Gaberbocchus Press.

Ramage, Daniel, David Leo Wright Hall, Ramesh Nallapati, and Christopher D. Manning (2009). "Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora". In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, EMNLP 2009, 6-7 August 2009, Singapore, A meeting of SIGDAT, a Special Interest Group of the ACL*. ACL, pp. 248–256. URL: http://www.aclweb.org/anthology/D09-1026.

Ricci, Francesco, Lior Rokach, and Bracha Shapira (2015). "Recommender Systems: Introduction and Challenges". In: *Recommender Systems Handbook*. Ed. by Francesco Ricci, Lior Rokach, and Bracha Shapira. Second Edition. Springer, pp. 1–34. ISBN: 978-1-4899-7636-9. URL: https://doi.org/10.1007/978-1-4899-7637-6_1.

Rubens, Neil, Dain Kaplan, and Masashi Sugiyama (2011). "Active Learning in Recommender Systems". In: *Recommender Systems Handbook*. Ed. by Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor. Second Edition. Springer, pp. 735–767. URL: https://doi.org/10.1007/978-0-387-85820-3_23.

Sacco, Giovanni Maria and Yannis Tzitzikas, eds. (2009). *Dynamic Taxonomies and Faceted Search: Theory, Practice, and Experience*. Springer.

Sadato, Norihiro, Tomohisa Okada, Manabu Honda, Ken-Ichi Matsuki, Masaki Yoshida, Ken-Ichi Kashikura, Wataru Takei, Tetsuhiro Sato, Takanori Kochiyama, and Yoshiharu Yonekura (2005). "Cross-modal integration and plastic changes revealed by lip movement, random-dot motion and sign languages in the hearing and deaf". In: *Cerebral Cortex* 15.8, pp. 1113–1122.

Salle, Alexandre, Marco Idiart, and Aline Villavicencio (2016). "Enhancing the LexVec Distributed Word Representation Model Using Positional Contexts and External Memory". In: *CoRR* abs/1606.01283. URL: http://arxiv.org/abs/1606.01283.

Sapir, Edward (1985). *Culture, language and personality: Selected essays*. Vol. 342. ISBN: 978-0520011168.

Schafer, J. Ben, Dan Frankowski, Jonathan L. Herlocker, and Shilad Sen (2007). "Collaborative Filtering Recommender Systems". In: *The Adaptive Web, Methods and Strategies of Web Personalization*. Ed. by Peter Brusilovsky, Alfred Kobsa, and Wolfgang Nejdl. Vol. 4321. Lecture Notes in Computer Science. Springer, pp. 291–324. URL: https://doi.org/10.1007/978-3-540-72079-9_9.

Schein, Andrew I., Alexandrin Popescul, Lyle H. Ungar, and David M. Pennock (2002). "Methods and metrics for cold-start recommendations". In: *SIGIR 2002: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 11-15, 2002, Tampere, Finland*. Ed. by Kalervo Järvelin, Micheline Beaulieu, Ricardo A. Baeza-Yates, and Sung-Hyon Myaeng. ACM, pp. 253–260. URL: https://doi.org/10.1145/564376.564421.

Scientific and Technical Information Department – CNRS (2016). *White Paper – Open Science in a Digital Republic*. http://books.openedition.org/oep/1635. OpenEdition Press. DOI: 10.4000/books.oep.1635.

Shao, Yang (2017). "HCTI at SemEval-2017 Task 1: Use convolutional neural network to evaluate Semantic Textual Similarity". In: *Proceedings of the 11th International Workshop on Semantic Evaluation, SemEval@ACL 2017, Vancouver, Canada, August 3-4, 2017*. Ed. by Steven Bethard, Marine Carpuat, Marianna Apidianaki, Saif M. Mohammad, Daniel M. Cer, and David Jurgens. Association for Computational Linguistics (ACL), pp. 130–133. URL: https://doi.org/10.18653/v1/S17-2016.

Shen, Yelong, Xiaodong He, Jianfeng Gao, Li Deng, and Grégoire Mesnil (2014). "A Latent Semantic Model with Convolutional-Pooling Structure for Information Retrieval". In: *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM 2014, Shanghai, China, November 3-7,*

2014. Ed. by Jianzhong Li, Xiaoyang Sean Wang, Minos N. Garofalakis, Ian Soboroff, Torsten Suel, and Min Wang. ACM, pp. 101–110.

Shepard, Roger N. and Jacqueline Metzler (1971). "Mental Rotation of Three-Dimensional Objects". In: *Science, New Series* 171.3972, pp. 701–703.

Shin, Youhyun, Yeonchan Ahn, Hyuntak Kim, and Sang-goo Lee (2015). "Exploiting synonymy to measure semantic similarity of sentences". In: *Proceedings of the 9th International Conference on Ubiquitous Information Management and Communication, IMCOM 2015, Bali, Indonesia, January 08 - 10, 2015*. Ed. by Dongsoo S. Kim, Sang-Wook Kim, Suk-Han Lee, Lajos Hanzo, and Roslan Ismail. ACM, 40:1–40:4. URL: https://doi.org/10.1145/2701126.2701219.

Simon, Herbert A. (1996). *Models of My Life*. MIT Press.

