PhD-FSTC-2019-13
The Faculty of Sciences, Technology and Communication

University of Bologna
Law School

# DISSERTATION

Defence held on 28/03/2019 in Bologna, Italy

to obtain the degree of

# DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG

# EN INFORMATIQUE

by

Rohan NANDA
Born on 17 September 1990 in Merrut Cantt UP (India)

# AUTOMATED IDENTIFICATION OF NATIONAL IMPLEMENTATIONS OF EUROPEAN UNION DIRECTIVES WITH MULTILINGUAL INFORMATION RETRIEVAL BASED ON SEMANTIC TEXTUAL SIMILARITY

## Dissertation defence committee

Dr Erich Schweighofer, Chairman
*Professor, University of Vienna*

Dr Paulo Quaresma, Member
*Professor, University of Evora*

Dr Leon van der Torre, Dissertation Supervisor
*Professor, University of Luxembourg*

Dr Guido Boella, Dissertation Supervisor
*Professor, University of Turin*

Dr Francesco Guerra, Member
*Professor, University of Modena and Reggio-Emilia*

Dr Luigi Di Caro, Member
*Professor, University of Turin*

## ABSTRACT

The effective transposition of European Union (EU) directives into Member States is important to achieve the policy goals defined in the Treaties and secondary legislation. National Implementing Measures (NIMs) are the legal texts officially adopted by the Member States to transpose the provisions of an EU directive. The measures undertaken by the Commission to monitor NIMs are time-consuming and expensive, as they resort to manual conformity checking studies and legal analysis. In this thesis, we developed a legal information retrieval system using semantic textual similarity techniques to automatically identify the transposition of EU directives into the national law at a fine-grained provision level. We modeled and developed various text similarity approaches such as lexical, semantic, knowledge-based, embeddings-based and concept-based methods. The text similarity systems utilized both textual features (tokens, N-grams, topic models, word and paragraph embeddings) and semantic knowledge from external knowledge bases (EuroVoc[1], IATE[2] and Babelfy[3]) to identify transpositions. This thesis work also involved the development of a multilingual corpus of 43 directives and their corresponding NIMs from Ireland (English legislation), Italy (Italian legislation) and Luxembourg (French legislation) to validate the text similarity based information retrieval system. A gold standard mapping (prepared by two legal researchers) between directive articles and NIM provisions was prepared to evaluate the various text similarity models. The results show that the lexical and semantic text similarity techniques were more effective in identifying transpositions as compared to the embeddings-based techniques. We also observed that the unsupervised text similarity techniques had the best performance in case of the Luxembourg Directive-NIM corpus.

We also developed a concept recognition system based on conditional random fields (CRFs) to identify concepts in European directives and national legislation. The results indicate that the concept recognitions system improved over the dictionary lookup program by tagging the concepts which were missed by dictionary lookup. The concept recognition system was extended to develop a concept-based text similarity system using word-sense disambiguation and dictionary concepts. The performance of the concept-based text similarity measure was competitive with the best performing text similarity measure. The labeled corpus of 43 directives and their corresponding NIMs was utilized to develop supervised text similarity systems


1 http://eurovoc.europa.eu/
2 http://iate.europa.eu
3 http://babelfy.org/

by using machine learning classifiers. We modeled three machine learning classifiers with different textual features to identify transpositions. The results show that support vector machines (SVMs) with term frequency-inverse document frequency (TF-IDF) features had the best overall performance over the multilingual corpus. Among the unsupervised models, the best performance was achieved by TF-IDF Cosine similarity model with macro average F-score of 0.8817, 0.7771 and 0.6997 for the Luxembourg, Italian and Irish corpus respectively. These results demonstrate that the system was able to identify transpositions in different national jurisdictions with a good performance. Thus, it has the potential to be useful as a support tool for legal practitioners and Commission officials involved in the transposition monitoring process.

**Keywords** : Text similarity, Transposition, European Law, Machine Learning, Concept Recognition

# ACKNOWLEDGMENTS

I would like to thank the Joint International Doctoral (Ph.D.) Degree in Law, Science and Technology and the Education, Audiovisual and Culture Executive Agency of the European Commission for providing funding to carry out this doctoral thesis. I would like to express my immense gratitude to Prof. Luigi Di Caro and Prof. Guido Boella for their excellent guidance and unconditional support during my PhD research. I am thankful to Prof. Leon van der Torre and Prof. Francesco Costamagna for their valuable guidance and feedback over the last three years. A very special thanks to Prof. Monica Palmirani for providing all the facilities for both administrative and research purposes. I would like to thank Prof. Martin Theobald for his feedback and suggestions for the improvement of this work. A special thanks to Prof. Antoni Roig and Prof. Pablo Noriega for their research and administrative support for my research stay in UAB Barcelona. I would also like to thank Dr. Livio Robaldo for his immense support for my academic, research and administrative well being.

A very special thanks to our collaborators from APIS Bulgaria, Hristo Konstantinov, Tenyo Tyankov, Daniel Traykov and Hristo Hristov. I would like to thank Llio Humphreys for her useful advice and suggestions for improving the ICAIL paper. I would like to thank Prof. Corrado Roversi for the interesting discussions during my stay at CIRSFID, Bologna. I would also like to thank Giovanni Siragusa for his collaboration. A very special thanks to Dina Ferrari and Virginie Mucciante for their immense support and help for adminsitrative issues.

I would like to thank a few special friends who were always there for me: Khoi, Stefanos, Silvestro, Vitor, Yukai, Iqbal, Arianna, Narine, Sara and Izarne.

Torino, July 20, 2018

Rohan Nanda

# CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# 1 INTRODUCTION

## 1.1 INTRODUCTION

The effective application of European Union (EU) Law at the national level is important to achieve the objectives of the Treaties and smooth functioning of the EU. Member States are responsible for the correct and timely implementation of EU law. The European Commission (EC) is responsible for monitoring the national implementations to ensure their compliance with EU law. The Commission also has the responsibility to examine the application of EU law under the control of the Court of Justice of the European Union (CJEU) [35].

Among the three major EU legal instruments, we are interested to study the transposition of directives into the national law. This is because directives are not directly applicable and Member States need to pass legislations to implement them into national law. Regulations are directly applicable in Member States and do not require transposition into national law. Decisions are binding only to those to whom they are addressed. Directives are binding as per the results to be achieved, but they provide national legislators of each Member State some discretion in the choice of methods and forms for implementation. A directive comes into effect only after it has been transposed into national law by the Member States [112]. Transposition is therefore quite important for effective implementation of EU policies across the Member States. The transposition of directives is also a legal duty of Member States as per the Article 288 of the Treaty on the Functioning of the European Union (TFEU). Delayed or incorrect transposition of directives hinder the EU policy objectives and the potential benefits they bring with them for European citizens [35]. Each directive is associated with a deadline by which Member States must implement national transposition measures which take into account the obligations of the Directive. In this thesis, we will refer these transposition measures as national implementing measures (NIMs).

Member States send the texts of NIMs to the Commission. The Commission then examines these texts to ensure that Member States have taken appropriate measures to achieve the objectives of the directive. The main goal of the Commission is to ensure that the NIMs are compliant with the directive. The Commission outsources the monitoring of NIMs to subcontractors and legal consulting firms. For instance, Milieu, a legal consultancy firm based in Brussels has been carrying out conformity checking studies of NIMs in different Member States since 2003 to study the transposition of several directives [27]. These

studies carried out by a team of competent legal experts, comprise legal analysis and concordance tables for studying the transposition of directive into the national law. Specifically, the concordance tables identify the specific provisions of NIMs which implement a particular article of the directive. Each row represents the transposition of a particular section of a directive into a specific provision of the NIM.

These legal measures undertaken by the Commission to monitor NIMs are time consuming and expensive [25]. For instance, to make a concordance table lawyers need to read several NIMs for each directive and then understand which provision of a particular NIM implements a particular article of the directive. This becomes more cumbersome for the Commission and lawyers doing cross-border or comparative legal research because they need to study the transposition in several Member States. Therefore, it is quite challenging for the Commission to monitor the application of EU Directives in the Member States. The EUR-Lex portal provides a list of NIMs adopted by the Member States and notified to the Commission. However, this provides only an outline of the intersection between European and national legislation. The list of NIMs do not provide a detailed understanding of the semantic correspondence between directives and NIMs at provision level. The identification of the transposed provisions is crucial for legal professionals and Commission officials to evaluate whether the obligations of the directive have been correctly transposed or not.

## 1.2 RESEARCH OBJECTIVES

There is clearly a need for a technological approach, which utilizes text mining and natural language processing (NLP) techniques, to assist the Commission and legal professionals in studying and evaluating the transposition of directives at a fine-grained provision level. The main objective of this approach is to develop an information retrieval system for identifying the relevant provisions of NIMs for a particular article of the Directive. This would make the process of monitoring NIMs much smoother and faster by supporting the manual work of identifying transpositions. It would assist lawyers and Commission officials to study and evaluate the transposition of directives more efficiently. To the best of our knowledge, this is the first approach which uses natural language processing and machine learning techniques to identify the transposition of EU directives into the national law of Member States. The major objectives of our research work are described as follows:

- Development of legal information retrieval systems based on text similarity approaches for identification of NIM provisions which transpose a particular article of the directive.

- Investigation and evaluation of various text similarity approaches with different textual features for the identification of transpositions.

The major research question addressed in this thesis is:

- How can text similarity approaches be used to identify the transposition of EU directives into national law ?

The usage of text similarity techniques to identify transpositions is based on the hypothesis that transposed NIM provisions and directive articles are semantically similar. This thesis work investigates whether semantic similarity can be a good indicator of transposition. The text similarity models and their outcomes provide a reasonable answer to this question. The thesis also explores the following research question:

- How do the text similarity techniques based on various semantic textual representations perform for identifying transpositions?

In this thesis, we develop different text similarity techniques by utilizing different semantic representations such as vector space model with term frequency-inverse document frequency (TF-IDF), latent semantic analysis (LSA), latent dirichlet allocation (LDA), paragraph vectors and vectors based on bag-of-concepts. Each semantic representation captures different textual features for identifying transpositions. Therefore, we evaluate different text similarity techniques on a multilingual corpus of directives and NIMs to compare their performance. This question is particularly answered in the results and analysis sections 2.3, 3.7 and 3.6.

## 1.3 STEPS TAKEN BY COMMISSION TO CONTROL THE IMPLEMENTATION OF DIRECTIVES

In this section, we will discuss the steps taken by Commission to control the implementation of directives. We have identified the following three steps as shown in Figure 1.1:

1. Steps taken by Commission to ensure effective transposition
2. Steps taken by Commission to monitor NIMs
3. Pre-infringement and Infringement steps taken by Commission

Each directive has a deadline by which Member States must adopt NIMs to transpose it into national law. To address the issues of implementation and enforcement of EU law, the Commission has developed a policy to promote compliance measures for effective transposition of directives. These measures are discussed in section 1.3.1. After adopting the transposition measures, Member States send the text

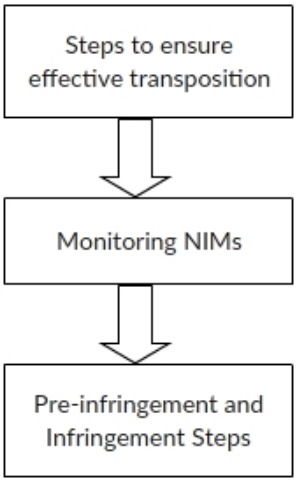


Figure 1.1: Steps taken by Commission to Control the Implementation of Directives

of NIMs to the Commission. The Commission then monitors the NIMs to ensure that they comply with the obligations of the directive. These monitoring measures are discussed in section 1.3.2. In case of non-compliance or breach of EU law, the Commission resorts to pre-infringement or infringement proceedings as discussed in Section 1.3.3 [35].

### 1.3.1 *Steps taken by EC to ensure effective transposition*

The Commission plays a pro-active role in order to make transposition process smooth and effective for the Member States. Some of the major steps taken by the Commission are discussed in this section.

**Guidelines:** Guidelines are non-legally binding documents which clarify the Commission's stand on the implementation and interpretation requirements of specific provisions of directives'. Guidelines are also specified by the Commission when the implementations of specific provisions of a directive vary quite a lot across Member States [35].

**Networks:** The Commission facilitates informal groups, called Networks which comprise representatives from Member States responsible for implementation of specific EU laws. Networks intend to achieve correct implementation of directives by enhancing cooperation between representatives of Commission and Member States.

**Implementation Plans:** The Commission prepares Transposition and Implementation Plans (TIPs) to assist Member States in transposing and implementing the directive. These plans identify the risks to correct and timely transposition of the directive and also provide appropriate measures to counter those risks. Member states are obliged

to plan appropriate measures to achieve the objectives of the directive and communicate them to Commission [35].

Other methods used by Commission to ensure effective transposition include inspection, fitness checks and legal review.

### 1.3.2 *Steps taken by EC to monitor NIMs*

After Member States have adopted NIMs, the Commission starts monitoring them to ensure the correct transposition of the directive [35]. Some of the steps taken by the Commission to monitor NIMs are discussed in this section.

**Correlation Tables:** Correlation tables identify the specific provisions of NIMs for each article of a directive. Correlation tables are made by Member States to ensure that the directive is fully transposed. These tables also assist the Commission to monitor the transposition of each provision of the directive. Correlation tables are generally not available to public as they are sent by Member States to the Commission as part of a confidential bilateral exchange. There is no agreed format or compulsory content for correlation tables. Therefore some Member States may provide detailed tables while others may just submit a list of some provisions [35].

**Conformity Checking:** The conformity checking reports are prepared by legal consulting firms such as, Milieu (based in Brussels). They have been discussed in the third paragraph of section 1.1. They provide a thorough legal analysis about the overview of the NIM provisions for a particular directive. They also provide a discussion about whether the implementing measures undertaken by the Commission correctly and completely transpose the directive or not. The conformity checking reports also provide details about the practical implementation of transposing measures at the national level by the authorities and the courts of the Member State. However, the presence of the concordance tables in these reports is most interesting for the scope of the thesis. A concordance table presents in a tabular format, the implementing provisions of NIM for a particular article or subarticle of the directive. They also provide a compliance assessment report about the implementation of the directive article by mentioning the associated comments or problems. A sample concordance table report [1] is shown in Figure 1.2.

### 1.3.3 *Pre-infringement and Infringement steps taken by Commission*

Despite the steps taken by Commission to ensure effective transposition of Directives there are several instances of non-compliance. The Commission itself acknowledges that the compliance of EU law in

1 https://publications.europa.eu/en/publication-detail/-/publication/eb746bfc-fa66-445a-9ec2-e135ccdc80a5/language-en

| Articles 16(1) and 16(2) | National provision (legal ref. & art.) | Text of national provision (in language of Member State) | Compliance assessment | Comments/Problems |
|---|---|---|---|---|
| **Article 16(1)** | | | | |
| Member States shall not make access to or exercise of a service activity in their territory subject to compliance with any requirements which do not respect the following principles:<br><br>(a) non-discrimination: the requirement may be neither directly nor indirectly discriminatory with regard to nationality or, in the case of legal persons, with regard to the Member State in which they are established; | European Union (Provision of Services) Regulations 2010, Regulation 6 (1) and 6 (2) (a) | Subject to Regulation 7 and Regulation 8, a relevant competent authority in the State may impose on a provider established in another Member State a requirement restricting the provider's freedom to provide a service in the State only if the requirement is non-discriminatory, necessary and proportionate.<br>(a) a requirement is non-discriminatory if it is neither directly nor indirectly discriminatory with regard to nationality or, in the case of a provider who is a legal person, with regard to the Member State in which the provider is established. | Yes | Note that Regulation 7 of the 2010 Regulations transposes Article 17 of the Services Directive, which lists a sizeable number of derogations (including services of general economic interest) from the principle of freedom to provide services.<br>Regulation 8 of the 2010 Regulations is virtually a mirror image of Article 18 of the Services Directive (which permits case-by-case derogations in respect of measures relating to the safety of services).<br>The Irish Regulations follow the wording of the Directive almost *verbatim*; therefore, there are no conformity problems as regards these Regulations. |
| (b) necessity: the requirement must be justified for reasons of public policy, public security, public health or the protection of the environment; | European Union (Provision of Services Regulations 2010, Regulation 6 (2)(b) | (b) a requirement is necessary if it is justified for reasons of public policy, public security, public health or the protection of the environment; | Yes | The Irish Regulations follow the wording used in the Directive almost *verbatim*; there are no conformity problems as regards these Regulations. |
| (c) proportionality: the requirement must be suitable for attaining the objective pursued, and must not go beyond what is necessary to attain that objective. | European Union (Provision of Services) Regulations 2010, Regulation 6 (2)(c) | (c) a requirement is proportionate if it is necessary for attaining the objective pursued and does not exceed what is necessary to attain that objective. | Ambiguous | Although the Service Directive refers to a requirement being "suitable", the Irish Regulations refer to a requirement being "necessary". There may, however, be a significant difference in the meaning of "suitable" and "necessary". There may be several "suitable" measures to achieve a goal (i.e. measures which, if adopted, can achieve the objective pursued), but not all of them will be necessary (i.e. essential for the attainment of the goal). Therefore, since the two terms do not have the exact same meaning, the transposition might be considered to be ambiguous. |

Figure 1.2: A sample concordance table for Ireland from the conformity checking report

Member States is still an unresolved issue [35]. The instances of non-compliance are identified while monitoring of NIMs (by Commission) and also by complaints from public, businesses and petitions from the European parliament [35].

**Pre-infringement Steps:** When the Commission detects a possible infringement of the EU directive, it resorts to pre-infringement tools to achieve out-of-court settlements by establishing a partnership with the Member States. EU Pilot is a pre-infringement tool developed by the Commission for resolving issues of non-compliance of EU law by carrying out informal bilateral discussion with the Member States. Any natural or legal person can lodge a complaint with the EU Pilot system against a Member State for any measure or practice not compatible with the EU law [35]. The complaints may also be registered upon Commission's own initiative.

**Infringement Steps:** If the pre-infringement EU pilot dialogue is unsuccessful, the Commission may launch formal infringement proceedings against the Member State under Article 258 of the Treat on the Functioning of the European Union (TFEU) [35] . The infringement may be launched against Member States under the following three conditions: failure to notify NIMs to Commission on time; non-conformity or non-compliance of national legislation with the requirements of EU directive; incorrect or no application of the directive. The Commission may start the litigation procedure by bringing the case to the Court of Justice of the European Union (CJEU).

## 1.4 USE CASES FOR AUTOMATED IDENTIFICATION OF NATIONAL IMPLEMENTING MEASURES

We provide two use cases where our system would assist legal practitioners and Commission by automatically identifying transpositions:

- Single jurisdiction legal research: A lawyer would like to see how Article $A_i$ of Directive $D$ is transposed in Member State $X$. In this case, the system retrieves the relevant NIM provisions (which transpose Article $A_i$ of Directive D) from Member State $X$. This is achieved by computing the similarity between directive articles and NIM provisions in the same language.

- Cross-border legal research: A lawyer would like to see how an Article $A_i$ of Directive $D$ is transposed in Member States $X$, $Y$, $Z$. In this case the system retrieves the relevant provisions of NIMs from each Member State by comparing directives and NIMs in the same language. This is achieved by using EU directives in the same language as the national language of the NIM and then computing the similarity between their articles and provisions.

## 1.5 RESEARCH CONTRIBUTION

The major research contributions of this thesis work are:

- This thesis presented the first work on automated identification of national implementation of European directives at provision level granularity by using text similarity approaches. A comprehensive set of text similarity techniques based on different semantic representations and textual features were developed for identifying transpositions. A multilingual corpus of 43 directives and their corresponding NIMs from Ireland, Italy and Luxembourg was prepared for validating the results of the text similarity techniques.

- A concept recognition system based on conditional random fields (CRFs) to identify and align concepts in European directives and national legislation. The system was extended to develop a concept-based text similarity system by using word-sense disambiguation and dictionary lookup.

The text similarity techniques were evaluated over the multilingual corpus of 43 directives and their corresponding NIMs. Our results show that the text similarity techniques are quite effective in identifying transpositions. The best macro average F-score values for TF-IDF based Cosine similarity measure were 0.8817, 0.7771 and 0.6997 for Luxembourg, Italy and Ireland legislation. These results indicate that the system has the potential to be used as a support tool in the transposition monitoring process (section 1.3.2) by assisting in the manual task of identifying transpositions.

## 1.6 THESIS OUTLINE

The rest of the thesis report is organized as follows. **Chapter 2** discusses the investigation of unsupervised lexical and semantic similarity techniques to identify transpositions. These techniques include cosine similarity based on vector space model, latent semantic analysis (LSA) and unifying similarity measure (USM). Knowledge-based similarity measures were also developed for both cosine similarity and LSA. The different similarity measures were first evaluated on a small corpus and compared with state-of-the-art methods. A more thorough evaluation was carried out on a multilingual corpus of 43 directives and their corresponding NIMs.

**Chapter 3** presents the unsupervised similarity techniques based on word and paragraph embeddings learned by shallow neural networks. We trained word embeddings on a legal corpus comprising both European directives and national legislations for three different languages (national legislations from Ireland, Luxembourg and

Italy). The trained word embeddings were utilized to develop provision vectors for directives and NIMs to compute text similarity. We also developed paragraph vector models which learn a dense vector representation for texts of different length. A discussion about the performance of different similarity techniques on the multilingual corpus was also presented in this chapter.

**Chapter 4** presents a machine learning based concept recognition system to identify concepts in European and national legislation. We utilized a corpus of directives and national legislation in English. A dictionary lookup program was developed to generate an annotated corpus with concepts from the IATE (Inter-Active Terminology for Europe) vocabulary. A few named entities were tagged by using a state-of-the-art named entity recognizer. A conditional random fields (CRF) classifier was trained on the annotated data. The system was able to identify concepts in both European and national legislation with decent performance. The advantage of using the CRF over dictionary lookup programs was illustrated by some examples. Further, the concept recognition system was used to develop a text similarity model by integrating word-sense disambiguation and named entity linking from Babelfy. This chapter also presents the results of the concept-based similarity system on the multilingual corpus.

**Chapter 5** describes the supervised text similarity models to identify transposition of European directives. We utilize the gold standard mapping from the legal annotators to develop a machine learning based text similarity model. The system is modeled as a binary classifier. Given a directive article and a NIM provision as an input, it classifies the pair as similar or not similar. The model utilizes the textual features such as TF-IDF, latent semantic analysis (LSA) and latent dirichlet allocation (LDA) from both directive and NIM provisions. We test our approach by implementing different machine learning classifiers such as Naive Bayes, Support Vector Machines (SVMs), Logistic regression and an Ensemble voting classifier. The models were evaluated with different feature sets for all three languages.

**Chapter 6** presents the literature review of related research works which automate or semi-automate certain legal tasks by means of text similarity, machine learning and concept-based techniques. We identify different legal-tech use cases and approaches for each domain.

**Chapter 7** presents the conclusion and future work.

## 1.7 PUBLICATIONS

This thesis work builds upon the papers co-authored over the past three years. Parts of this thesis have appeared in the following publications:

- Rohan Nanda, Luigi Di Caro and Guido Boella. **A Text Similarity Approach for Automated Transposition Detection of**

**European Union Directives**. In Proceedings of the 29th International Conference on Legal Knowledge and Information Systems (JURIX), Pages:143-148, Volume 294, December 2016, Sophia Antipolis, France.

This paper establishes the need for automated identification of national implementing measures and motivations behind this work [88]. It presents the investigation of lexical, semantic and knowledge-based similarity techniques on the English corpus of five directives. This work has been elaborated in Chapter 1 and Chapter 2 of this thesis.

- Rohan Nanda, Luigi Di Caro, Guido Boella, Hristo Konstantinov, Tenyo Tyankov, Daniel Traykov, Hristo Hristov, Francesco Costamagna, Llio Humphreys, Livio Robaldo and Michele Romano. **A Unifying Similarity Measure for Automated Identification of National Implementations of European Union Directives**. In Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law (ICAIL), ACM, Pages:149-158, June 2017, London, United Kingdom.

  This paper presented a unifying similarity measure (USM) for automated identification of national implementing measures [89]. USM utilized textual features such as common words, common sequences and partial string matches. The proposed similarity measure was evaluated on a multilingual corpus of four directives in English, French and Italian. This work was extended and the similarity measure was evaluated on a larger corpus of 43 directives and NIMs as discussed in Chapter 2 of the thesis.

- Rohan Nanda, Giovanni Siragusa, Luigi Di Caro, Martin Theobald, Guido Boella, Livio Robaldo and Francesco Costamagna. **Concept Recognition in European and National Law**. In Proceedings of the 30th International Conference on Legal Knowledge and Information Systems (JURIX), Pages:193-198, Volume 302, December 2017, Luxembourg.

  This paper presented a system to identify concepts in European directives and national legislation [90]. The system derived concepts from IATE (Inter-Active Terminology for Europe) and used conditional random fields (CRFs) to tag concepts in legislative texts. This work was extended to develop a semantic similarity measure based on IATE concepts and word-sense disambiguation (from Babelfy). This work is discussed in Chapter 4 of the thesis.

- Rohan Nanda, Giovanni Siragusa, Luigi Di Caro, Guido Boella, Lorenzo Grossio, Marco Gerbaudo and Francesco Costamagna. **Unsupervised and Supervised Text Similarity Systems for Automated Identification of National Implementing Measures of**

**European Directives**. In Journal of Artificial Intelligence and Law. October 2018.

This paper presented a thorough evaluation of various unsupervised and supervised text similarity techniques used to identify transpositions on a multilingual corpus of 43 directives and their corresponding NIMs from Ireland, Luxembourg and Italy [92]. New unsupervised text similarity techniques based on word and paragraph learned by shallow neural networks were introduced. Also supervised text similarity techniques based on machine learning classifiers were presented in this paper. This work has been elaborated mainly in Chapter 3 and Chapter 5. Section 2.3 of Chapter 2 also presents some results fo this paper.

- Rohan Nanda, Adebayo Kolawole John, Luigi Di Caro, Guido Boella and Livio Robaldo. **Legal Information Retrieval Using Topic Clustering and Neural Networks**. In Proceedings of COLIEE 2017. 4th Competition on Legal Information Extraction and Entailment at the International Conference on Articial Intelligence and Law (ICAIL) 2017.

  This paper presented a description about the methodology used in the information retrieval task in the Competition on Legal Information Extraction/ Entailment (COLIEE) 2017 task [91]. A text similarity system was developed using approximate string matching and topic clustering methods. The system was evaluated on the COLIEE question-answering dataset consisting of queries and Japanese civil law codes. This paper is relevant to the overall theme of the thesis.

# 2

# UNSUPERVISED LEXICAL AND SEMANTIC TEXT SIMILARITY MODELS

In this chapter, we present the unsupervised text similarity models to identify the transpositions of European directives. In section 2.1 , we utilize both lexical and semantic similarity techniques and supplement them with knowledge from EuroVoc[1] to identify transpositions. We then evaluate our approach by comparing the results with the correlation tables (gold standard) for five directives and their corresponding national implementing measures (NIMs) in English. Our results indicate that both similarity techniques proved to be effective in identifying transpositions.

In section 2.2, we present a unifying text similarity measure (USM) for automated identification of national implementations of European Union (EU) directives. USM incorporates methods for matching common words, common sequences of words and approximate string matching. It was used for identifying transpositions on a multilingual corpus of four directives and their corresponding national implementing measures (NIMs) in three different languages : English, French and Italian. We further utilized a corpus of four additional directives and their corresponding NIMs in English language for a thorough test of the USM approach. We evaluated the model by comparing our results with a gold standard consisting of official correlation tables (where available) or correspondences manually identified by legal experts. Our results indicate that USM was able to identify transpositions with average F-score values of 0.808, 0.736 and 0.708 for French, Italian and English Directive-NIM pairs respectively in the multilingual corpus. A comparison with state-of-the-art methods for text similarity illustrates that USM achieves a higher F-score and recall across both the corpora.

In section 2.3, we present a thorough investigation of the lexical and semantic text similarity models to identify transpositions on a large multilingual corpus of 43 directives and their corresponding NIMs from Ireland, Italy and Luxembourg. The results indicate that TF-IDF cosine outperformed other methods in terms of F-score. The performance of latent semantic analysis (LSA) and USM was comparable. We also observe that the Luxembourg Directive-NIM corpus achieved the highest F-score as compared to Ireland and Italian legislation.

1 http://eurovoc.europa.eu/

## 2.1 TF-IDF COSINE AND LATENT SEMANTIC ANALYSIS

In this section, we describe our approach for automated transposition identification of EU directives using unsupervised text similarity models, based on TF-IDF (term frequency-inverse document frequency) Cosine and latent semantic analysis (LSA). While transposing EU directives the Member States have a certain amount of discretion in the choice of methods. For example, in the UK, the national legislators generally use two broad approaches for transposing a directive. The first approach is called 'copy-out', where the NIM provision uses similar wording as that of the directive. In this approach, the NIM may also cross refer to the relevant directive provision. Due to this reason we chose TF-IDF cosine similarity approach to detect such kind of transpositions. In the second approach, called 'elaboration', the provisions of NIM use different language from the wordings of the directive. This is done to clarify the meaning of NIM provision for legal or domestic policy reasons [45]. We chose latent semantic analysis (LSA) due to its ability to extract semantics of terms by analyzing their usage in different documents. This might be helpful in detecting cases of 'elaboration' transposition.

We utilized a corpus of five directives and their corresponding NIMs in English (from Ireland or the United Kingdom). First of all, each group of directive and NIMs are stored in a format to adhere to the structure of their particular correlation table. This would enable us to compare our results with that of the correlation tables. This was carried out for each group of directive and NIM because correlation tables have no standard way of structuring the directives and NIMs. Sometimes they mention the transposed provisions for a complete article of the directive and sometimes only for a paragraph of the article. Thus, the articles and provisions of each Directive-NIM group are stored as a corpus. From here on the term provision refers to both article (of Directive) and provision (of NIM). Figure 2.1 presents the system architecture. Each article of a directive is compared with all the provisions of its corresponding NIMs to retrieve the most semantically similar provisions. The retrieved provisions are compared with the gold standard (correlation tables) to evaluate the performance of the system.

The next step is pre-processing of the data. This consists of a number of steps to remove noise from the text. It is important to select only suitable and relevant terms as the performance of information retrieval systems is dependent on these pre-processing methods. The punctuation was removed and the text was converted to lowercase. Then tokenization was carried out to extract single words from the text. The stop words were removed using NLTK's corpus of stopwords for English. We used NLTK's part-of-speech tagger (POS tagger) to filter out nouns, verbs and adjectives from the remaining set of tokens

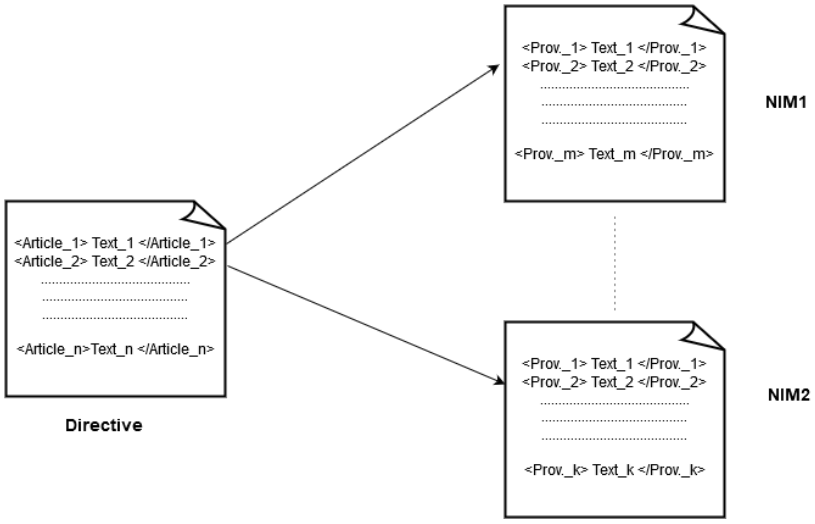


Figure 2.1: Articles of a directive are compared with NIM provisions to retrieve the most semantically similar provisions

[9]. The manual reading of several directive and NIM provisions suggested that nouns, verbs and adjectives contained the most informative features.

In the next step, the tokens obtained after pre-processing were enriched with the knowledge from EuroVoc[2], a multilingual thesaurus of the European Union. EuroVoc seemed to be an ideal choice because it not only contains documentary information of the EU institutions but also covers a wide range of fields representing both Community and national points of view. We utilized all the 127 microthesauri of EuroVoc to cover all possible domains for our corpus. We made use of the equivalence relationship between preferred and non-preferred terms in EuroVoc [97]. This relationship defines the indexing term when more than one terms represent the same concept. The EuroVoc microthesauri contains several instances of equivalence relationship. We used this relationship for synonyms and near-synonyms terms. For the implementation purpose, the 127 microthesauri were consolidated as a python dataframe. The list of tokens obtained after pre-processing was compared with each element of the consolidated EuroVoc microthesauri. In case of a match, the token in our corpus was replaced by the preferred indexing term as per equivalence relationship of EuroVoc. Afterwards, the set of new tokens are stemmed to reduce the inflectional forms of words.

### 2.1.1 *TF-IDF Cosine*

The output from the previous section is a bag-of-words representation. It is basically a list of each token and its count in a particular provi-

2 http://eurovoc.europa.eu/

sion. A provision-term matrix is then constructed with a collection of all provision vectors in the corpus. The rows of the matrix consist of the terms and the columns correspond to the provisions. This representation of documents or provisions as vectors in a common vector space is called as vector space model (VSM). We applied Term Frequency-Inverse Document Frequency (TF-IDF) weighting method to the provision-term matrix [111]. The TF-IDF measure evaluates the importance of each token, by offsetting its frequency in the provision with its frequency in the entire corpus. The TF-IDF weight of term $t$ in provision $p$ is given as follows:

$$tf - idf_{t,p} = (tf_{t,p}) \cdot \log \frac{N}{\mathrm{pf}_t} \tag{2.1}$$

where $tf_{t,p}$ is the term frequency of term $t$ in provision $P$, $N$ is the number of provisions in the corpus and $pf_t$ is the provision frequency of term $t$ in the corpus. The cosine similarity measure between article vector $A$ and provision vector $P$ is computed as follows:

$$CS(A, P) = \frac{A.P}{|A||P|} \tag{2.2}$$

The dot product of the article and provision vector is divided by the product of their lengths (lengths computed by Euclidean distance).

### 2.1.2 *Latent Semantic Analysis*

One of the major drawbacks of utilizing the vector space model (VSM) is its inability to deal with polysemy and synonymy. Synonymy implies that two different words can have the same meaning. VSM fails to model the relationship between synonymous terms. Therefore, it underestimates the true semantic similarity between a query and a document containing synonymous terms. Polysemy implies that a term has more than one meaning. In this case, the computed similarity between the query and the document containing the polysemic term would overestimate the actual semantic relevance [78].

Latent Semantic Analysis (LSA) is a popular indexing method in information retrieval which is used to produce a low-rank approximation matrix for the document-term matrix (provision-term matrix in our case) by using word co-occurrence [30]. The derived features of LSA have been shown to capture polysemy and synonymy to some extent [30]. LSA uses singular value decomposition (SVD) to project the provision vectors into a reduced latent space [43]. SVD decomposes the provision-term matrix into separate matrices which capture the similarity between terms and provisions across different dimensions in space. The relationship between terms is represented in a subspace approximation of the original vector space to reduce noise and find latent relations between terms and documents. The original provision-term matrix $X$ is reduced to a lower rank approximation matrix, $X_k$,

where the rank $k$ is much smaller than the original rank of matrix $X$. The approximation is represented as follows:

$$X_k = U\Sigma_k V^T \tag{2.3}$$

The $\sum$ matrix represents the singular values of $X$. $U$ and $V$ represent the left singular vector and right singular vector respectively. The truncated matrix $(V')^T$ represents the provisions in the reduced k-dimensional space. The query, $A_i$ (directive article) is also transformed into the LSA space as follows:

$$A_{ik} = \Sigma_k^{-1} U_k^T A_i \tag{2.4}$$

The cosine similarity values can be computed between the directive article and the corresponding NIM provisions to retrieve the most similar NIM provisions. We experimented with different number of latent dimensions on our dataset and the best performance was observed at 50 dimensions (chosen value for results).

Figure 2.2 presents the system architecture for TF-IDF Cosine and LSA for automated identification of national implementing measures (NIMs). The query (specific article of directive) is also transformed through the pre-processing and vector transforms. Since we want to evaluate the influence of adding knowledge from EuroVoc and also compare the performance of CS and LSA, we divide the evaluation into four cases : (i) TFIDF-Cosine (ii) TFIDF-Cosine with EuroVoc, (iii) Latent semantic analysis (LSA), (iv) LSA with EuroVoc. It is important to note that dotted block of EuroVoc in Figure 2.2 is considered only for case (ii) and (iv). Similarly, the dotted block of SVD is considered only for case (iii) and (iv). For case (i) and (ii), cosine similarity is calculated as cosine of the angle between the transformed query vector (in TF-IDF representation) and each provision vector in the corpus (also in TF-IDF representation). The matching NIM provisions with similarity values greater than or equal to the threshold value are retrieved by the system. Similarly, for case (iii) and (iv), the similarity is measured by the cosine of the angle between the query vector and each provision vector in the reduced-dimensional space.

### 2.1.3 *Results and Analysis*

In this section, we study the results of automated identification of national implementing measures of five directives using TF-IDF Cosine and LSA. The directives and NIMs used in this corpus for studying transposition are as follows (CELEX number of directives). Directive1: 32011L0085. NIM1 (NIM of Ireland transposing Directive1): European Union (Requirements for budgetary frameworks of Member

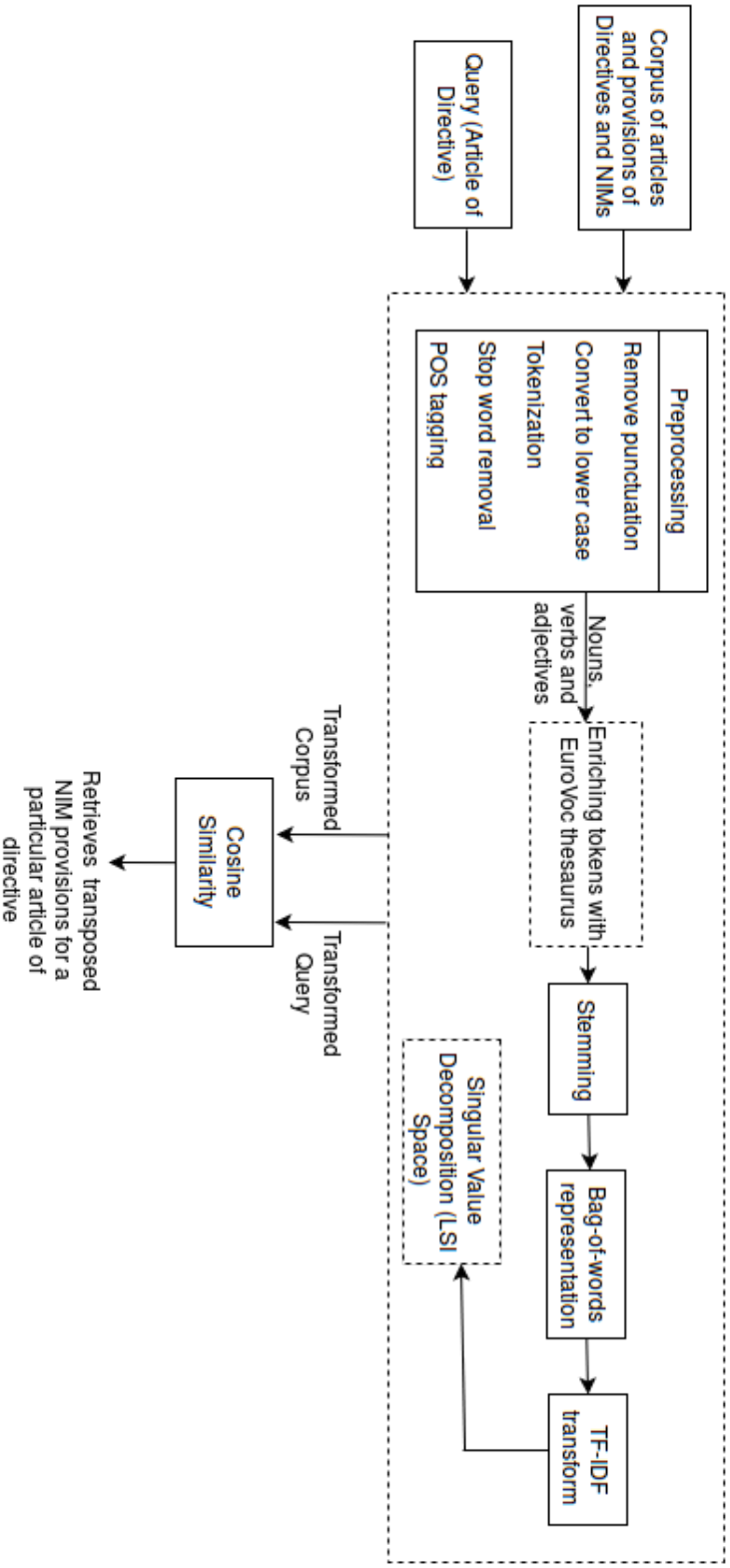


Figure 2.2: System architecture for automated identification of national implementing measures

States) Regulations 2013. Directive2: 32001L0024. NIM2 (NIM of Ireland transposing Directive2): European Communities (Reorganisation and Winding-Up of Credit Institutions) Regulations 2004. Directive3: 31999L0092. NIM3 (NIM of UK transposing Directive3): The Dangerous Substances and Explosive Atmospheres Regulations 2002. Directive4: 32003L0010. NIM4 (NIM of UK transposing Directive4): The Control of Noise at Work Regulations 2005. Directive5: 31998L0024. NIM5 (NIM of UK transposing Directive5): The Control of Substances Hazardous to Health Regulations 2002. NIM6 (NIM of UK transposing Directive5): The Control of Lead at Work Regulations 2002. NIM7 (NIM of UK transposing Directive5): Control of Asbestos at Work Regulations 2002. NIM8 (NIM of UK transposing Directive5): The Dangerous Substances and Explosive Atmospheres Regulations 2002.