Stokols, Daniel (2006). "Toward a Science of Transdisciplinary Action Research". In: *American Journal of Community Psychology* 38, pp. 63–77.

Suchanek, Fabian M, Gjergji Kasneci, and Gerhard Weikum (2007). "Yago: a core of semantic knowledge". In: *Proceedings of the 16th international conference on World Wide Web*. ACM, pp. 697–706.

Tai, Kai Sheng, Richard Socher, and Christopher D. Manning (2015). "Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks". In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*. The Association for Computer Linguistics, pp. 1556–1566. URL: http://aclweb.org/anthology/P/P15/P15-1150.pdf.

Tian, Junfeng, Zhiheng Zhou, Man Lan, and Yuanbin Wu (2017). "ECNU at SemEval-2017 Task 1: Leverage Kernel-based Traditional NLP features and Neural Networks to Build a Universal Model for Multilingual and Cross-lingual Semantic Textual Similarity". In: *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pp. 191–197.

Tlauka, Michael (2006). "Orientation dependent mental representations following real-world navigation". In: *Scandinavian Journal of Psychology* 47, pp. 171–176.

Trask, Andrew, Phil Michalak, and John Liu (2015). "sense2vec - A Fast and Accurate Method for Word Sense Disambiguation In Neural Word Embeddings". In: *CoRR* abs/1511.06388. URL: http://arxiv.org/abs/1511.06388.

Tsatsaronis, George, Iraklis Varlamis, and Michalis Vazirgiannis (2010). "Text Relatedness Based on a Word Thesaurus". In: *Journal of Artificial Intelligence Research (JAIR)* 37, pp. 1–39. URL: https://doi.org/10.1613/jair.2880.

Wieting, John and Kevin Gimpel (2017). "Revisiting Recurrent Networks for Paraphrastic Sentence Embeddings". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*. Ed. by Regina Barzilay and Min-Yen Kan. Association for Computational Linguistics (ACL), pp. 2078–2088. URL: https://doi.org/10.18653/v1/P17-1190.

Wieting, John, Mohit Bansal, Kevin Gimpel, and Karen Livescu (2015). "From Paraphrase Database to Compositional Paraphrase Model and Back". In: *Transactions of the Association for Computational Linguistics TACL* 3, pp. 345–358. URL: https://tacl2013.cs.columbia.edu/ojs/index.php/tacl/article/view/571.

– (2016). "Charagram: Embedding Words and Sentences via Character n-grams". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*. Ed. by Jian Su, Xavier Carreras, and Kevin Duh. The Association for Computational Linguistics (ACL), pp. 1504–1515. URL: http://aclweb.org/anthology/D/D16/D16-1157.pdf.

Wilcoxon, Frank (1945). "Individual comparisons by ranking methods". In: *Biometrics bulletin* 1.6, pp. 80–83.

Wood, David, Marsha Zaidman, Luke Ruth, and Michael Hausenblas (2014). *Linked Data*. Manning Publications Co.

Wu, Hao, Heyan Huang, Ping Jian, Yuhang Guo, and Chao Su (2017). "BIT at SemEval-2017 Task 1: Using semantic information space to evaluate semantic textual similarity". In: *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pp. 77–84.

Yang, Yinfei, Steve Yuan, Daniel Cer, Sheng-yi Kong, Noah Constant, Petr Pilar, Heming Ge, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil (2018). "Learning Semantic Textual Similarity from Conversations". In: *Proceedings of The Third Workshop on Representation Learning for NLP, Rep4NLP@ACL 2018, Melbourne, Australia, July 20, 2018*. Ed. by Isabelle Augenstein, Kris Cao, He He, Felix Hill, Spandana Gella, Jamie Kiros, Hongyuan Mei, and Dipendra Misra. Association for Computational Linguistics (ACL), pp. 164–174. URL: https://aclanthology.info/papers/W18-3022/w18-3022.

Zezula, Pavel, Giuseppe Amato, Vlastislav Dohnal, and Michal Batko (2006). *Similarity search: the metric space approach*. Vol. 32. Springer Science & Business Media.

Zhao, Han, Zhengdong Lu, and Pascal Poupart (2015). "Self-Adaptive Hierarchical Sentence Model". In: *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*. Ed. by Qiang Yang and Michael Wooldridge. AAAI Press, pp. 4069–4076. URL: `http://ijcai.org/Abstract/15/571`.

Zhou, Ping, Jiayin Wei, and Yongbin Qin (2013). "A Semi-Supervised Text Clustering Algorithm with Word Distribution Weights". In: *2013 the International Conference on Education Technology and Information System (ICETIS 2013)*. Atlantis Press, pp. 1024–1028.

Ziegler, Cai-Nicolas, Sean M. McNee, Joseph A. Konstan, and Georg Lausen (2005). "Improving recommendation lists through topic diversification". In: *Proceedings of the 14th international conference on World Wide Web, WWW 2005, Chiba, Japan, May 10-14, 2005*. Ed. by Allan Ellis and Tatsuya Hagino. ACM, pp. 22–32.

Zighed, Djamel A, Stéphane Lallich, and Fabrice Muhlenbach (2005). "A statistical approach to class separability". In: *Applied Stochastic Models in Business and Industry* 21.2, pp. 187–197.