The implementation was carried out in Python and utilized NLTK and Gensim libraries [100][71]. The system is evaluated by computing the standard metrics of precision, recall and F-score used in information retrieval. Precision measures how many identified transpositions are correct transpositions. Recall measures how many actual transpositions are identified by the system. In other words, precision evaluates the system on its ability to retrieve the correct transpositions. Recall evaluates the system on its ability to retrieve all the correct transpositions. If a system identifies only 1 (but correct) transposition out of 10 existing transpositions, then precision would be 1 (the identified transposition is correct). However, the recall would be 0.1 (1/10) because the system could retrieve only 1 transposition out of 10. The F-score combines both precision and recall into a single measure. It is computed as the harmonic mean of precision and recall. The values of precision and recall are computed by recording true positives (TP), false positive (FP), true negative (TN) and false negative (FN). The following equations represent the computation of precision (P), recall (R) and F-score (F):

$$P = \frac{TP}{TP + FP} \tag{2.5}$$

$$R = \frac{TP}{TP + FN} \tag{2.6}$$

$$F = \frac{2PR}{P + R} \tag{2.7}$$

We did not consider accuracy as we have very different number of true positives and true negatives resulting in an unbalanced dataset. In such cases, accuracy is not a fair metric for evaluation. We model and evaluate the system by considering these four cases: (i) TF-IDF Cosine, (ii) TF-IDF Cosine with EuroVoc, (iii) Latent semantic analysis (LSA),

(iv) Latent semantic analysis (LSA) with EuroVoc. Figure 2.3 (Results of Directive 1, Directive 2 and Directive 3) and Figure 2.4 (Results of Directive 4 and Directive 5) present the results of automated identification of NIMs by utilizing TF-IDF cosine and LSA for the five directives being considered. Directive1, Directive2, Directive3 and Directive4 are each transposed by 1 NIM. Directive 5 is transposed by 4 NIMs. Appropriate threshold levels for identifying transpositions for both TF-IDF Cosine and LSA were determined through experimentation on the dataset. A threshold of 0.35 and 0.40 was selected for TF-IDF Cosine and LSA, respectively for first four directives. For Directive5, the chosen threshold values were 0.25 (for TF-IDF Cosine) and 0.30 (for LSA). This is due to the fact that Directive 5 and its corresponding NIMs have much higher number of total provisions (Table 2.1) as compared to other four directives. The same thresholds are used when both CS and LSA models are supplemented with knowledge from EuroVoc as the length of dictionary of tokens was almost similar to the original one.

The results in Figures 2.3 and 2.4 indicate no clear winner in terms of performance. However, we do make a few interesting observations. In terms of F-Score, TF-IDF Cosine achieves the best performance across all 5 directives. The performance of LSA was similar to TF-IDF Cosine in Directive1 and Directive2. However, it was outperformed by TF-IDF Cosine in Directive3, Directive4 and Directive5. This is because, LSA has been shown to perform well when a large corpus is available to extract the latent relationships between different terms with same meaning in different documents. LSA needs a large corpus to derive the semantics of a word by analyzing its relationship with other words [28]. In a small corpus (like in our case), there is not enough text to extract the relationships between different words. Also the application of SVD causes some important features (needed for text similarity) to be lost, which results in higher false negatives (system is unable to detect the transposition, even though its present). This results in LSA systems achieving lower recall as compared to TF-IDF Cosine system (as recall depends on false negatives). The same is observed through the graphs of Figures 2.3 and 2.4. In Directive3, Directive 4 and Directive 5 the recall of LSA is always lower than TF-IDF Cosine due to these higher false negatives. In Directive1 and Directive2, TF-IDF Cosine has the same number of false negatives as LSA resulting in similar recall. The low recall of LSA systems is compensated by the higher precision due to the trade-off. The precision values of LSA were equal to or higher than TF-IDF Cosine in Directive1, Directive2, Directive3 and Directive5. However, the precision values of TF-IDF Cosine are quite close to LSA. Overall in terms of all three metrics TF-IDF Cosine has the best performance due to higher recall and F-score and decent precision in all the directives.

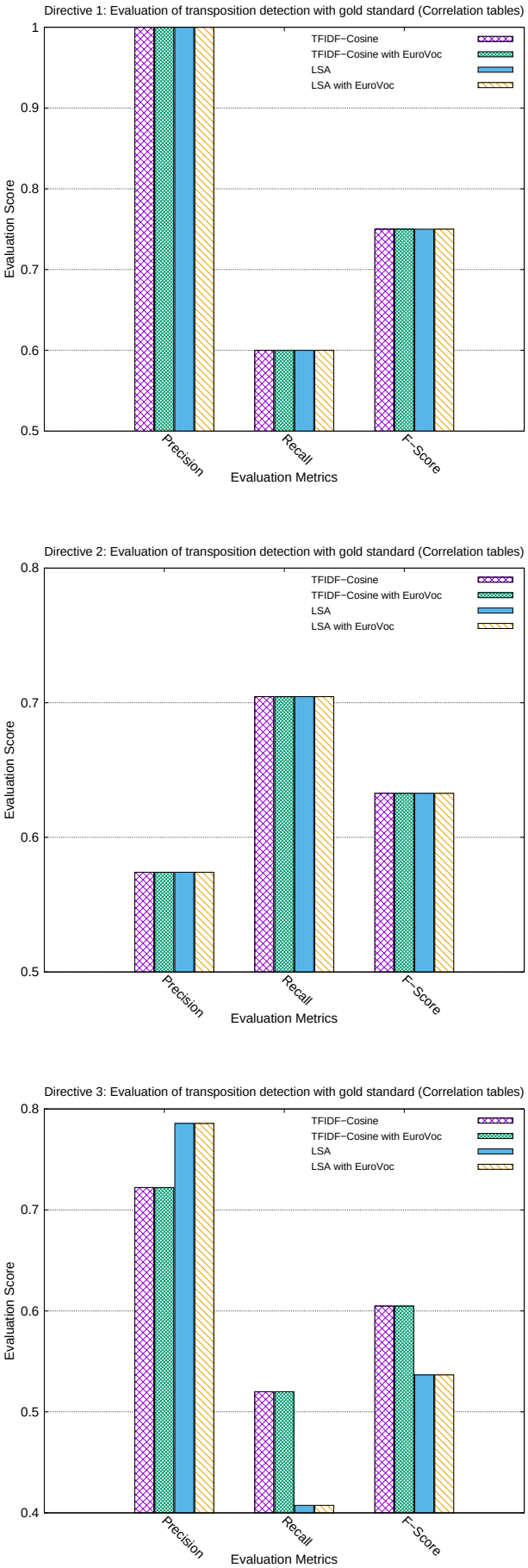


Figure 2.3: Evaluation of transposition identification for Directive1, Directive 2 and Directive 3

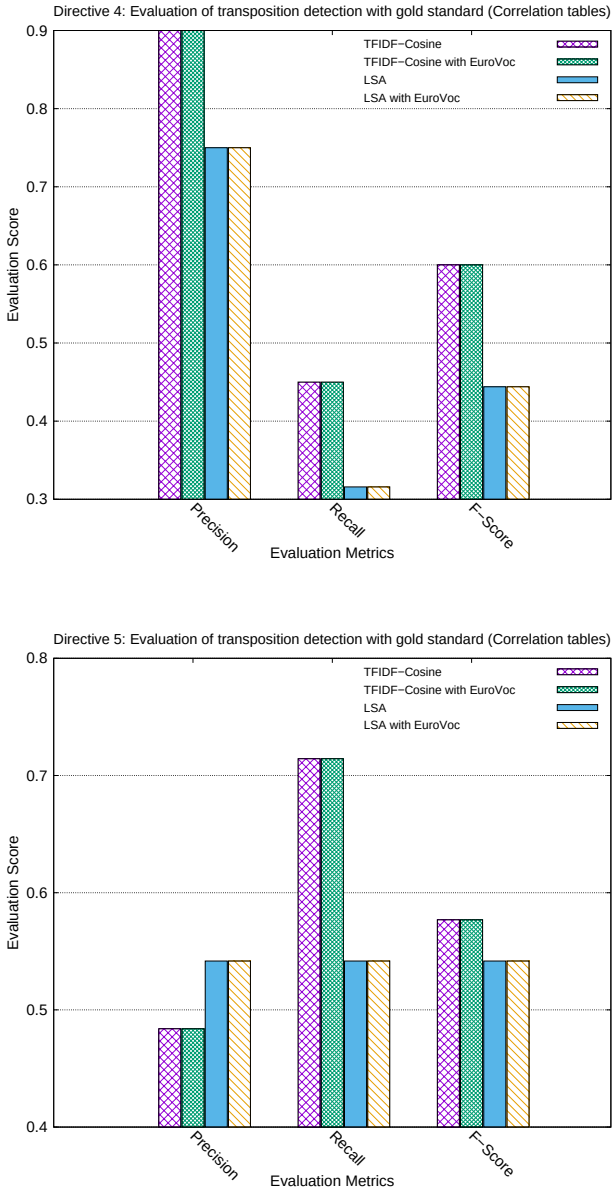


Figure 2.4: Evaluation of transposition identification for Directive 4 and Directive 5

In case of Directive5, there were several cases where an article was transposed by multiple provisions of different NIMs (NIM5, NIM6, NIM7 and NIM8). This made it pretty challenging for both TF-IDF Cosine and LSA to retrieve all the matching provisions from different NIMs, thus resulting in very few matches. Also the total number of NIM provisions in this case was 306 (Table 2.1), much higher than the other four cases.

Table 2.1: Statistics of Directives and NIMs under consideration

| Directive-NIM | Number of provisions of Directive | Number of provisions of NIM/s | Total number of provisions |
|---|---|---|---|
| (Directive1, NIM1) | 31 | 17 | 48 |
| (Directive2, NIM2) | 68 | 88 | 156 |
| (Directive3, NIM3) | 37 | 67 | 104 |
| (Directive4, NIM4) | 24 | 50 | 74 |
| (Directive5, NIM5, NIM6, NIM7, NIM8) | 41 | 306 | 347 |

We also observe from the results that the addition of knowledge from EuroVoc does not improve the performance of both TF-IDF Cosine and LSA. In most cases, the precision and recall values of TF-IDF Cosine with EuroVoc and LSA with EuroVoc are same as those of TF-IDF Cosine and LSA respectively. We found that in our corpus there were several provisions of both directives and NIMs where some terms were enriched from EuroVoc thesaurus. However, the terms added from EuroVoc to a particular article of a directive did not match any terms present in the transposing provision and vice versa. This is why the knowledge from EuroVoc does not help to improve the existing TF-IDF Cosine and LSA results. For instance, in an article of directive, the terms 'worker', 'evaluation','health' and 'risk' are replaced by EuroVoc terms, 'labour force', 'evaluation method', 'health policy' and 'insured risk' respectively. The provision transposing this article does not contain any of these EuroVoc terms. Also, in the provision, the terms 'employee' and 'evaluation' are replaced by 'wage earner' and 'evaluation method' respectively. The term 'evaluation' was already common between both article and provision. Its replacement by 'evaluation method' does not make any difference with regard to text similarity. The term 'wage earner' is not present in the article. Thus,

the knowledge from EuroVoc did not help to improve the results of text similarity in the present corpus of directives and NIMs.

In summary, our results indicate the fact that both TF-IDF Cosine and LSA similarity techniques are effective in identifying transposition of EU directives. There was no notable performance improvement by incorporating the knowledge from EuroVoC. The performance of the system is subject to the requirements of legal professionals studying the transposition. In majority of the cases, LSA achieves higher precision (except in Directive4). While TF-IDF Cosine always achieves higher recall (except Directive1 and Directive2, where they have same recall). In terms of F-score, TF-IDF Cosine outperforms LSA (except Directive1 and Directive2, where they have same F-score).

## 2.2 A UNIFYING TEXT SIMILARITY MEASURE (USM) FOR AUTOMATED IDENTIFICATION OF NATIONAL IMPLEMENTATIONS OF EUROPEAN UNION DIRECTIVES

### 2.2.1 *The Proposed Model*

In the previous section, we investigated the application of two state-of-the-art unsupervised text similarity techniques to identify transpositions. The results were satisfactory but needed improvement for the system to be useful in legal research. Therefore, in this section we propose, develop and evaluate a unifying text similarity measure (USM) for automated identification of transposed NIM provisions of EU directives in different Member States. The proposed model was used for identifying transpositions at a fine-grained provision level in a multilingual corpus of four directives and their corresponding NIMs across three different languages: English, French and Italian (for the national legislation of Ireland, United Kingdom, Luxembourg and Italy). We further utilized a corpus of four additional directives and their corresponding NIMs in English language for a thorough performance analysis of our model. We evaluated the model by comparing our results with a gold standard consisting of official correlation tables (where available) or correspondences manually identified by legal experts. Our results indicate that USM was able to identify transpositions with average F-score values of 0.808, 0.736 and 0.708 for French, Italian and English Directive-NIM pairs respectively in the multilingual corpus. It also achieved an average F-score of 0.712 on the second corpus (of four additional directives and their corresponding NIMs in English language).

Manual analysis of the directive articles and their corresponding NIM provisions provided the following observations:

1. The presence of common words and phrases in many articles and their corresponding NIM provisions.

2. The presence of common sequences of words in some articles and their corresponding NIM provisions.

3. NIM provisions rarely transpose the entire article of the directive. Either they go into more detail, or they partially transpose the article. In some of these cases, an article is transposed by two or more provisions.

(1)-(3) motivated us to develop a dedicated model for automated identification of NIM provisions. We define a similarity measure for each observation and then propose a unifying similarity measure (USM) to take into account (1)-(3). USM is proposed in order to benefit from the complementarity of different similarity measures and it would be useful to identify different kinds of transpositions which are not identified by a single similarity measure.

**Cosine similarity:** To address the first observation we utilize the cosine similarity measure as it has been shown to perform well in identifying semantically similar texts in the presence of common words and phrases [52]. The cosine similarity between the vectors of Article $A$ and provision $P$ is computed as follows:

$$C(A,P) = \frac{A.P}{|A||P|} \tag{2.8}$$

The numerator represents the dot product of the vectors. The denominator is the product of their lengths, given by the Euclidean distance. The effect of the document length is compensated by the denominator which normalizes the similarity value. The cosine similarity ranges from 0 to 1 as TF-IDF weights are non-negative.

**N-gram similarity:** The second observation is addressed by incorporating the N-gram similarity measure. N-gram models are useful in identifying transpositions in the presence of common sequences of words in articles and NIM provisions. This is because the N-gram model generates a contiguous sequence of words for a given text. The presence of shared N-grams in article and NIM provisions may imply transposition. For an Article $A$ and a NIM provision $P$, the N-gram similarity is defined as [5]:

$$N(A,P) = \frac{sharedgrams}{totalgrams} \tag{2.9}$$

Here, *sharedgrams* is the number of character N-grams shared by $A$ and $P$. *totalgrams* is the total number of N-grams present in both $A$ and $P$. However, another N-gram similarity metric is considered in order to compensate for the low similarity values of short strings by using a warp variable and computing the similarity as follows [5]:

$$N(A,P) = \frac{totalgrams^{warp} - unsharedgrams^{warp}}{totalgrams^{warp}} \tag{2.10}$$

where,

$$unsharedgrams = totalgrams - sharedgrams \tag{2.11}$$

The term *unsharedgrams* is the number of N-grams which are not shared by $A$ and $P$. The warp values are between 1 and 3. Since most provisions are short and precise texts, we used the warp to compute the N-gram similarity using Eq.2.10. We chose N-grams for N=4 as they provided the best results. The value of warp was chosen as 2 to moderately elevate the similarity values of short texts.

**Approximate String Matching:** The third observation is addressed by incorporating an approximate string matching algorithm. The two texts $A$ and $P$ are first tokenized. Each group of tokens in $A$ and $P$ is considered as a set [118]. Then the intersection set, $I$ of sorted tokens in $A$ and $P$ is computed as:

$$I = A \cap P \tag{2.12}$$

Set $A$ is then represented as the union of the tokens in the intersection set $I$ and the remaining tokens in the remainder article set $R_A$.

$$A = I \cup R_A \tag{2.13}$$

Similarly, the provision set $P$ is represented as the union of the intersection set $I$ and the remainder provision set $R_P$.

$$P = I \cup R_P \tag{2.14}$$

Now we compute three similarity measures for (I,A), (I,P) and (A,P). The similarity measure $AS$ between two sets is computed as $2.0 * M/T$, where T is the total number of elements in both sets and M is the number of matches [118]. The similarity is in the range of [0,1]. The maximum similarity value of the three is considered as the final output. The major significance of this method is that the intersection set $I$ is the same in both sets $A$ and $P$. $A$ and $P$ have high similarity values when set $I$ is the larger part of either $A$ or $P$.

**The Unifying Similarity Measure (USM):** We observed that the above three different similarity measures have their own unique way of estimating the similarity of two texts. These three measures were identified on the basis of the manual analysis of articles and corresponding NIM provisions. We propose a novel unifying similarity measure which benefits from the complementarity of the above three similarity measures. The major advantage of this measure is its ability to identify transpositions which were not identified previously with the use of a single similarity measure. The unifying similarity measure, *USM(A,P)* between $A$ and $P$ is defined as the weighted arithmetic mean of cosine similarity *CS(A,P)*, N-gram similarity *N(A,P)* and approximate similarity *AS(A,P)* as follows:

$$USM(A,P) = \frac{w_1 * CS(A,P) + w_2 * N(A,P) + w_3 * AS(A,P)}{w_1 + w_2 + w_3} \tag{2.15}$$

Here $w_1$, $w_2$ and $w_3$ are the weights assigned to cosine similarity, N-gram similarity and approximate similarity respectively. All three similarity measures used in the unifying measure are in the range of [0,1]. The weights are assigned by using the inverse-variance weighting method [46]. Each similarity measure is weighted in inverse proportion to its variance. The weight $w_i$ for each similarity measure is thus given as:

$$w_i = \frac{1}{\sigma_i^2} \tag{2.16}$$

Here, $\sigma_i^2$ is the variance of a particular similarity measure. The range of USM is also in [0,1]. We identified a similar weighted measure which used jaccard similarity as the weighting measure for computing pearson correlation, cosine similarity and manhattan similarity [17]. The integration of knowledge-based measures in USM was not considered because they are language dependent. Though EuroVoc should be an ideal choice due to its multilingual nature, it did not prove useful in identification of transpositions in English legislation in practice [88].

### 2.2.2 *Pre-processing and vectorization*

A multilingual NLP pipeline was developed for processing the corpus. The directive and NIM documents in XML format are processed to extract the legal provisions. Each legal provision is linked to a unique label (article or provision number). The next step involves pre-processing the text. Pre-processing helps in removing noise and generating a high quality representation of text for semantic similarity. First of all sentence tokenization is carried out to segment provisions into sentences. Then word tokenizers are used to extract words from sentences. The obtained tokens are converted into lowercase. We utilized spaCy's [3] list of stopwords for French and Italian to filter out common words in directives and NIMs. For English, we used NLTK's [71] stopwords list. The punctuation was also removed. The remaining tokens were tagged with part-of-speech (POS) tags (POS tag of a token is taken as an input by the lemmatizer to correctly lemmatize it). For English we utilized NLTK's WordNet lemmatizer. For French and Italian we used spaCy's default lemmatizer. Our experiments in feature selection indicate that keeping only specific POS tags like nouns, verbs and adjectives leads to loss of essential features which are necessary in short text similarity. Other POS tags also contain important semantic information which must be preserved. Therefore, we do not filter out tokens for any particular POS tag. Each provision in the corpus is thus represented in a bag-of-words format. It is a list of each token and its count in a particular provision.

3 https://spacy.io/

Then we applied the Term Frequency-Inverse Document Frequency (TF-IDF) weighting scheme to all the provisions [111]. Each provision is represented as a vector in TF-IDF representation. The cosine similarity is computed as the cosine of the transformed query vector (article of directive) and each NIM provision vector in the corpus. The N-gram similarity was computed on the Directive-NIM corpus obtained after pre-processing. N-grams were generated for each provision in the corpus and the similarity between an article and a NIM provision was computed as discussed in section 2.2.1. The approximate similarity was computed as discussed in section 2.2.1. The unifying similarity measure (USM) was computed as the weighted arithmetic mean of all three similarity measures. For a particular query (article), the matching NIM provisions with USM values greater than or equal to the threshold value are retrieved. The metrics precision, recall and F-score were computed for each directive by incrementing threshold values from 0 to 1 at intervals of 0.01. The threshold which provides the best F-score was chosen.

### 2.2.3 *Results and Analysis*

This section presents the results of identification of NIM provisions using the USM approach. A multilingual corpus (consisting of four directives and their corresponding NIMs in English, French and Italian languages) was used to verify whether the USM approach was able to identify transpositions in different languages. Table 2.2 displays the directives and NIMs considered in the multilingual corpus. The extended English language corpus (four additional directives) was used to thoroughly evaluate the performance of USM on additional directives. The English NIMs were taken from Ireland and the UK. The French NIMs were taken from Luxembourg legislation. The Italian NIMs were taken from Italian legislation. In our research, we found official correlation tables (prepared by Member States) for certain Directive-NIM pairs for the UK and Ireland. Therefore, we were restricted to study the identification of NIM provisions in these directive-NIM pairs only. The correlation tables (where not available) for Directive-NIM pairs were prepared by a legal researcher with in-depth knowledge of the legislation at both EU and national levels. The tables were checked and reviewed by another trained legal researcher. Any differences and inconsistencies in the identified transposed provisions were resolved.

Figure 2.5 shows the results of automated identification of NIM provisions by the proposed model on the multilingual corpus. LUX refers to Directive-NIM pairs in French (with NIM from Luxembourg). ITA refers to Directive-NIM pairs in Italian (with NIM from Italy). EN refers to Directive-NIM pairs in English (with NIMs from UK in CELEX 32003L0010 and 31999L0092 and NIMs from Ireland in case of CELEX 32002L0044 and 32001L0024). We observe that the Luxembourg

Table 2.2: Directives and NIMs in the multilingual corpus

| Directive-NIM group | Directives (CELEX number) | NIMs (English) | NIMs (French) | NIMs (Italian) |
|---|---|---|---|---|
| (Directive1, NIM1) | 32003L0010 | United Kingdom (Statutory Instrument No. 1643 of 28/06/2005) | Luxembourg (Memorial A,Number:23, 02/03/2007) | Italy (Decreto Legislativo, Number 195/2006) |
| (Directive2, NIM2) | 32002L0044 | Ireland (Statutory Instrument No. 370/2006) | Luxembourg (Memorial A,Number:23, 02/03/2007) | Italy (Decreto Legislativo, Number 187/2005) |
| (Directive3, NIM3) | 32001L0024 | Ireland (Statutory Instrument No. 198/2004) | Luxembourg (Memorial A,Number:45, 29/03/2004) | Italy (Decreto Legislativo, Number 197/2004) |
| (Directive4, NIM4) | 31999L0092 | United Kingdom (Statutory Instrument No. 2776 of 7/11/2002) | Luxembourg (Memorial A,Number:39, 05/04/2005) | Italy (Decreto Legislativo, Number 233/2003) |

Directive-NIM pair achieves the highest F-score and recall for three directives (CELEX: 32003L0010, 32002L0044 and 31999L0092). For CELEX 31999L0092, the Italian Directive-NIM pair too achieves the highest F-score along with the Luxembourg pair. The English Directive-NIM pair achieved the highest recall and F-score only in CELEX 32001L0024. We also computed the macro-average precision, recall and F-score measures for USM across all four directives (Figure 2.6).

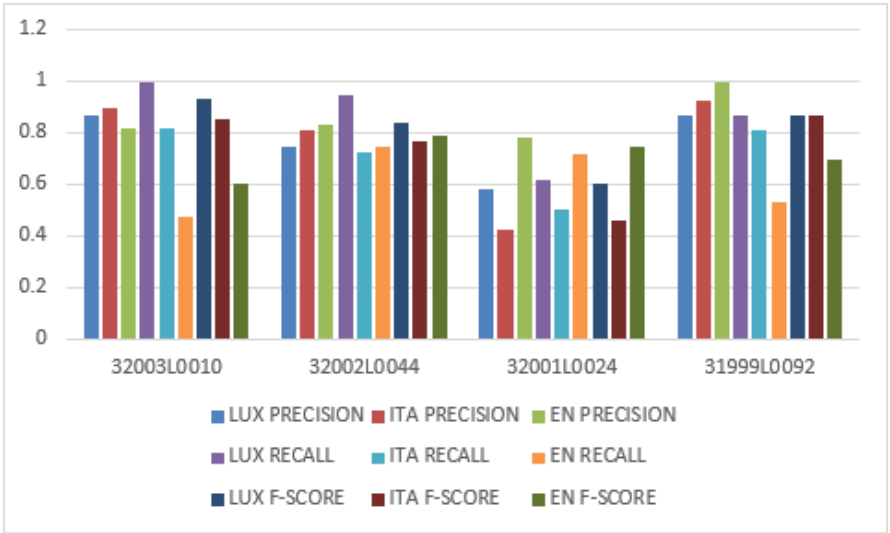


Figure 2.5: Results of automated identification of NIM provisions by USM on the multilingual corpus of four directives

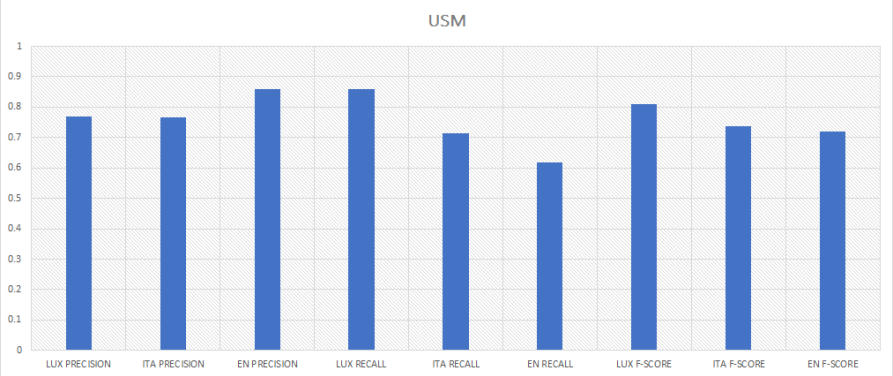


Figure 2.6: Macro-average precision, recall and F-score for USM across all four directives in the multilingual corpus

The macro-average values of the evaluation metrics across all four directives indicate that the Luxembourg Directive-NIM pairs consistently achieved better recall and F-score than their Italian and English counterparts. This implies that our system was able to identify a greater number of transposing provisions per directive for Luxembourgish legislation. This is because the Luxembourg NIM provisions used wordings and terminologies similar to the European directives. We consider Article 5.1 of Directive CELEX 32002L0044 and their corresponding transposing provisions for Luxembourg, Italian and Irish legislation as per the correlation tables (Figures 2.7, 2.8 and 2.9 respectively).

| Directive 32002L0044 | Luxembourg NIM provision |
|---|---|
| 1. En tenant compte du progrès technique et de la disponibilité de mesures de maîtrise du risque à la source, les risques résultant de l'exposition aux vibrations mécaniques sont supprimés à leur source ou réduits au minimum.<br>La réduction de ces risques se base sur les principes généraux de prévention figurant à l'article 6, paragraphe 2, de la directive 89/391/CEE. | 1.En tenant compte du progrès technique et de la disponibilité de mesures de maîtrise du risque à la source, les risques résultant de l'exposition aux vibrations mécaniques sont supprimés à leur source ou réduits au minimum.<br>La réduction de ces risques se base sur les principes généraux de prévention figurant à L. 312-2, (2), du Code du travail. |

Figure 2.7: Article 5.1 of Directive CELEX 32002L004 and corresponding NIM provision 5.1 of Luxembourg

In the case of Luxembourg (Figure 2.7), the presence of many similar words between the directive and the NIM facilitates the transposition identification by the system. However, in the Italian case (Figure 2.8), both the article and NIM provision have partly similar meaning. Both the article and provision talk about reducing and eliminating the risks, but miss out some key information. The NIM does not mention mechanical vibration (referred to as "vibrazioni meccaniche" in the article), while the article does not mention exposure limit values (re-

| Directive 32002L0044 | Italian NIM provision |
|---|---|
| Tenendo conto del progresso tecnico e della disponibilità di misure per controllare il rischio alla fonte, i rischi derivanti dall'esposizione alle vibrazioni meccaniche sono eliminati alla fonte o ridotti al minimo.<br>La riduzione di tali rischi si basa sui principi generali di prevenzione di cui all'articolo 6, paragrafo 2, della direttiva 89/391/CEE. | Fermo restando quanto previsto dall'articolo 3 del decreto legislativo 19 settembre 1994, n. 626, il datore di lavoro elimina i rischi alla fonte o li riduce al minimo e, in ogni caso, a livelli non superiori ai valori limite di esposizione. |

Figure 2.8: Article 5.1 of Directive CELEX 32002L004 and corresponding NIM provision 5.1 of Italy

| Directive 32002L0044 | Ireland NIM provision |
|---|---|
| Taking account of technical progress and of the availability of measures to control the risk at source, the risks arising from exposure to mechanical vibration shall be eliminated at their source or reduced to a minimum.<br>The reduction of such risks shall be based on the general principles of prevention set out in Article 6(2) of Directive 89/391/<br>EEC. | Having regard to the general principles of prevention in Schedule 3 to the Act, an employer shall ensure so far as is reasonably practicable that risk from the exposure of his or her employees to mechanical vibration is either eliminated at source or reduced to a minimum. |

Figure 2.9: Article 5.1 of Directive CELEX 32002L004 and corresponding NIM provision 6.1 of Ireland

ferred to as "valori limite di esposizione" in the provision). The NIM provision also refers to a national measure instead of the European directive. Due to these factors, the system was not able to identify this transposition. In the case of Ireland (Figure 2.9), both the directive and NIM provision convey the same meaning, but the NIM does not mention technical progress and availability of measures. The NIM also refers to another national measure. However, due to the presence of two common sequences, "general principles of prevention" and "reduced to a minimum" and a few common words like "mechanical vibration" and "eliminated" the system is able to identify this transposition (as the proposed model utilizes N-grams for sequences and cosine similarity for common words).

The above example illustrates the differences in transposing the directives in different Member States. The Luxembourg legislation had more instances where the provisions share common words and sentence structures with the directives, thus resulting in higher recall. The English and Italian legislation had only few such cases. The English Directive-NIM pairs had lower average recall and F-score than the Italian and Luxembourg pairs. This is because in many cases in English NIMs, the provisions and articles use different words and sentence structures. The average F-score values for Luxembourg,

Italian and English Directive-NIM pairs were 0.808, 0.736 and 0.708 respectively.

We briefly discuss the content of the directives and their corresponding NIMs in the English language for the multilingual corpus. Directive CELEX 32001L0024 focuses on the measures to be taken by Member States on the reorganisation and winding up of credit institutions. The corresponding NIM (Reorganisation and Winding-Up of Credit Institutions Regulations 2004) is coherent with the directive and and provides precise implementation of the directive articles. For instance, one article in the directive states that an "administrative or judicial authority" must inform the competent authorities of other host Member States about the opening of proceedings. The corresponding transposing provision states that the "Bank" must inform the competent authority by any available means about the opening of proceedings. Thus, we observe that NIM implementations are more specific and takes into account the national legal framework. Similar observations were recorded for Directive CELEX 31999L0092 which discusses the minimum requirements for improving the safety and health protection of workers at risk from explosive atmospheres.

Directives CELEX 32003L0010 and 32002L0044 have a very similar structure as both are focussed on the minimum health and safety requirements regarding the exposure of workers to risks arising from physical agents. CELEX 32003L0010 considers noise whereas CELEX 32002L0044 considers vibration. Both directives share some common article headings like, "Determination and assessment of risks", "Provisions aimed at avoiding or reducing exposure", "Worker information and training". However, the articles are focused on their respective domains, ie. noise and mechanical vibration. Figure 2.10 shows the transposition of two very similar articles from these two directives. We observe that the content of both articles is almost the same. We also observe that articles of both directives recommend Member States to adopt provisions for health surveillance of workers in case of a risk to their health.

The UK and Ireland NIM provisions are more specific than the directive articles and explicitly mention risks from hearing and mechanical vibration respectively. However, the directive articles in Figure 2.10 do not make a distinction between the risks arising from noise and vibration (even though CELEX 32003L0010 and 32002L0044 consider risks arising from noise and vibration respectively). The similar structure and presence of a few common words and sequences between Ireland NIM provision and directive CELEX 32002L0044 facilitates the transposition identification and results in a relatively higher F-score than CELEX 32003L0010 and the UK NIM provision.

Figure 2.6 shows the evaluation metrics averaged over all directives. These results indicate that our model was able to identify transpositions with good performance on legislation written in three different

| Directive 32003L0010 Article 10.1 | UK NIM Provision 9.1 |
|---|---|
| Without prejudice to Article 14 of Directive 89/391/EEC, Member States shall adopt provisions to ensure the appropriate **health surveillance** of workers where the results of the **assessment** and measurement provided for in Article 4(1) of this Directive indicate **a risk to** their **health**. Those provisions, including the requirements specified for health records and their availability, shall be introduced in accordance with national law and/or practice. | If the risk **assessment** indicates that there is **a risk to** the **health** of his employees who are, or are liable to be, exposed to noise, the employer shall ensure that such employees are placed under suitable **health surveillance**, which shall include testing of their hearing. |

| Directive 32002L0044 Article 8.1 | Ireland NIM Provision 8.1 |
|---|---|
| **Without prejudice** to Article 14 of Directive 89/391/EEC, Member States shall adopt provisions to **ensure** the **appropriate health surveillance** of workers with reference to the outcome of the **risk assessment** provided for in Article 4(1) of this Directive where it indicates **a risk to their health.** Those provisions, including the requirements specified for health records and their availability, shall be introduced in accordance with national laws and/or practice. | **Without prejudice** to section 22 of the Act, it shall be the duty of an employer to **ensure** that **appropriate health surveillance** is made available to those employees for whom a **risk assessment** referred to in Regulation 5 reveals **a risk to their health**, including employees exposed to mechanical vibration in excess of an exposure action value. |

Figure 2.10: Two articles from directives CELEX 32003L0010 and CELEX 32002L0044 transposed by UK NIM and Ireland NIM provision respectively

languages. This demonstrates that our model could be scalable for identifying transpositions in an automated way in different legal systems.

### 2.2.4 *Comparison of USM with state-of-the-art methods on the Multilingual corpus*

In this section, we compare the results of the unifying similarity measure with state-of-the art text similarity measures on the multilingual corpus of four directives. We implemented Euclidean similarity, Manhattan similarity, Latent Semantic Analysis (LSA) and Latent Dirichlet allocation (LDA) methods and evaluated their results on the multilingual corpus of four directives and their corresponding NIMs in English, French and Italian languages. Figure 2.11 show the comparison of USM with other state-of-the-art methods.

#### 2.2.4.1 *Italian Legislation Results*

In the case of Italian Directive-NIM pairs, USM outperforms other methods in terms of F-score in all four directives. It also achieved a higher recall than other methods in three directives (CELEX: 32003L0010, 32002L0044 and 31999L0092). USM further achieved the highest precision for CELEX 32003L0010 and 31999L0092. However, LDA achieved better precision than USM in CELEX 32002L0044. This is because LDA


Figure 2.11: Comparison of the Unifying Similarity Measure (USM) with Euclidean, Manhattan, Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA) similarity measures on the multilingual corpus

could retrieve very few transpositions and had the lowest recall among all methods for CELEX 32002L0044. So, it was able to identify those few transpositions with a higher precision. We further computed the macro-average precision, recall and F-score values across all directives for different similarity measures. The results are shown in Figure 2.12. For the Italian Directive-NIM pairs , we observe that USM outperforms other state-of-the-art methods in all three metrics: precision, recall and F-score. The macro-average F-score for USM was 0.738. USM was also able to retrieve a greater number of transpositions than other methods as it achieved a higher recall.

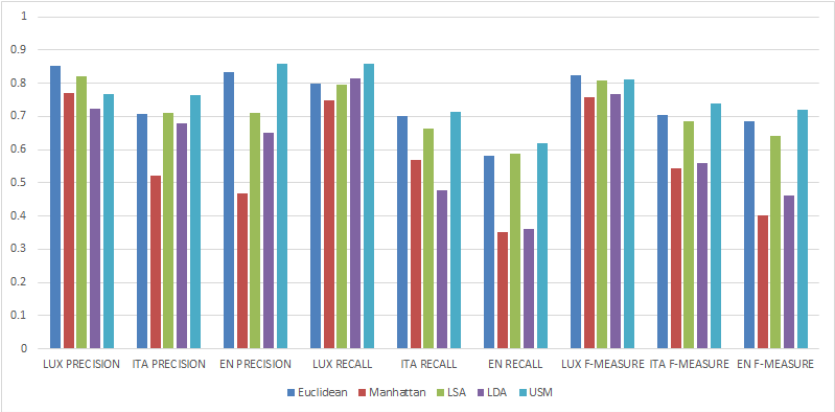


Figure 2.12: Comparison of USM with state-of-the-art similarity measures for macro-average precision, recall and F-score across all four directives

#### 2.2.4.2 *Luxembourg Legislation Results*

In the case of the Directive-NIM pairs written in French, USM achieved the best F-scores in CELEX 320003L0010 and 32001L0024. However in CELEX 32002L0044 and 31999L0092, Euclidean similarity achieved the best F-score, although the F-score of USM is very close to Euclidean similarity in both directives and both values are above 0.8. So there is only a small difference.

Now we closely examine the reasons for this performance. For CELEX 32002L0044, the number of obtained true positives were same for both USM and Euclidean. Also the recall of USM was much higher than Euclidean. So, the higher F-score of Euclidean is because of its perfect precision. One key motivation for proposing USM was to increase the recall (to identify as many transpositions as possible, by incorporating complementary similarity measures). However, one of the limitations of such a weighted mean is an increase in the number of false positives (in some cases). This is because our model takes into account three different similarity measures (which check for three different features) and sometimes the presence of just a few matching features may not result in a true positive. The same explanation also

holds true for CELEX 31999L0092 (where the recall of USM and Euclidean is the same, but Euclidean achieves higher precision). The results of comparison of average values (Figure 2.12) indicate that Euclidean similarity achieved the best average F-score, while USM was second best with a very minute difference. In terms of recall, USM outperformed other methods. However, Euclidean similarity was successful in achieving a higher average precision than USM (due to more false positives by USM).

#### 2.2.4.3 *English Legislation Results*

In this section, we discuss the results of Directive-NIM pairs in English. For three Directives, CELEX: 32002L0044, 32001L0024 and 31999L0092, USM achieves a higher F-score than other methods. For CELEX 32003L0010, both USM and Euclidean similarity achieve the best F-score. Also the recall of USM was higher than other methods for CELEX 32002L0044 and 31999L0092. In the case of CELEX 32001L0024 and 32003L0010, USM achieved the second highest recall. In terms of the average comparison of evaluation metrics (Figure 2.12), USM also achieved the best performance in F-score, recall and precision.

We observed from the results that USM achieved the highest recall in all three cases of Luxembourg, Italian and English legislation. This shows that USM was able to identify more transpositions than other methods. This is possible because USM checks for multiple features when comparing texts, while other methods just look for one. USM benefits from the complementary nature of different similarity techniques. We illustrate one example in Figure 2.13 where USM was able to identify the transposition but other methods - Euclidean, LSA, LDA and Manhattan failed. It can be observed that though the NIM provision transposes the corresponding article, the language in the NIM is quite different from that of directive. The NIM also does not mention anything about reviewing the derogations every four years. The presence of a common sequence, "that the resulting risks are reduced to a minimum" and a few common words like "health surveillance" and "special circumstances" were enough for the USM model to identify this transposition. The other methods failed to identify such cases of transposition as they did not consider approximate matching and N-gram similarity.

It is also interesting to observe that latent semantic analysis (LSA) had a decent performance in evaluation. It was chosen because of its ability to extract the meaning of words by analyzing patterns in word usage across different provisions, so that it would be useful to identify cases of transposition where NIM and directives use different words. The application of singular value decomposition (SVD) may cause some important features (relevant for text similarity) to be lost, thus resulting in low recall. This was also evident in English, Luxembourg and Italian legislation where LSA achieved lower recall than USM (Fig-

| Directive 32002L0044 | Ireland NIM |
|---|---|
| The derogations referred to in paragraphs 1 and 2 shall be granted by Member States after consultation of the two sides of industry in accordance with national laws and practice. Such derogations must be accompanied by conditions which guarantee, taking into account the special circumstances, that the resulting risks are reduced to a minimum and that the workers concerned are subject to increased health surveillance. Such derogations shall be reviewed every four years and withdrawn as soon as the justifying circumstances no longer obtain. | The Authority shall not grant any exemptions under this Regulation unless-<br>(a) it consults the employers and the employees concerned or their representatives, or both,<br>(b) it applies conditions to any such exemption, taking into account the special circumstances, to ensure that the resulting risks are reduced to a minimum, and<br>(c) the employees concerned are subject to appropriate health surveillance |

Figure 2.13: Article 10.3 from dir. CELEX 32002L0044 and corresponding NIM provision 10.3 of Ireland identified by USM

ure 2.12). The performance of Latent Dirichlet Allocation was poorer in terms of recall as compared to LSA, USM and Euclidean similarity in English and Italian legislation. LDA is a generative model which discovers a latent distribution of topics in a corpus of documents. It is based on the assumption that a document can be represented as a mixture of hidden topics [10]. LDA is a probabilistic topic model characterized by a conditional word by document probability distribution, $p(w|d)$ [8]. This distribution is a combination of topic by document distribution p(z | d) and word by topic distribution $p(w|z)$:

$$p(w|d) = \sum_{z} p(w|z)p(z|d) \tag{2.17}$$

Thus, each document $d$ is represented as a multinomial distribution of latent topics $z$, and each topic $z$ is represented as a multinomial distribution of words $w$. The LDA transform is applied over the tf-idf provision term matrix to obtain provision-topic matrix. Each provision vector is thus represented in a reduced dimension as a topic distribution. LDA considers each provision as a mixture of hidden topics and each topic as a mixture of words. The topics generated by LDA (in the articles and NIM provisions) were quite different when the articles and NIM provisions used different words. This influenced the similarity values and resulted in a lower recall for Italian and English legislation (where the directive and NIM have different wordings in many cases), as shown in Figure 2.12. In case of Luxembourg legislation, the recall of LDA was high enough as the wordings are more similar. Our experiments with different number of topics suggested that LDA's performance improved with the increase in number of topics. We chose 500 topics for the LDA model.

The Euclidean similarity measure is based on the Euclidean distance. Its a lexical similarity measure which was applied to the tf-idf vectors to compute similarity. It achieved a higher recall than other methods for Luxembourg legislation as there were many similar words.

However, for Italian and English legislation the achieved recall was lower than USM. Manhattan similarity is a similarity measure based on the Manhattan distance. The value of Manhattan distance is higher than Euclidean distance and thus the similarity values are much lower. The Manhattan distance follows a grid-like path and the computed distance between two provision vectors may not provide a reasonable estimate of their similarity.

### 2.2.5 *Results on the extended English corpus*

The results of transposition identification on the multilingual corpus suggested that there was greater linguistic variability in the English transpositions, whereas the French and Italian texts had more words and phrases in common. The English Directive-NIM pairs had a lower average F-score of 0.708 as compared to 0.736 and 0.808 of Italian and French Directive-NIM pairs respectively. Therefore the English text was deemed the most challenging and appropriate for further in-depth evaluation of USM compared to other models.

In this section, we evaluate the performance of USM on an additional corpus of 4 directives and their corresponding NIMs in the English language[4]. The NIMs were taken from the legislation of Ireland. Table 2.3 shows the results of automated identification of NIM provisions. We observe that USM clearly outperforms other state-of-the-art text similarity measures in terms of F-score and recall. Thus, USM model achieved encouraging results over the multilingual and the extended English language corpus. A more thorough evaluation of different lexical and semantic techniques on a larger corpus is presented in section 2.3.

## 2.3 EVALUATION OF LEXICAL AND SEMANTIC UNSUPERVISED TEXT SIMILARITY MODELS ON A MULTILINGUAL CORPUS OF 43 DIRECTIVES

The first two sections presented encouraging results on small corpora (both monolingual corpus in English from Section 2.1 and a multilingual corpus from Section 2.2). In this section, we evaluate the previously discussed lexical and semantic similarity models (TF-IDF Cosine, USM, LSA and LDA) on a larger multilingual corpus of 43 directives and their corresponding NIMs.

4 the corresponding list of NIMs from Ireland in order of the directives mentioned in Table 2.3 are : SI No. 619/2001, SI No. 572/2013, SI No.875/2005, SI No.176/2010, where SI refers to Statutory Instrument

Table 2.3: Comparison of USM with state-of-the-art text similarity methods on the extended English corpus

| | Precision | | | | | Recall | | | | | F-Score | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Directives | Euclidean | Manhattan | LSA | LDA | USM | Euclidean | Manhattan | LSA | LDA | USM | Euclidean | Manhattan | LSA | LDA | USM |
| 31998L0024 | 0.9565 | 1 | 0.9545 | 0.607 | 0.92 | 0.6666 | 0.147 | 0.6363 | 0.6296 | 0.6969 | 0.7857 | 0.2564 | 0.7636 | 0.6181 | 0.7931 |
| 32000L0054 | 0.5937 | 0.7 | 0.6785 | 0.6666 | 0.6764 | 0.6333 | 0.1707 | 0.5588 | 0.5454 | 0.7187 | 0.6129 | 0.2745 | 0.6129 | 0.6 | 0.6969 |
| 32003L0122 | 0.7272 | 0.6428 | 0.8 | 0.7777 | 0.6923 | 0.5714 | 0.6428 | 0.5333 | 0.5 | 0.6923 | 0.64 | 0.6428 | 0.64 | 0.6086 | 0.6923 |
| 32006L0025 | 0.6923 | 0.5714 | 0.8181 | 0.4166 | 0.6667 | 0.5625 | 0.4705 | 0.5294 | 0.3846 | 0.6667 | 0.6206 | 0.5161 | 0.6428 | 0.4 | 0.6667 |

### 2.3.1 *Corpus Preparation*

The first step involved creating a corpus of directives and NIMs. We prepared a multilingual parallel corpus of 43 directives and their corresponding NIMs from Ireland, Luxembourg and Italian legislation. Table 2.4 presents the CELEX numbers of the directives and NIMs as per EUR-Lex. The chosen directives were taken from different subject matters. This is because only a very few directives belong to a particular subject matter. We needed a bigger corpus for a thorough evaluation of our system on directives from different subject matters. Only the directives for which all the three Member States (Ireland, Luxembourg and Italy) have communicated the NIMs to the Commission were chosen. This information was obtained from EUR-Lex which provides a list of NIMs communicated by each Member State for a particular directive. We were restricted in our choice of directives due to the fact that many directives did not have NIMs communicated from all three Member states (Luxembourg, Ireland and Italy). Due to the highly time-consuming process of preparing the gold standard mapping we did not include directives with a large number of NIMs. This is an aspect which deserves further investigation and can be addressed in future work. The consolidated version of the directives was used in the corpus. Each legislative document was stored in a proprietary XML format with each XML element representing a legal provision (directive article or NIM provision). A gold standard mapping between directive articles and NIM provisions was prepared by two legal researchers with expertise in European law. An inter-annotator agreement was computed for each language corpus (of 43 directives and their corresponding NIMs) using Cohen's Kappa [80]. The mean Kappa scores for English, French and Italian corpus are 0.4812, 0.79 and 0.6065 respectively. This indicates that the agreement was highest in the French Directive-NIM corpus and the lowest in English Directive-NIM corpus.

### 2.3.2 *Pre-processing and vectorization*

We utilize the same pre-processing pipeline as discuss in section 2.2.2.

### 2.3.3 *Results and Analysis of Lexical and Semantic Unsupervised Text Similarity Models on the Multilingual corpus of 43 directives*

In this section, we evaluate the models discussed in the above sections on the multilingual corpus of 43 directives and their corresponding NIMs. The metrics precision, recall and F-score were computed for each directive by incrementing threshold values from 0 to 1 at intervals of 0.01. The threshold which provides the best F-score was chosen. We then computed the macro-average precision, recall and F-score

Table 2.4: The CELEX numbers of directives and NIMs in the multilingual corpus

| Sno | Directives | NIMs (Ireland) | NIMs (Luxembourg) | NIMs (Italy) |
|---|---|---|---|---|
| 1 | 32010L0024 | 72010L0024IRL_188115 | 72010L0024LUX_194845 | 72010L0024ITA_195371 |
| 2 | 32009L0128 | 72009L0128IRL_190844 | 72009L0128LUX_222878<br>72009L0128LUX_222460 | 72009L0128ITA_195369 |
| 3 | 31994L0011 | 71994L0011IRL_97765 | 71994L0011LUX_97767 | 71994L0011ITA_97762 |
| 4 | 31996L0040 | 71996L0040IRL_103146 | 71996L0040LUX_103149 | 71996L0040ITA_103143 |
| 5 | 31996L0093 | 71996L0093IRL_104711 | 71996L0093LUX_104713 | 71996L0093ITA_104707 |
| 6 | 31997L0043 | 71997L0043IRL_106134 | 71997L0043LUX_106145 | 71997L0043ITA_106123 |
| 7 | 31998L0058 | 71998L0058IRL_107788 | 71998L0058LUX_107790 | 71998L0058ITA_107789 |
| 8 | 31998L0084 | 71998L0084IRL_108777 | 71998L0084LUX_108779 | 71998L0084ITA_108778 |
| 9 | 31999L0105 | 71999L0105IRL_111554 | 71999L0105LUX_126553<br>71999L0105LUX_126554 | 71999L0105ITA_111555 |
| 10 | 32009L0021 | 72009L0021IRL_184902 | 72009L0021LUX_189874 | 72009L0021ITA_186849 |
| 11 | 31999L0002 | 71999L0002IRL_109429 | 71999L0002LUX_109431 | 71999L0002ITA_109430 |
| 12 | 32009L0020 | 72009L0020IRL_188439 | 72009L0020LUX_189875 | 72009L0020ITA_194551 |
| 13 | 31999L0095 | 71999L0095IRL_111630 | 71999L0095LUX_111632 | 71999L0095ITA_125921 |
| 14 | 32009L0033 | 72009L0033IRL_183965 | 72009L0033LUX_183231 | 72009L0033ITA_179616 |
| 15 | 32000L0036 | 72000L0036IRL_112636 | 72000L0036LUX_112638 | 72000L0036ITA_112637 |
| 16 | 32000L0055 | 72000L0055IRL_113427 | 72000L0055LUX_113429 | 72000L0055ITA_113428 |
| 17 | 32001L0110 | 72001L0110IRL_116005 | 72001L0110LUX_116006 | 72001L0110ITA_30057 |
| 18 | 32008L0090 | 72008L0090IRL_168455 | 72008L0090LUX_168629 | 72008L0090ITA_170924 |
| 19 | 32001L0112 | 72001L0112IRL_116042 | 72001L0112LUX_116043 | 72001L0112ITA_29334 |
| 20 | 32001L0113 | 72001L0113IRL_116060 | 72001L0113LUX_116062 | 72001L0113ITA_116061 |
| 21 | 32007L0002 | 72007L0002IRL_170884 | 72007L0002LUX_170775 | 72007L0002ITA_167690 |
| 22 | 32007L0043 | 72007L0043IRL_170239 | 72007L0043LUX_170162 | 72007L0043ITA_173275 |
| 23 | 32007L0033 | 72007L0033IRL_170294 | 72007L0033LUX_170795 | 72007L0033ITA_173410 |
| 24 | 32001L0111 | 72001L0111IRL_116024 | 72001L0111LUX_116026 | 72001L0111ITA_116025 |
| 25 | 32005L0094 | 72005L0094IRL_142403 | 72005L0094LUX_131762 | 72005L0094ITA_167074 |
| 26 | 32001L0081 | 72001L0081IRL_115688<br>72001L0081IRL_194972 | 72001L0081LUX_115689 | 72001L0081ITA_29985 |
| 27 | 32001L0095 | 72001L0095IRL_28698 | 72001L0095LUX_135144 | 72001L0095ITA_29986<br>72001L0095ITA_135265 |
| 28 | 32004L0023 | 72004L0023IRL_131105 | 72004L0023LUX_147977 | 72004L0023ITA_150656<br>72004L0023ITA_150706 |
| 29 | 32001L0096 | 72001L0096IRL_115977<br>72001L0096IRL_115978 | 72001L0096LUX_115979 | 72001L0096ITA_35623 |
| 30 | 32002L0092 | 72002L0092IRL_34868 | 72002L0092LUX_126481<br>72002L0092LUX_123898 | 72002L0092ITA_125142 |
| 31 | 32003L0094 | 72003L0094IRL_33063 | 72003L0094LUX_33944 | 72003L0094ITA_132883 |
| 32 | 32014L0028 | 72014L0028IRL_239853 | 72014L0028LUX_243958 | 72014L0028ITA_237982 |
| 33 | 32015L0413 | 72015L0413IRL_250326 | 72015L0413LUX_234950 | 72015L0413ITA_214698 |
| 34 | 32013L0053 | 72013L0053IRL_245865 | 72013L0053LUX_243962<br>72013L0053LUX_243961 | 72013L0053ITA_233695<br>72013L0053ITA_233693 |
| 35 | 32006L0088 | 72006L0088IRL_157218 | 72006L0088LUX_153017 | 72006L0088ITA_158323 |
| 36 | 32008L0057 | 72008L0057IRL_185250 | 72008L0057LUX_169960 | 72008L0057ITA_173702 |
| 37 | 32008L0096 | 72008L0096IRL_186546 | 72008L0096LUX_190526 | 72008L0096ITA_180588<br>72008L0096ITA_180158 |
| 38 | 32008L0043 | 72008L0043IRL_161791 | 72008L0043LUX_161581<br>72008L0043LUX_161580 | 72008L0043ITA_166919 |
| 39 | 32005L0062 | 72005L0062IRL_137665 | 72005L0062LUX_129420 | 72005L0062ITA_150819<br>72005L0062ITA_150669<br>72005L0062ITA_150695 |
| 40 | 31999L0092 | 71999L0092IRL_111679 | 71999L0092LUX_120249 | 71999L0092ITA_111680 |
| 41 | 32001L0024 | 72001L0024IRL_180124<br>72001L0024IRL_28393 | 72001L0024LUX_114418 | 72001L0024ITA_30729 |
| 42 | 32002L0044 | 72002L0044IRL_133618 | 72002L0044LUX_142436 | 72002L0044ITA_124474 |
| 43 | 32003L0010 | 72003L0010IRL_133619 | 72003L0010LUX_142437 | 72003L0010ITA_132468 |

metrics for each legislation corpus (Ireland, Luxembourg and Italy). The macro-average precision is computed by taking the average of the precision values for the 43 directives (for a particular legislation). The macro-average recall is computed by taking the average of the recall values for the 43 directives (for a particular legislation). The macro-average F-score is the harmonic mean of the macro-average precision and macro-average recall.

We implemented two variants of USM, USM_chars, with character N-grams and USM_tokens, with token N-grams. We utilized 4-grams for both cases and N-gram similarity was computed as discussed in section 2.2.1. Figure 2.14 presents the macro-average precision, recall and F-score of the lexical and semantic unsupervised text similarity models over the multilingual corpus. We observe that the Luxembourg Directive-NIM corpus achieves a higher precision, recall and F-score than the English and Italian corpus for each similarity measure. This is because of the presence of common words and phrases in European directives and the Luxembourg legislation. The Irish and Italian legislation had more linguistic variation with respect to the European directives.

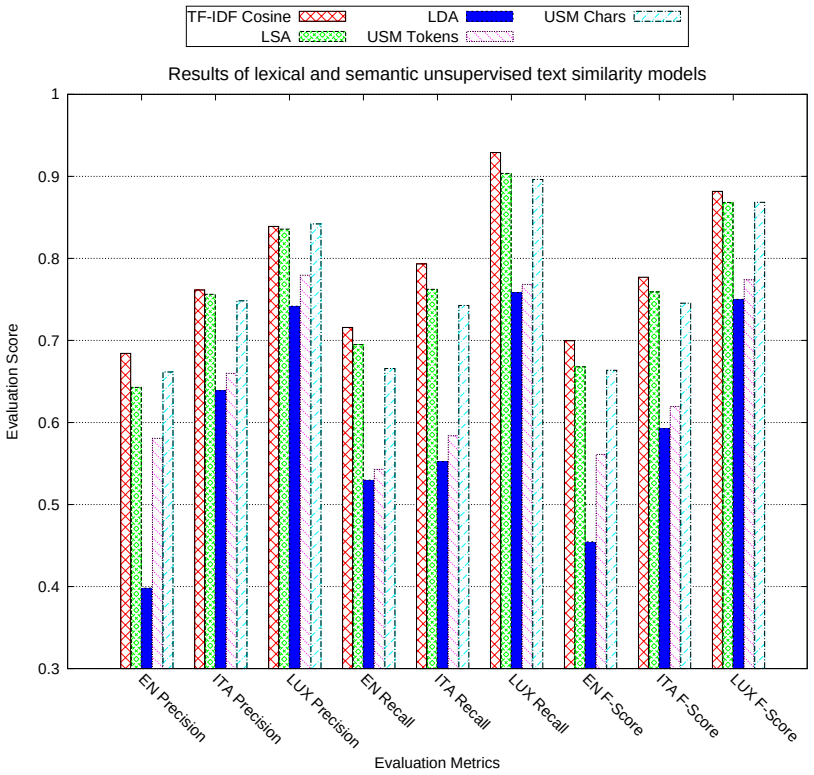


Figure 2.14: Results of the lexical and semantic unsupervised text similarity models on a multilingual corpus of 43 directives

TF-IDF cosine similarity measure achieved the best F-score for all three corpora. The performance of LSA and USM_chars model was

Table 2.5: Article 3.2 of Directive (CELEX Number: 32008L0096) and its implementing NIM provision 4.2 from Ireland legislation (CELEX Number: 72008L0096IRL_186546)

| Article 3.2 of Directive | Provision 4.2 of Ireland NIM |
|---|---|
| The road safety impact assessment shall be carried out at the initial planning stage before the infrastructure project is approved. In that connection, Member States shall endeavour to meet the criteria set out in Annex I. | The road safety impact assessment shall be carried out at the initial planning stage of the infrastructure project, before— (a) in the case of an infrastructure project coming within Part IV of the Act of 1993, submitting a scheme to An Bord Pleanála, pursuant to sections 47 and 49 of the Act of 1993, as amended by sections 9 and 11 of the Act of 2007, or (b) in any other case, submitting an application for consent for the infrastructure project under the Planning and Development Act 2000 (No. 30 of 2000) and Regulations made under Part XI of that Act. |

comparable and they were the second best methods after TF-IDF cosine in terms of F-score. LSA has a slightly better performance (F-score) than USM_chars for English and Italian corpus.

These results indicate that the application of dimensionality reduction techniques, such as LSA and LDA do not improve the performance of the text similarity system. The idea behind such techniques is to reduce the variability in word usage and thus highlight the latent relations between words and documents which were obscured by noise [28]. However, in case of short texts, such as legal provisions the reduction of dimensions results in loss of key features which maybe relevant for semantic similarity. This is also demonstrated in Irish, Italian and Luxembourg legislation corpus where LSA achieved a lower recall than TF-IDF cosine (Figure 2.14). In terms of precision, the performance of LSA is almost equivalent to TF-IDF cosine (in Luxembourg and Italian legislation). The overall performance of LDA was poorer as compared to other methods. In case of short texts (such as tweets), they have been outperformed by TF-IDF based models [50].

USM_chars model had a decent performance over the multilingual corpus. There were some transpositions which were identified by USM_chars but missed by other methods. Table 2.5 presents one such example. It can be observed that the only similar part in directive article and NIM provision is about the road safety impact assessment being carried at the planning stage of the infrastructure project. The NIM provision then goes in further details which are not mentioned in the directive article. The N-gram and approximate string

matching features of USM facilitate the identification of such cases of transposition.

## 2.4 SUMMARY

This chapter presented the investigation of unsupervised lexical and semantic similarity techniques for automated identification of national implementing measures (NIMs) of European Union directives. We presented a thorough evaluation of the different similarity techniques with detailed results and analysis on two smaller corpora and a multilingual corpus of 43 directives and their corresponding NIMs in English, French and Italian. The major text similarity methods presented in this chapter include: TF-IDF Cosine, Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA) and the Unifying Similarity Measure (USM). Our results indicate that TF-IDF cosine outperformed other unsupervised text similarity models in terms of F-Score over the multilingual corpus of 43 directives and NIMs. In the next chapter, we discuss unsupervised text similarity models based on words and paragraphs embeddings.

# 3

# UNSUPERVISED TEXT SIMILARITY MODELS BASED ON WORD AND PARAGRAPH EMBEDDINGS LEARNED BY SHALLOW NEURAL NETWORKS

In the previous chapter, we investigated the application of lexical and semantic unsupervised text similarity models to identify the transpositions of European directives into national implementing measures (NIMs). The results indicate that TF-IDF Cosine achieved the best performance over the multilingual corpus of 43 directives and their corresponding NIMs. In this chapter, we investigate word and paragraph embedding models learned by shallow neural networks to identify the transposition of directives. The word embeddings obtained from Word2vec model have been utilized in many natural language processing applications. Word embeddings could be highly useful in a short text similarity task as they can be used to enrich the texts with external semantic knowledge learned from a large corpus [56]. Enriching directive and NIM provisions with external legal vocabularies could also be useful to identify transpositions because European and national law may have different terminologies. However our results in section 2.1.3 indicate that addition of EuroVoc[1] terms did not improve over the TF-IDF and latent semantic analysis models. Therefore, in this chapter we will utilize word embeddings to develop semantic similarity models for identifying transpositions.

## 3.1 WORD2VEC

Word2vec is one of the most common model used to generate word embeddings from large unlabelled corpora [82]. It is a shallow neural network which learns a distributed representation of words. It was developed to reduce the computational complexity of traditional neural networks to allow for more efficient training. Word2vec comprises two models: continuous bag-of-words (CBOW) and skip-gram. In the continuous bag-of-words (CBOW) model, the surrounding context words are used to predict the center word. Word2vec consists of one hidden layer with one input layer and one output layer. The number of neurons in the input layer is set to the number of words in the vocabulary ($V$) of the training corpus. The input layer is one-hot encoded vector representation of the words. This means for a particular word, only one entry is set to 1. The rest of the entries in the vector are set to 0. The size of the hidden layer is set to the length of the word embedding dimension ($N$). Output layer is the same size as

1 http://eurovoc.europa.eu

the input layer. The weights from the input layer to the hidden layer of the network is represented by $V \times N$ matrix $W_i$. Each row of the matrix $W_i$ represents the word vector, $V_w$ of a particular word $w$ in the input layer. For a particular context word, $k$ the one-hot encoded input vector will have $x_k = 1$ and rest of the entries, $x_{k'} = 0$, for $k' \neq k$. The matrix for hidden layer, $h$ is then given as follows:

$$h = W_i^T x = W_{(k,.)}^T = V_w^T \tag{3.1}$$

The weights from the hidden layer to the output matrix are given by the matrix $W_o$ of size $N \times V$. In this case, the score for a particular word in the vocabulary is computed as:

$$u_j = v_{w_j}^T h \tag{3.2}$$

where $v_{w_j}$ is column $j$ of matrix $W_o$.

.

CBOW is based on the idea of bag-of-words: given a word at position $t$, CBOW generates a vector averaging the embedding in the windows $[t - d, t + d]$, where $d$ is the size of the window; the averaged vector is then multiplied by the hidden layer to predict the next word.

Skip-gram, instead, is the opposite of CBOW: given a word in position $t$, it predicts the surrounding words in a windows of size $[t - d, t + d]$. The current word is used as an input to a log-linear classifier with continuous projection layer and is used to predict surrounding words in a particular range around the current word. We used both skip-gram and CBOW to generate word embeddings. We set embedding dimension to 128, number of negative samples to 16, context windows to 5, and the learning rate to 0.38.

## 3.2 FASTTEXT

FastText [14] is a word embedding model developed by Facebook. It allows to train models quickly on a large corpus. It also offers the advantage of computing the word vectors of words which were not in the vocabulary of the training set. It substantially differs from Word2Vec in terms of the loss function and the way it computes the embedding of a word. For the loss, instead of using cross-entropy, it uses a binary logistic loss, randomly sampling negative words from the vocabulary. The embedding matrix contains character n-grams embedding (of size 3, 4, 5 and 6). Given a word, the n-grams embedding that compose the word are retrieved from the matrix, summed together and multiplied by the hidden layer. The resulting vector is then passed to the loss function. Finally, the learned n-gram embedding is used to define the word embedding of all words inside the vocabulary. We utilize both CBOW and skip-gram models of fastText.

## 3.3 SYSTEM DESCRIPTION FOR TEXT SIMILARITY MODELS BASED ON WORD AND PARAGRAPH EMBEDDINGS

We require a large amount of unlabelled legal text data to train a word embeddings model. Word embeddings trained on a legal domain corpus have shown better performance on legal datasets than generic embeddings trained on Google News and Wikipedia [18]. This is because the data used to train the embeddings is quite different from the test data (legal data) on which embeddings have to be evaluated. Therefore, we collected a corpus of European directives and national legislation to train word embeddings. The European part consists of a multilingual parallel corpus of 4300 directives in English, French and Italian. The national part consists of the national legislation from 1960 to 2018 from Ireland, Luxembourg and Italy. The number of documents were 27365, 14365 and 16233 in Ireland, Luxembourg and Italian legislation respectively. The embeddings were trained on this combined corpus of European directives and national legislation. The NLP pre-processing pipeline discussed in section 2.2.2 was utilized to clean the corpus before training word embeddings. Table 3.1 presents the most similar words for four given words as per the word embeddings trained on the English Directive-NIM corpus.

Table 3.1: Most similar words for a given word as per Word2vec embeddings

| Word | Nearest words |
|---|---|
| board | vessel, master, passenger, ship |
| requirement | condition, satisfy, meet, minimum |
| notice | document, notification, collate, file |
| contract | offer, agreement, entity, purchase |

## 3.4 COMPUTATION OF PROVISION VECTORS

In order to utilize word embeddings for text similarity of legal provisions, we need to compute provision vectors. This could be done in two ways: word-sum and word-average. In word-sum, the provision vector is generated by adding the vector of the words in the provision. Given a sequence of N words, the resulting vector $e_{sum}$ is computed as follows:

$$e_{sum} = \sum_{i=1}^{N} e_i \tag{3.3}$$

where $e_i$ is the embeddings of i-th word.

In word-average, the sum of the word embeddings in a provision is divided by the provision length. The resulting average vector $e_{emb}$ is computed as follows:

$$e_{avg} = \frac{\sum_{i=1}^{N} e_i}{N} \tag{3.4}$$

We also experiment with inverse document frequency (IDF) and word-sum. Since some words in a text are more relevant compared to others, we multiply each word embedding by the IDF of the word. The average-idf provision vector, $e_{idf}$ is computed as follows:

$$e_{avg_idf} = \frac{\sum_{i=1}^{N} e_i * idf_{w_i}}{N} \tag{3.5}$$

where $idf_{w_i}$ is the IDF value of i-th word in the provision.

The formula in Equation 3.5 is very similar to TF-IDF, with the only exception that term-frequency is substituted by the embedding of the word.

## 3.5 PARAGRAPH VECTOR MODEL

We also utilized paragraph vector, an unsupervised model which learns a fixed-length distributed vector representation for texts of variable length, such as sentences, paragraphs and documents [65]. Paragraph vector model can be seen as an extension of word2vec. Word2vec involves predicting the target word given the context. The training data comprises context and target word pairs. The context may comprise not only the words but also other suitable features (for instance, part-of-speech tags of context words) which may help to predict the target word. Paragraph vector model adds a paragraph token to the context. This token represents the document or a paragraph as an additional context. This token also acts as the document or paragraph identifier. While training the word vectors, the paragraph vector is also trained. After the training is finished, the paragraph vector represents a distributed vector representation of the paragraph. The concatenation of word vectors with the paragraph vector is used to predict the next word. This model is called the Distributed Memory Model of Paragraph Vectors (PV-DM). Another variant of paragraph vector model is called Paragraph Vector without word ordering: Distributed bag of words (PV-DBOW). This method ignores the input context words which were used by the PV-DM method. It uses the paragraph vector along with an input word to predict other words in the paragraph. This model does not require to store word vectors and is thus much faster. Previous experiments have demonstrated that a paragraph vector obtained as a combination of PV-DM and PV-DBOW

models achieves a better performance as compared to utilizing paragraph vectors obtained individually from each model [65]. We also utilized a combination of PV-DM and PV-DBOW to develop provision vectors for the directive-NIM corpus. The paragraph vector model was trained on the combined unlabelled corpus of European directives and national legislation. We used the same dimension size of 128 as word2vec and fastText provision vectors.

## 3.6 RESULTS OF TEXT SIMILARITY MODELS BASED ON WORD AND PARAGRAPH EMBEDDINGS

Figure 3.1 displays the results of the word2vec model (for different provision vectors) for the multilingual corpus of 43 directives and their corresponding NIMs. We observe that the Luxembourg Directive-NIM corpus achieves the best precision, recall and F-score for different word2vec models. This result is coherent with the results of the similarity measures discussed in section 2.3.3. The performance of both skip-gram and CBOW models of word2vec is comparable across the multilingual corpus. But the CBOW model slighlty outperforms the skip-gram model in terms of F-score for all three languages. The legal datasets of European directives and national legislation used in this paper to train word embeddings are quite small as compared Wikipedia or Google News datasets which are generally used to train the embeddings. The CBOW model smoothes most of the distributional information as it models the entire context as one observation [1]. As a result, CBOW achieves better performance than skip-gram in smaller datasets. The skip-gram model on the other hand considers each word-context pair as a new observation. Therefore, the skip-gram model works better in case of a larger dataset as it provides a larger number of observations. The performance of different provision vector models for the CBOW model is comparable. The average-idf provision vector performs slightly better than other vectors in the English corpus. In French and Italian corpus, both average and average-idf vectors have the similar performance and slightly outperform the sum vector. Overall, we conclude that the average-idf had the best performance in the CBOW model. In case of the skip-gram model, all the provision vectors have similar performance.

Figure 3.2 displays the results of the fastText model for the multilingual corpus of directives and NIMs. In this case also the Luxembourg Directive-NIM corpus achieves a higher F-score than English and Italian corpus. We also observe that the skip-gram model of fastText slightly outperforms the CBOW model. This is because the skip-gram model in word2vec predicts the context only from the vectors of words present in the training corpus. Whereas the skip-gram model of fastText utilizes the vectors of the word and also vectors of the n-grams comprising the word. The presence of n-grams results in achieving

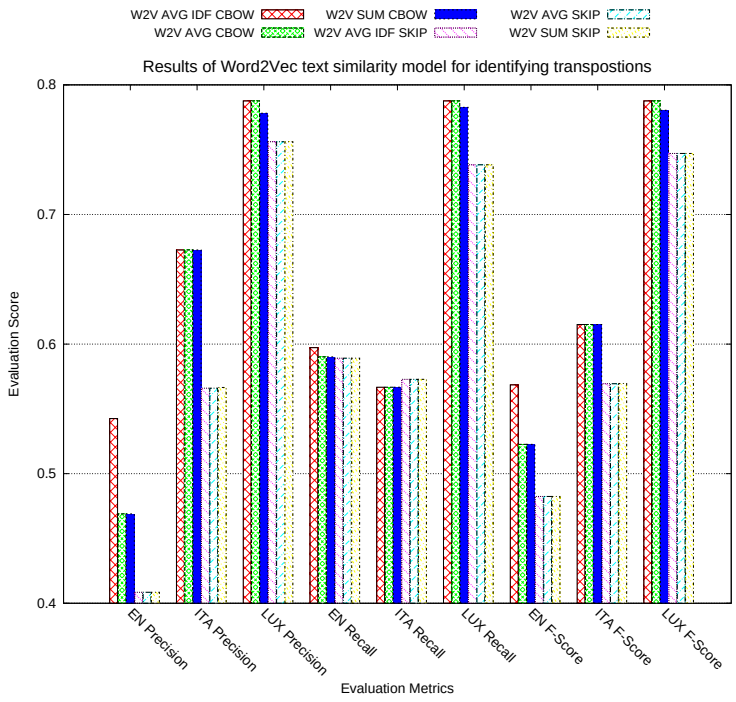


Figure 3.1: Macro-average Precision, Recall and F-score values for Skip-gram and CBOW Word2vec models

a better performance for syntactic tasks due to the addition of morphological information [14]. The performance of different provision vectors for both skip-gram and CBOW models is very similar. The average-idf vector has a slightly better performance than other vectors in case of the English corpus.

We also evaluate the paragraph vector on the multilingual corpus of 43 directives and their corresponding NIMs. Figure 3.3 displays the results of the paragraph vector and the best performing provision vectors of word2vec (average-idf of the CBOW model) and fastText (average-idf for the skip-gram model) model. The results indicate that the paragraph-vector model outperforms both word2vec and fastText in terms of F-score. One advantage of using paragraph vectors is that they take into account the word order though in a small context [65]. The provision vectors developed by the sum and average of word vectors lose the word order. Therefore, paragraph vector models show better performance to identify transpositions as compared to provision vector models of fastText and word2vec.

We also present a two-dimensional visualization of provision vectors generated by fastText and latent semantic analysis (LSA) models as shown in Figure 3.4 (fastText vectors are represented by the top plot and LSA vectors are represented by the bottom plot). The visualization is generated by using t-Distributed Stochastic Neighbour Embedding (t-SNE) ([74]). It is a dimensionality reduction algorithm

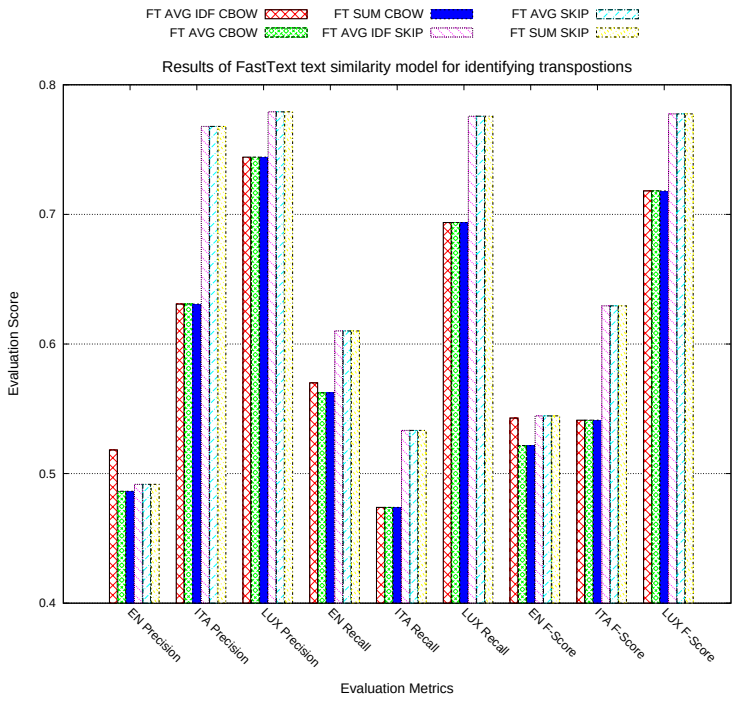


Figure 3.2: Macro-average Precision, Recall and F-score values for Skip-gram and CBOW models of FastText

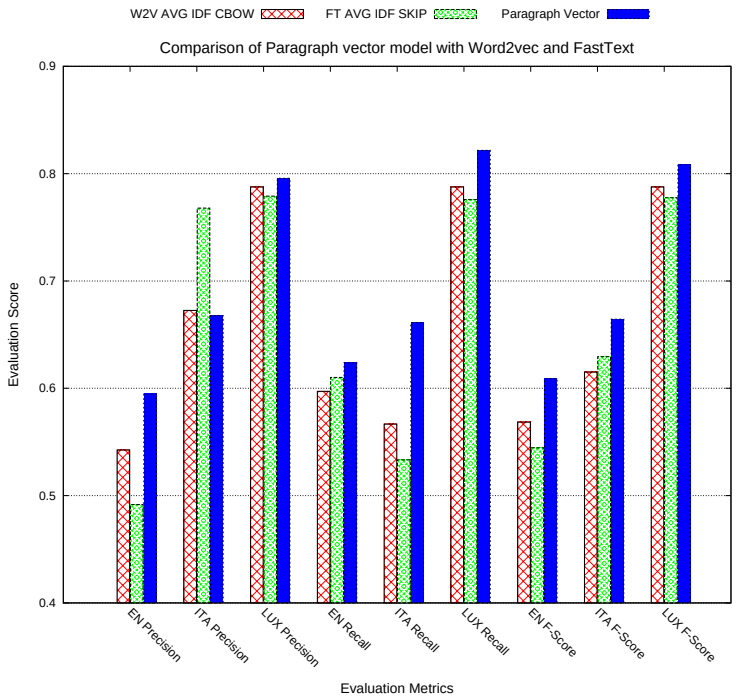


Figure 3.3: Comparison of Paragraph vector model with word2vec and fastText

which converts high-dimensional data into a low-dimensional (two or three-dimensions) space for visualization. In Figure 3.4, the labels *A* and *P* represent the directive articles and NIM provisions respectively. We encircle some article and provision pairs in both plots which are very close to each other. We observe that the pairs encircled with blue colour (A10.1, P14.1), (A3, P2.1), (A9.1, P13) and (A2, P3.2) are clustered together in both fastText and LSA plots. These pairs of transposition were correctly identified by both fastText and LSA. In the LSA plot, we also encircle the pair (A7, P8.3), with light green colour, which was correctly identified by LSA but missed by fastText. In the fastText plot, points A7 and P8.3 are far away and not clustered together. We observe that semantically similar provisions are mostly clustered together in the visualization. Moreover, we can also find correspondences between similar provisions from the same legislative document (for instance NIM provisions P11, P10.2, P6.2 and P8.5 are clustered together in both plots).

## 3.7 COMPARISON OF TEXT SIMILARITY MODELS BASED ON WORD AND PARAGRAPH EMBEDDINGS WITH LEXICAL AND SEMANTIC SIMILARITY TECHNIQUES

In this section, we compare the performance of word2vec, fastText and paragraph vector text similarity models with the lexical and semantic similarity techniques discussed in section 2.3.3. Figure 3.5 presents the results of the best performing unsupervised text similarity models. We observe that TF-IDF cosine model had the best performance in terms of F-score for all three corpora. It was closely followed by LSA and USM_chars model. The lexical and semantic similarity methods outperform the word and paragraph embedding models. This is probably because a large number of transpositions can be identified by highlighting important terms using TF-IDF and modeling their relationships through LSA. The results of the embedding-based models are encouraging and probably with improvements in provision vector representation their performance may improve. There were some cases where they were able to identify the complex cases of transposition which were missed by the best performing methods.

Table 3.2 presents an example of a transposition which was identified by paragraph vector and word2vec models but missed by all other methods, such as TF-IDF cosine, USM, LSA, LDA and fastText. We observe that the NIM provision only partly implements the directive article. The second part of NIM provision talks about the proof of insurance which is not mentioned in the directive article. The NIM also does not mention anything about compliance and conformity with international law as mentioned in the directive. The proximity of word vector pairs (trained on the legal corpus), such as 'owners' and

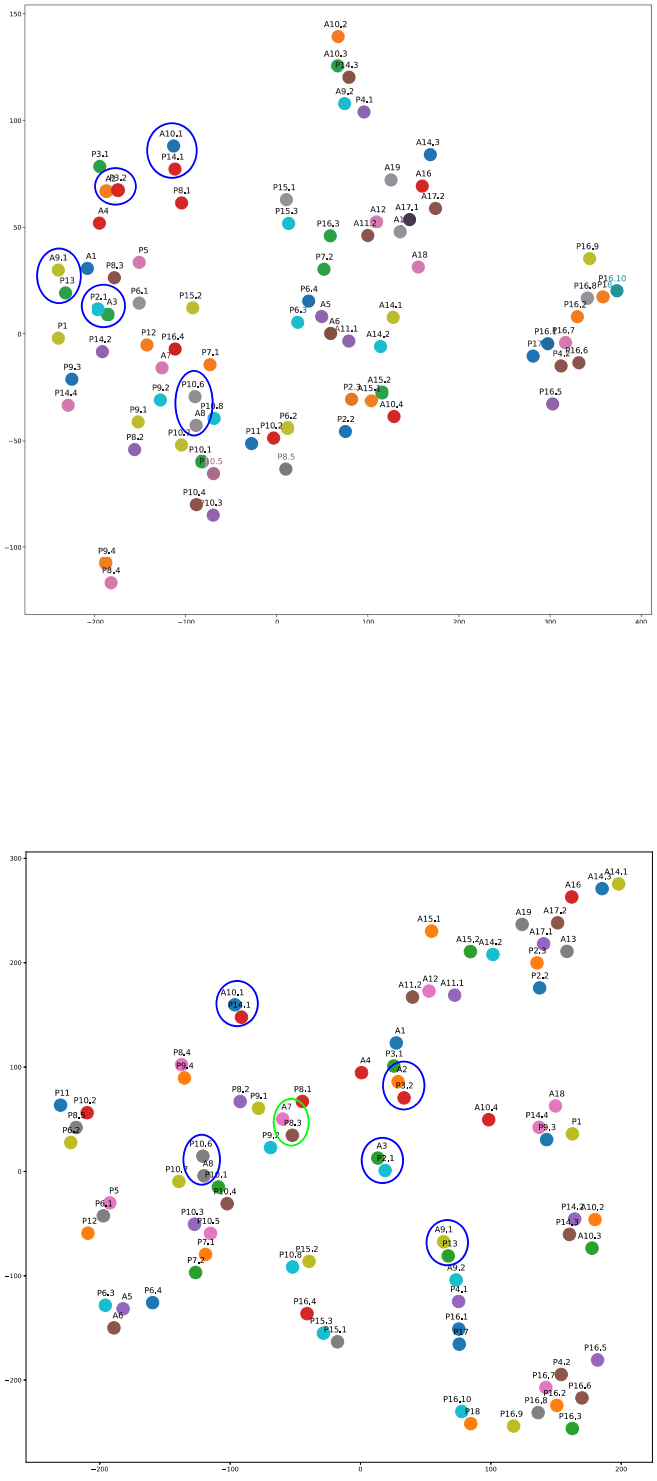

Figure 3.4: Two-dimensional visualization of fastText (top plot) and LSA (bottom plot) provision vectors using t-SNE for Directive CELEX 32001L0096 and Ireland NIM 72001L0096IRL_115977

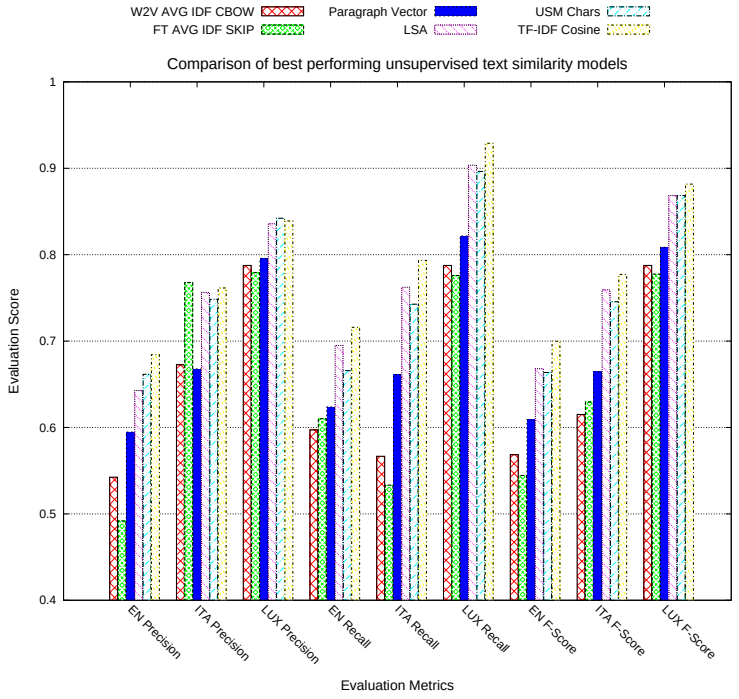


Figure 3.5: Comparison of the best performing unsupervised text similarity models

'shipowners', 'in place' and 'in force' facilitates the identification of this transposition.

Table 3.2: Article 4.2 of Directive (CELEX Number: 32009L0020) and its implementing NIM provision 4.3 from Ireland legislation (CELEX Number: 72009L0020IRL_188439)

| Article 4.2 of Directive | Provision 4.3 of Ireland NIM |
|---|---|
| Each Member State shall require shipowners of ships flying a flag other than its own to have insurance in place when such ships enter a port under the Member State's jurisdiction. This shall not prevent Member States, if in conformity with international law, from requiring compliance with that obligation when such ships are operating in their territorial waters. | The owner of a ship flying a flag other than that of the State— (a) shall have insurance in force in respect of the ship when it enters a port in the State, and (b) shall ensure that proof of such insurance in the form of a certificate or certificates referred to in Regulation 5(2) is carried on board the ship. |

## 3.8 SUMMARY

In this chapter, we presented unsupervised text similarity techniques based on word and paragraph embeddings to identify transpositions. We utilized word2vec and fastText embeddings trained on the legal corpus to develop provision vectors for text similarity. We also utilized paragraph vector model to represent provisions in a dense paragraph vector. Our results indicate that the paragraph vector model outperformed fastText and word2vec to identify transpositions. Further, we also compared the performance of word and paragraph embedding techniques with lexical and semantic unsupervised techniques such as, latent semantic analysis (LSA), unifying similarity measure (USM), latent dirichlet allocation (LDA) and TF-IDF Cosine. In the next chapter, we will study about identifying concepts in European and national legislation. We will also discuss the development of two similarity methods using word sense disambiguation and legal dictionaries.

# 4

# CONCEPT RECOGNITION IN EUROPEAN DIRECTIVES AND NATIONAL LEGISLATION

This chapter presents a concept recognition system for European directives and national legislation. Current named entity recognition (NER) systems do not focus on identifying concepts which are essential for interpretation and harmonization of European and national law. We utilized the IATE (Inter-Active Terminology for Europe) vocabulary, a state-of-the-art named entity recognition system (spaCy)and Wikipedia to generate an annotated corpus for concept recognition. We applied conditional random fields (CRF) to identify concepts on a corpus of European directives and Statutory Instruments (SIs) of the United Kingdom. The CRF-based concept recognition system achieved an F-score of 0.71 over the combined corpus of directives and SIs. Our results indicate the usability of a CRF-based learning system over dictionary tagging. We also developed a concept-based text similarity system by utilizing Babelfy and IATE. The system was used for identifying transpositions and was evaluated it on a multilingual corpus of 43 directives and their corresponding NIMs.

## 4.1 INTRODUCTION

With the increasing volume of European and national legislation available online, the identification of domain concepts in legal texts is very important for the development of legal information retrieval systems. The identification of domain concepts provides a deeper insight into the interpretation and understanding of texts. The recognition of concepts in legal texts would also be useful for the harmonization and integration of European and national law. Research in this domain has mainly focused on identification of named entities like person, organization and location names. However, European and national legislation contains very few instances of named entities. They primarily comprise legal and domain-specific jargon which can be represented by concepts.

We develop a system for concept recognition in European directives and national law (statutory instruments of the United Kingdom). We chose directives as they are not directly applicable and need to be transposed into national law. Therefore, they may have more similar concepts with the national legislation than regulations or decisions. We chose statutory instruments (SIs) from the United Kingdom (UK) national law for our experiments. Most European Union (EU) directives are transposed by statutory instruments in the UK legislation [84].

Therefore, most of the SIs comprise NIMs and may have similar domain concepts with directives. We chose the SI's from the UK because the SI's of Ireland were not available at the time of this work. However, in the concept-based similarity system (Section 4.5.1) we experimented with the Ireland NIMs (as the Ireland NIMs were available after the concept recognition system was implemented).

The concept recognition system was used for automatically identifying concepts in a corpus of 2884 directives and 2884 SIs. We generated an annotated corpus using a semi-supervised approach to save human effort and time for evaluation of our system. Further, we also generated a mapping to link similar terms in directives and SIs under the same concept.

## 4.2 CONCEPT RECOGNITION SYSTEM

In this section, we describe the concept recognition system. In the legal domain, concepts are generally represented using ontologies or vocabularies. Previous NER systems (based on the concepts represented in LKIF ontology) demonstrated that LKIF level of generalization was not suitable [18]. This is because NER systems could not clearly distinguish between the classes defined in LKIF. Therefore, in this paper we investigate the use of vocabularies for developing our concept recognition system . We utilize Inter-Active Terminology for Europe [1] (IATE), which is EU's inter-institutional terminology database. IATE is highly suitable for developing concept recognition systems because it is based on a concept-oriented approach where each term is mapped to a concept. It provides both mono-and multilingual mapping between terms and concepts and thus can also be used to develop multilingual concept recognition systems for future work. IATE has 21 subject domains with one sub-level. The EuroVoc thesaurus also offers the same 21 subject domains but with upto 6 sub-levels. Our initial hypothesis was to determine if we are able to recognize concepts and classify them to these 21 domains. For future work, we intend to utilize also the sub-domains to achieve a fine-grained hierarchical concept recognition. IATE consists of 1.3 million entries in English. Every entry in IATE is mapped to a subject domain. We filtered out some irrelevant entries in IATE (stopwords and concepts mapped to NO DOMAIN).

### 4.2.1 *Annotated Corpus Generation*

We utilized a corpus of 2884 directives and 2884 statutory instruments for our experiments. Since training data was not available, we utilized a semi-supervised approach to generate an annotated corpus. The development of NER or concept recognition systems require a large amount of manually annotated datasets, which is expensive to

1 http://iate.europa.eu

obtain [69]. We manually annotated few documents with IATE subject domains. Then we developed a dictionary lookup program to tag terms (both words and phrases) in the text with IATE subject domains. Each term in the text was compared to entries in IATE vocabulary and matching terms were tagged with the relevant subject domains. IATE consists of a mapping composed of the set of <multi-word expression, domain> pairs. We consistently improved the dictionary lookup program to match different multi-word expressions present in the IATE vocabulary. The IATE dictionary lookup program as a function $\phi$ for a document $d$ produces a set of candidate subject domains $\phi(d) = \{d_1, d_2, \ldots, d_n\}$. We also used spaCy[2], a state-of-the-art NER system to annotate some entities like time, date and money. After this tagging by IATE and spaCy, we observed that some candidate domains in the set $\{d_1, d_2, \ldots, d_n\}$ were incorrect. This is because some entries and domains in IATE vocabulary do not seem to be semantically similar and are not reliable for annotation. For instance, the term "apply" is mapped to domain "AGRICULTURE, FORESTRY AND FISHERIES". The downloaded version of IATE dictionary did not include any context information to assist our dictionary lookup program for correct annotation of documents. Therefore, we filtered out such candidate entities by using Dexter [116], a Wikipedia entity linker.[3] The application of Dexter $\psi$ on a document $d$ produced a set of Wikipedia entities $\psi(d) = \{w_1, w_2, \ldots, w_n\}$. We used them to filter the subject domains $\{d_1, d_2, \ldots, d_n\}$, taking only the domains present in $\psi(d)$. Thus, we annotated all the documents in the corpus. In the next step, each document is transformed into a collection of <word, label> pairs as input for concept recognition system. In the <word, label> pair, *word* represents a word in the document, while *label* represents the IATE subjct domain or a spaCy NER tag associated with the word. In cases when *word* does not belong to any class, *label* is assigned to 'O' tag (concept or word does not belong to any subject domain).

### 4.2.2 *Corpus Statistics*

After generating the annotated corpus for both directives and SIs we divided each dataset into 80% training and 20% test set to build the concept recognition system. Table 4.1 shows the number of documents, tokens and vocabulary size for both directive and SI datasets respectively. We observe that SIs have a much larger vocabulary than directives. Table 4.2 shows the number of tokens labeled with IATE or spaCy tags or with a 'O' tag (tokens not belonging to any class). Table 4.3 represents the number of tagged tokens for each IATE subject

2 https://spacy.io/

3 Wikipedia Entity Linkers find named entities in the text that can be linked to a Wikipedia page.

Table 4.1: Number of documents, number of tokens and the vocabulary size (|V|) for directives (left) and SIs (right)

| Dataset | # docs | # tokens | \|V\| |
|---|---|---|---|
| Train | 2,307 | 4,646,286 | 24,522 |
| Test | 577 | 1,226,338 | 14,127 |
| Total | 2,884 | 5,872,624 | 38,649 |

| Dataset | # docs | # tokens | \|V\| |
|---|---|---|---|
| Train | 2,307 | 4,189,157 | 83,172 |
| Test | 577 | 1,096,246 | 33,757 |
| Total | 2,884 | 5,285,403 | 116,929 |

domain and spaCY NER in train and test set of directive and statutory instruments.

Table 4.2: Number of tagged (IATE or spaCy tags) and untagged tokens (O tag).

| Dataset | Directives | | Statutory Instruments | |
|---|---|---|---|---|
| | IATE/spaCy Ner tags | O tag | IATE/spaCy Ner tags | O tag |
| Train | 238,929 | 4,407,357 | 169,609 | 4,019,548 |
| Test | 64,678 | 1,161,660 | 45,854 | 1,050,392 |

Table 4.3: Number of tagged tokens for IATE subject domains and named entities in directives and SIs corpus

| IATE Subject Domains and spaCy Named Entities | | | | |
|---|---|---|---|---|
| **IATE Subject Domains** | Directive Train | Directive Test | SI Train | SI Test |
| FINANCE | 16,366 | 2,838 | 10,504 | 2,564 |
| POLITICS | 3,878 | 1,138 | 10,566 | 3,136 |
| ENVIRONMENT | 8,478 | 3,294 | 3,560 | 1,045 |
| EDUCATION AND COMMUNICATIONS | 13,419 | 4,066 | 9,936 | 3,340 |
| LAW | 55,366 | 13,767 | 37,851 | 8,240 |
| INTERNATIONAL ORGANISATIONS | 269 | 60 | 98 | 22 |
| EMPLOYMENT AND WORKING CONDITIONS | 4,069 | 922 | 7,033 | 1,399 |
| AGRI-FOODSTUFFS | 3,213 | 1,066 | 1,573 | 337 |
| INDUSTRY | 17,831 | 5,946 | 11,063 | 4,170 |
| PRODUCTION, TECHNOLOGY AND RESEARCH | 9,371 | 2,462 | 5,000 | 1,724 |
| BUSINESS AND COMPETITION | 18,356 | 4,047 | 9,405 | 3,027 |
| ENERGY | 9,585 | 2,515 | 2,599 | 590 |
| TRANSPORT | 12,402 | 2,964 | 11,181 | 3,223 |
| EUROPEAN UNION | 2,449 | 699 | 969 | 249 |
| AGRICULTURE, FORESTRY AND FISHERIES | 14,085 | 4,840 | 8,832 | 3,588 |
| SOCIAL QUESTIONS | 19,531 | 5,995 | 21,878 | 5,539 |
| ECONOMICS | 3,767 | 1,095 | 2,810 | 614 |
| GEOGRAPHY | 341 | 73 | 5,325 | 740 |
| INTERNATIONAL RELATIONS | 956 | 347 | 801 | 205 |
| SCIENCE | 6,214 | 1,575 | 2,943 | 773 |
| TRADE | 18,788 | 4,886 | 5,243 | 1,217 |
| **spaCy Named Entities** | Directive Train | Directive Test | SI Train | SI Test |
| QUANTITY | 4 | 0 | 4 | 5 |
| MONEY | 2 | 0 | 6 | 4 |
| ORDINAL | 1 | 0 | 2 | 2 |
| TIME | 106 | 22 | 46 | 16 |
| DATE | 81 | 61 | 381 | 85 |
| O | 4,407,357 | 1,161,660 | 4,019,548 | 1,050,392 |

### 4.2.3 *CRF-based Concept Recognition System*

The annotated corpus for both directives and SIs was divided into train (80%) and test (20%) sets to build and evaluate the concept recognition system. We utilized conditional random fields (CRFs) to build our concept recognition system as they have been known to work well in tasks which require labeling sequence data (especially natural language text). They are discriminative probabilistic models where each observation is a token from a sentence and the corresponding label (tag of subject domain or entity) represents the state sequence. We utilize the following features for our CRF model: word suffix, word identity (whether a word represents a subject domain/named-entity or not), word shape (captialized, lowercase or numeric) and part-of-speech (POS) tags. We used the Limited-memory BFGS training algorithm with L1+L2 regularization.

## 4.3 RESULTS AND ANALYSIS

In this section, we present the results of our system. We evaluate our CRF-based concept recognition system with standard information retrieval metrics of precision, recall and F-score. We did not consider accuracy as a fair metric for evaluation because in the training data we have very different number of mentions for each class. Thus, resulting in an unbalanced dataset. But we do present the precision, recall and F-score for each class of the concept recognition system. We carried out three runs of experiments to thoroughly evaluate the CRF concept recognition system:

- Directive Corpus (2884 documents): 80% train (2307 directives) and 20% test set (577 directives)
- SI Corpus (2884 documents) : 80% train (2307 SIs) and 20% test set (577 SIs)
- Combined Corpus (5768 documents) : 80% train (2307 directives + 2307 SIs) and 20% test set (577 directive + 577 SIs)

Table 4.4 reports the F-score of our CRF-based concept recognition model for each subject domain and entity class. We observe that all IATE subject domains are clearly distinguished due to the achievement of a reasonable F-score for each domain for each corpus. The low F-score of domain 'INTERNATIONAL ORGANISATIONS' is explained by a smaller number of tagged tokens, resulting in only a few training instances (as observed from Table 4.3). The other subject domains had sufficient training data and therefore were classified with a higher F-score. We also observe that CRF could not identify classes, 'QUANTITY','MONEY' and 'ORDINAL' of spaCY named entities because there were hardly any training instances for these classes (as observed

Table 4.4: Results (F-score) for concept recognition for each class by CRF-based concept recognition system

| Tag name | Directives | SIs | Directives + SIs |
|---|---|---|---|
| **IATE Subject Domains** | | | |
| FINANCE | 0.68 | 0.62 | 0.62 |
| POLITICS | 0.70 | 0.74 | 0.71 |
| ENVIRONMENT | 0.68 | 0.41 | 0.66 |
| EDUCATION AND COMMUNICATIONS | 0.68 | 0.72 | 0.71 |
| LAW | 0.92 | 0.81 | 0.89 |
| INTERNATIONAL ORGANISATIONS | 0.52 | 0.14 | 0.32 |
| EMPLOYMENT AND WORKING CONDITIONS | 0.70 | 0.68 | 0.70 |
| AGRI-FOODSTUFFS | 0.75 | 0.73 | 0.68 |
| INDUSTRY | 0.67 | 0.45 | 0.60 |
| PRODUCTION TECHNOLOGY AND RESEARCH | 0.69 | 0.67 | 0.69 |
| BUSINESS AND COMPETITION | 0.78 | 0.77 | 0.77 |
| ENERGY | 0.81 | 0.50 | 0.74 |
| TRANSPORT | 0.59 | 0.60 | 0.58 |
| EUROPEAN UNION | 0.79 | 0.77 | 0.76 |
| AGRICULTURE FORESTRY AND FISHERIES | 0.70 | 0.58 | 0.64 |
| SOCIAL QUESTIONS | 0.68 | 0.65 | 0.66 |
| ECONOMICS | 0.66 | 0.57 | 0.68 |
| GEOGRAPHY | 0.52 | 0.76 | 0.75 |
| INTERNATIONAL RELATIONS | 0.70 | 0.59 | 0.59 |
| SCIENCE | 0.60 | 0.48 | 0.59 |
| TRADE | 0.77 | 0.66 | 0.76 |
| **spaCy Named Entities** | | | |
| QUANTITY | 0.00 | 0.00 | 0.00 |
| MONEY | 0.00 | 0.00 | 0.00 |
| ORDINAL | 0.00 | 0.00 | 0.00 |
| TIME | 0.62 | 0.00 | 0.60 |
| DATE | 0.00 | 0.19 | 0.47 |

from Table 4.3).'TIME' and 'DATE' classes also had very few training instances, thus resulting in a lower F-score. These results also indicate that European and national legislation consists of very few named entities and are therefore more suited for concept recognition. The average F-scores of our CRF-based concept recognition system for directive, SI and combined corpus were 0.75, 0.66 and 0.71 respectively (Table 4.5). The lower average F-score for SI corpus is probably due to the larger vocabulary size of the SI corpus (Table 4.1). A larger vocabulary implies more diversity in the tokens assigned to each domain, thus also leading to fewer training instances. We also compare the performance of CRF with a baseline method (Most frequent class model). It is a simple model which computes the most frequent class assigned to each token in the training set, and it uses them to tag the new documents. If a word is not present in the training set, it assigns it to the class 'O'. We observe that CRF outperforms the baseline model. This is because the baseline model does not take into account the

Table 4.5: Results of concept recognition with CRF model and comparison with baseline (Most Frequent Class) and Stanford NER model

| Corpus | System | Precision | Recall | F-score |
|---|---|---|---|---|
| Directive Corpus | Most frequent class | 0.74 | 0.53 | 0.61 |
| | CRF | 0.80 | 0.71 | 0.75 |
| | Stanford NER | 0.80 | 0.71 | 0.75 |
| SIs Corpus | Most frequent class | 0.61 | 0.40 | 0.48 |
| | CRF | 0.73 | 0.61 | 0.66 |
| | Stanford NER | 0.68 | 0.53 | 0.59 |
| Combined Corpus (Directives + SIs) | Most frequent class | 0.66 | 0.47 | 0.54 |
| | CRF | 0.76 | 0.68 | 0.71 |
| | Stanford NER | 0.72 | 0.6 | 0.65 |

context information for a particular token while assigning it to a class. We also compared CRF with Stanford NER for both the Directive Corpus and SIs corpus. The CRF model had similar performance to Stanford NER in the directive corpus. However it outperformed the Stanford NER in the SIs corpus and the combined corpus by achieving a higher F-score. One of the drawbacks of Stanford NER is the large amount of training time required (several days). The CRF-based concept recognition system utilizes few important features and completes training under an hour.

### 4.3.1 *Discussion*

In this section, we discuss the advantages of developing and training a CRF over using a dictionary lookup program to automatically detect concepts. One drawback of using dictionary tagging to annotate corpus is that some terms are missed and not tagged due to inconsistent rules to accommodate different phrases or tokenization errors. Also since the downloaded IATE dictionary did not provide any context information, we could not utilize context information to assign the correct tag to a particular token in the corpus. CRFs on the other hand, use contextual information to learn and assign tags because of their Markov property. Therefore, CRF models have the potential to reduce false positives and false negatives in the dictionary lookup tagging. In IATE dictionary, an entry, 'integrated energy performance' is linked to subject domain, 'INDUSTRY'. Table 4.7 presents an example sentence with tagged labels of IATE dictionary and predicted CRF labels. CRF classifies both 'energy' and 'performance' to 'INDUSTRY' subject domain whereas dictionary missed them. This is because dictionary lookup utilizes state-of-the-art tokenizers which are not 100% accurate and may lead to incorrect tokenization resulting in a mismatch. Most other multi-worded phrases like 'national regulatory authority', 'Federal Motor Transport Authority' and several others were tagged correctly by dictionary. CRF on the other hand, had some training instances (as shown in Table 4.6) from which it learns that terms 'energy'

Table 4.6: Relevant training instances for CRF

| Terms | IATE subject domains |
|---|---|
| seasonal energy performance ratio | INDUSTRY |
| energy performance diagnosis | INDUSTRY |

and 'performance' are related to 'INDUSTRY'. Thus it was able to correctly classify them. Thus, training a CRF model is advantageous also on automatically annotated corpora because it can improve the tagging of dictionary by learning these semantic relations between terms and subject domains. Thus, it can be used to improve the quality of annotations and develop a better gold standard for further work.

Table 4.8 shows an example phrase from the SI corpus to compare the performance of CRF with the Most frequent class model, Stanford NER and the true labels (from the dictionary). The label 'EMP' here refers to the subject domain, 'EMPLOYMENT AND WORKING CONDITIONS' from the IATE dictionary. We observe that CRF correctly classifies all of the labels. The Most frequent class (baseline) model could classify only word 'sick' correctly and missed out on 'statutory' and 'pay'. This is because in the training set there only few instances of words 'statutory' and 'pay' for 'EMP' class and most instances are for 'O' class. The term 'statutory' had 324 instances of 'O' class while only 31 instances of 'EMP' class. Therefore 'O' class was the most frequent class. We also observe that Stanford NER correctly classified 'statutory' and 'sick'. However, the words 'the' and 'pay' were incorrectly classified.

## 4.4 ALIGNMENT OF SIMILAR TERMS ACROSS DIRECTIVE AND SIS

In order to utilize the concept recognition system, it is important to align similar terms across European and national law. This semantic alignment of terms is highly useful for legal professionals to understand the differences in terminologies at the European and national level. It is also beneficial for development of other legal information systems which utilize this semantic information. The concept recognition system tags each term in the text to a particular subject domain. As a result we have a large collection of terms under each subject domain from both directives and statutory instruments. We divided the terms under each subject domain into two lists : directive terms and SI terms. We computed the set difference of these two lists to obtain a list of terms present in directive but not in SIs. Similarly, we also obtained a list of terms present in SIs but not in directives. Then we computed text similarity (using Levenshtein distance) to find the most semantically similar term in SIs (not present in directive) for a particular term in directive. Table 4.9 shows a few examples

Table 4.7: Comparison of CRF output with the dictionary tagging

| | CRF predicted labels | Dictionary |
|---|---|---|
| The | O | O |
| general | O | O |
| framework | O | O |
| for | O | O |
| a | O | O |
| methodology | ECONOMICS | ECONOMICS |
| of | O | O |
| calculation | O | O |
| of | O | O |
| the | O | O |
| integrated | O | O |
| **energy** | **INDUSTRY** | **O** |
| **performance** | **INDUSTRY** | **O** |
| of | O | O |
| buildings | O | O |

Table 4.8: An example phrase to compare different models against the true values

| | Most frequent class | Stanford NER | CRF | True Labels |
|---|---|---|---|---|
| the | O | EMP | O | O |
| statutory | O | EMP | EMP | EMP |
| sick | EMP | EMP | EMP | EMP |
| pay | O | O | EMP | EMP |
| up | O | O | O | O |

of such terms. Figure 4.1 shows the first example from Table 4.9. The SIs use the term "vocational qualification" instead of "professional qualification" which is used in the directives.

Table 4.9: Aligned terms from European and national law

| Subject Domain | Aligned terms (*Directive→SIs*) |
|---|---|
| EMPLOYMENT AND WORKING CONDITIONS | *professional qualification→vocational qualification*<br>*seniority→job security*<br>*occupational disease→industrial disease* |
| FINANCE | *life assurance→endowment assurance*<br>*financial institution→financial administration*<br>*dividend→tax on dividends* |

Figure 4.1: An example of aligned terms under the same subject domain (Employment and Working Conditions): professional qualification (from directives) and vocational qualification (from SIs)


## 4.5 CONCEPT AND WORD-SENSE DISAMBIGUATION-BASED TEXT SIMILARITY SYSTEM FOR IDENTIFYING TRANSPOSITIONS

In this section, we present a text similarity system which utilizes word-sense disambiguation and concept recognition from IATE dictionary. We utilize Babelfy for word-sense disambiguation [87]. Babelfy provides a unified framework to carry out both word-sense disambiguation and entity linking. The next two paragraphs discuss about word sense disambiguation and entity linking.

Word-sense disambiguation involves identifying the correct sense of the word depending on the context in which it is used. Many words have multiple meanings. In such cases, the correct meaning (or sense) of the word can be determined by its context. Word-sense disambiguation has been addressed by supervised, unsupervised and knowledge-based approaches. The supervised approaches utilize labelled training data (manually annotated sense dataset) to build machine learning classifiers. The classifier utilizes the linguistic features from the context to predict the sense for a particular word. Supervised disambiguation methods suffer from the knowledge acquisition bottleneck due to a lack of availability of large manually annotated datasets [42, 93]. Unsupervised methods are based on the assumption that words with similar senses have similar context (surrounding words). The sense of a word is then inferred from the input text by clustering word

occurrences. The new occurrences of a word are assigned to one of the existing clusters. The unsupervised methods do not required labelled training data and do not have a defined set of senses for a particular word (as they do not utilize a dictionary). Therefore, they assign the occurrences of a word into separate clusters (or classes), by estimating whether any two occurrences have the same sense or not [93]. The knowledge-based disambiguation approaches rely on resources such as dictionaries and thesauri to determine the appropriate sense of a word. Some knowledge-based methods utilize the overlap of sense definitions to disambiguate a pair of words [67]. The dictionary senses which have the highest overlap are chosen to be the correct disambiguated senses. Other knowledge-based approaches utilize semantic similarity and graph-based methods by exploiting the structure of semantic networks such as WordNet.

Entity linking involves identifying the named entity mentions in the text and then matching them to an entry in the knowledge base. Therefore, an entity linking system utilizes the context of the named entity mention and information about the entity from the knowledge base to link a mention to an appropriate entry (in the knowledge base) [108]. Word-sense disambiguation is quite similar to entity linking. In case of word-sense disambiguation, a word is linked to a sense (present in a semantic network or sense inventory such as WordNet, instead of a knowledge base). Entity linking systems typically use name-dictionary based approaches to generate candidate entities for the mention by filtering out irrelevant entities in the knowledge base. They also utilize surface form expansions techniques and supervised machine learning classifiers to link acronym mentions to entities. The candidate entities obtained after this step have to be ranked to determine the best match for the mention. Both supervised and unsupervised ranking approaches can be used to rank the candidate entity mentions. The supervised approaches required labelled training data to build a machine learning classifier to rank the candidate entities. The unsupervised approaches utilize information retrieval models based on vector space models [108].

Babelfy provides a common framework for both word-sense disambiguation and entity linking. It utilizes BabelNet (a multilingual semantic network developed by integrating Wikipedia and WordNet) [94]. BabelNet consists of both concepts and named entities as vertices. The edges represent the semantic relations between the vertices. Each vertex of the BabelNet semantic network is associated with an entity or concept or other related vertices (also known as semantic signatures). For a particular input text, part-of-speech (POS) tagging is applied to determine the potential textual fragments (candidates) which can be linked to an entry in the BabelNet semantic network. The words and phrases in the textual fragments are then mapped to their candidate meanings using the vertices of the semantic network. The candidate

meanings of the textual fragments are linked together by using the semantic signatures. This results in a holistic graph-based semantic representation of the whole text. Further, a dense subgraph representation is obtained to reduce ambiguity and link the best candidate mapping for each textual fragment [87].

### 4.5.1 *Text Similarity measure using Babelfy and IATE*

We queried the Babelfy Java API (application programming interface) for semantic annotation of the multilingual parallel coprus of 43 directives and their corresponding NIMs. Figure 4.2 shows the NLP pipeline to process a particular provision using Babelfy and IATE. We consider a sample provision from one of the directives in the corpus. This provision is used to query the Babelfy API. The API returns the disambiguated text (with both concepts and named entities). In our example, we do not have any named entities. The text retrieved from the Babelfy API consists of babelfy identifiers appended to the disambiguated words and phrases. In our example, directive is linked with identifier bn_:00893324n. This identifier refers to the directive of European Union in BabelNet and also refers to the directive entry in Wikipedia. The last letter, *n* of the identifier indicates Noun, the POS tag. The word auxiliary is correctly disambiguated as meaning , 'functioning in a supporting capacity'. Other words also assigned to the same identifier include 'subsidiary' and 'supplementary'. The next step involves tagging the remaining words and phrases with IATE using the dictionary lookup program.

We utilized the fastText embeddings trained on the legal corpus to remove some of the invalid entries in IATE dictionary. We compute the vector representation of both terms and concepts in the IATE dictionary using the fastText embeddings. The embeddings of multi word expressions were computed by summing the embeddings of individual words and dividing them by the number of words. A sample gold standard of correct and incorrect dictionary entries was created manually. The cosine similarity values between term and concepts was computed. We computed the accuracy for different threshold values (the entries with similarity values greater than or equal to the threshold were classified as correct). The best threshold was then used to label correct and incorrect entries in the entire dictionary of IATE. The entries with incorrect labels (low similarity values) were removed from the IATE dictionary.

In the given example, the phrase 'public service' was tagged with the domain concept, 'LAW' from IATE. The identifier, 'IE' has been added as a prefix to the IATE domain concepts. In the next step we utilize regular expressions and tokenization to replace words by their Babelfy or IATE concept identifiers. Stopwords are removed. The words which have not been linked to either Babelfy or IATE are left

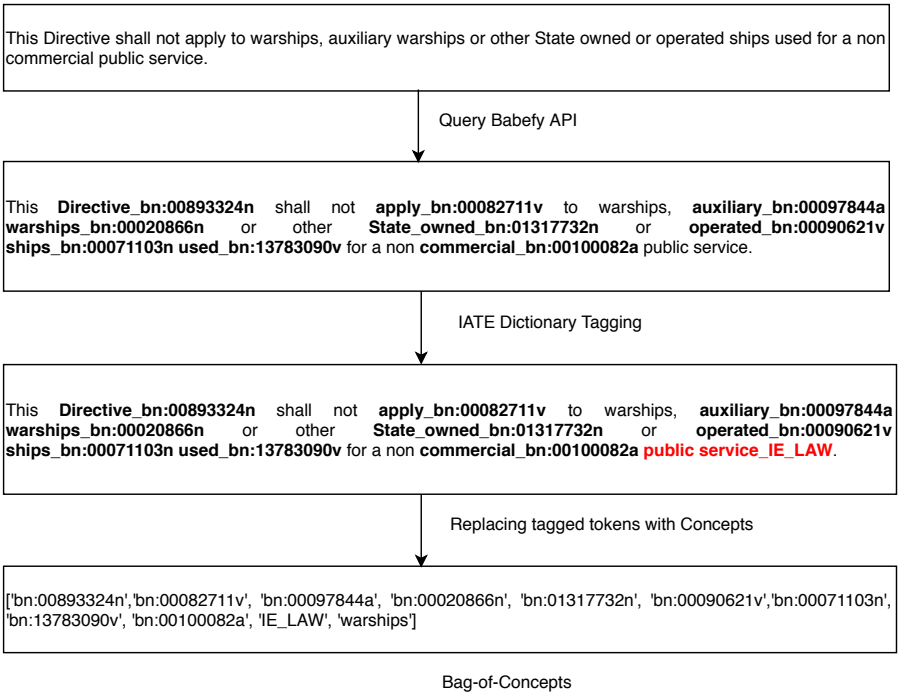


Figure 4.2: NLP pipeline for producing bag-of-concepts representation of legal provisions

as it is. Eventually, we have a bag-of-concept identifiers which is an improved semantic representation over the bag-of-words model. Then we applied the TF-IDF transform to the bag-of-concepts as per the vector space model. A cosine similarity measure is computed between a directive article and all the NIM provisions to retrieve the most semantically similar provision.

The bag-of-concepts provision vectors were computed for the entire multilingual corpus of 43 directives and their corresponding NIMs. This was possible because of the multilingual entries in both Babelfy and IATE. We develop two versions of the similarity system. The first version uses only Babelfy to tag concepts. The second version uses both Babelfy and IATE as explained in Figure 4.2. We present the macro average precision, recall and F-score for the concept-based similarity system on the multilingual corpus of 43 directives and their corresponding NIMs (Figure 4.3). The results indicate that the Luxembourg Directive-NIM corpus has the best performance. This is consistent with the previous results from other unsupervised text similarity systems. We also observed that the Babelfy-based similarity measure outperforms the similarity measure with the combination of Babelfy and IATE. This illustrates that the concepts associated from IATE are not very relevant for text similarity. Some of the term-concept entries in the IATE dictionary are not semantically related. The use of a similarity measure based on fastText embeddings to remove some of the unrelated entries in the IATE dictionary was partially successful.

In the future work, we will resort to manual cleaning of the IATE dictionary for improving the results.

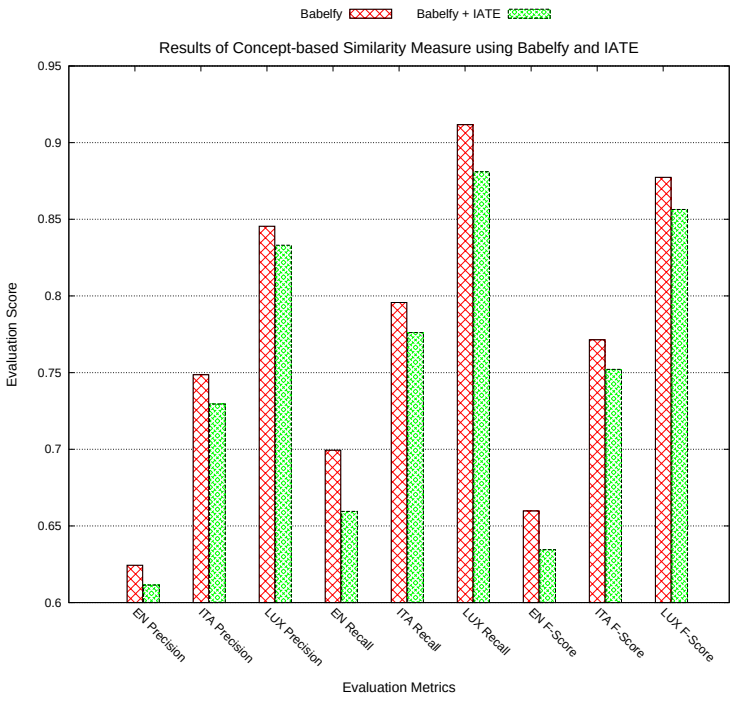


Figure 4.3: Results of the concept-based similarity using Babelfy and IATE

## 4.6 SUMMARY

In this chapter, we developed and evaluated a CRF-based concept recognition system for European and national law. We generated a labeled corpus of directives and statutory instruments with subject domains of IATE vocabulary, Wikipedia and a state-of-the-art named entity recognition system. We evaluated the system on both European and national law corpus and analyzed its performance with respect to a baseline model and Stanford NER. Our results indicate that the concept recognition system is able to identify concepts in both directives and UK statutory instruments with a F-score of 0.71 over the combined corpus. It can also be used to iteratively improve the dictionary lookup tagging from IATE. We demonstrated that concept recognition systems are useful to align legal terminology at European and national level to assist legal practitioners and domain experts. Further, we also developed concept-based similarity measures using Babelfy and IATE dictionary. We evaluated the concept-based similarity measures on the multilingual corpus of 43 directives and their corresponding NIMs. In the next chapter, we investigate the application of supervised text similarity models for identifying transpositions of European directives.

# 5 SUPERVISED TEXT SIMILARITY MODELS

In the previous chapter, we presented a concept recognition system for European directives and national legislation. We also proposed a concept and word-sense disambiguation based text similarity system to identify transpositions. In this chapter, we use supervised machine learning approaches for automated identification of national implementing measures (NIMs). We utilize the labeled training data from the multilingual corpus of 43 directives and their corresponding NIMs. We evaluate the performance of machine learning classifiers with different textual features.

## 5.1 MODELING TEXT SIMILARITY AS A SUPERVISED MACHINE LEARNING TASK

In this section, we utilize supervised machine learning approaches to identify semantically similar legal provisions. The techniques discussed in previous chapters are unsupervised as they utilize an unlabeled dataset. The objective is to find the transposing NIM provisions for a particular article of the directive. We utilize the labeled training data from the gold standard for this purpose. If a directive article, A is transposed by a NIM provision, P then they are considered to be similar provisions (represented by "True" label). The provisions which are not similar are represented by the "False" label. The "False" label also implies that the NIM provision, P does not transpose the directive article, A. Therefore, this is a binary classification problem with two classes, "True" and "False". We select an equal number of "True" and "False" label pairs from the corpus to develop a balanced dataset. Both "True" and "False" label pairs were selected from the intersection set of both annotators. Table 5.1 shows the format of the dataset used for this classification task. The directive articles A and NIM provisions P represent the text of each article and provision respectively. A machine learning classifier is then trained on the labeled training dataset. The classifier is evaluated by comparing its predictions with the ground truth on a test set. In the next section, we discuss the supervised machine learning models.

## 5.2 SUPERVISED MACHINE LEARNING

Machine learning classifiers have been widely used in text categorization and textual entailment tasks [107, 123]. They utilize labeled training data which consists of sample inputs with known outcomes

Table 5.1: Dataset format for supervised classification of provisions

| Directive Article | NIM Provision | Transposition |
|---|---|---|
| A1 | P1 | True |
| A2 | P2 | True |
| A3 | P3 | False |
| A4 | P4 | True |
| ..... | ..... | ..... |
| A101 | P43 | Classifier Predicts ? True/False |

(class labels). The classifier consists of a function to map a set of input features to a target variable. The target variable represents the set of output class labels. The classifier models the relationship between the input features and target labels by learning the classification function from the labeled data. The learned model is then used to predict the outcomes of new (unseen) instances. Machine learning classifiers are thus data-driven models. The input features can be statistically computed measures or manual rules inferred from the data or an expert. In this chapter, we utilize three machine learning classifiers to identify transpositions. These include Naive Bayes, Logistic Regression and Support Vector Machines.

### 5.2.1 *Naive Bayes Classifier*

Bayesian classifiers have been quite effective to predict the outcome of uncertain events. They are based on the Bayes' theorem which provides a methodology to update the prior beliefs by taking into account the new evidence. Bayesian classifiers are modeled through a Bayesian network and are used for developing probabilistic reasoning systems [102]. In order to understand the probabilistic inference of Naive Bayes classifier, we first need to study the Bayes' theorem.

The conditional probability for an event $B$, given the occurrence of event $A$, is given as $P(B|A)$. It is defined as:

$$P(B|A) = \frac{P(B \cap A)}{P(A)} \tag{5.1}$$

The prior probability of a hypothesis $H$ represents how likely is the occurrence of $H$ without any evidence $E$. It is represented by $P(H)$. Bayes' theorem computes the conditional probability of the hypothesis $H$, given the evidence $E$ as follows :

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)} \tag{5.2}$$

Bayesian networks are directed acyclic graphs (DAGs). They represent the probabilistic relationship between the nodes. The nodes

represent continuous or discrete random variables. The probabilistic conditional dependency between the nodes is represented by the edges. A directed edge from node $A$ to node $B$ implies that $A$ has a direct influence on $B$. It also indicates that $A$ is the parent of $B$. The conditional probability distribution for a random variable $A_i$ is represented as $P(A_i|Parents(A_i))$. Therefore, the conditional probability distribution for node $A$ with parent $B$ is represented as $P(A|B)$. We should also note that a node in a Bayesian network is conditionally independent of all the other nodes, except the parent nodes.

The Naive Bayes classifier assumes conditional independence among the features for a given class. Even then it has shown state-of-the-art performance with other classifiers in machine learning classification tasks [41, 64]. In the classification task, each input feature thus contributes individually to the classification result. The Naive Bayes classifier is a Bayesian network with one parent node and atleast one child node. The child nodes are conditionally independent with respect to the parent node. Figure 5.1 represents a Naive Bayes network with class label $C$ and child nodes $X_1$, $X_2$, ... $X_n$ as features. Bayes' theorem is utilized to compute the probability of the class label $C$, for each of the features $X_1$, $X_2$, ... , $X_n$.

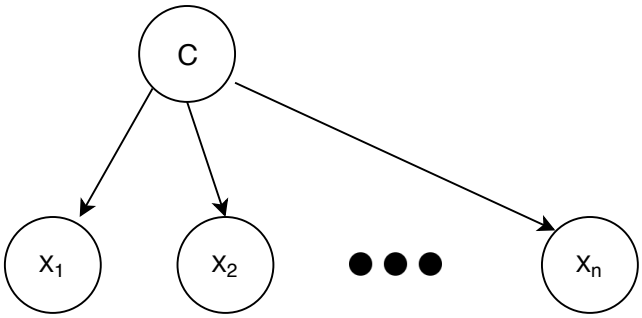


Figure 5.1: Naive Bayes Classifier

The conditional probability for a class label $C$ given a feature $X_1$ is given by the following equation:

$$P(C|X_1) = \frac{P(C)P(X_1|C)}{P(X_1)} \tag{5.3}$$

The joint probability distribution of the Bayesian network with the conditional independence assumption is defined as:

$$P(C, X_1, X_2, ..., X_n) = P(C)P(X_1, X_2, ..., X_n|C) \tag{5.4}$$

$$P(C, X_1, X_2, ..., X_n) = P(C)P(X_1|C)P(X_2|C)...P(X_n|C) \tag{5.5}$$

Assuming that the class label $C$ has $m$ possible outcomes such that $m = 1, 2, ..., M$, the previous equation for the joint probability distribution of the Naive Bayes network can be written as:

$$P(C_m, X_1, X_2, ..., X_n) = P(C_m) * \prod_{i=1}^{n} P(A_i|C_m) \tag{5.6}$$

The hypothesis with the maximum probability for the class label $C_m$ is chosen by Naive Bayes network.

$$y = \arg \max_{m \epsilon \{1,...,M\}} P(C_m) * \prod_{i=1}^{n} P(A_i|C_m) \tag{5.7}$$

### 5.2.2 *Logistic Regression*

Logistic Regression has been widely used in machine learning problems with discrete output variables. It can also be used to model the input variables which are not continuous. Logistic regression utilizes the natural logarithm function to model the relationship between a binary dependent variable and an independent variable $X$. This relationship is defined as follows:

$$P = \frac{e^{a+bX}}{1 + e^{a+bX}} \tag{5.8}$$

Here, $P$ refers to the probability that $y = 1$, which implies that event $y$ occurs, $e$ is the base of natural logarithm and $a$ and $b$ are parameters of the model. The probability function thus represents a non-linear relationship between $X$ and $P$ [15]. The above equation can be simplified as follows:

$$P = \frac{1}{1 + e^{-(a+bX)}} \tag{5.9}$$

Logistic regression utilizes odds to compute the probability. The odds are given by the ratio of the probability of the occurrence of an event to its non-occurrence. They are defined by the following equation:

$$odds = \frac{P}{1 - P} \tag{5.10}$$

The dependent variable is defined as a logit by taking the natural log of the odds as follows:

$$logit(P) = log(odds) = \ln \frac{P}{1 - P} \tag{5.11}$$

The logit function of probability is linear with respect to the independent variable, $X$. It is given as:

$$logit(P) = a + bX \tag{5.12}$$

From equations 5.11 and 5.12, we have:

$$\ln \frac{P}{1-P} = a + bX \tag{5.13}$$

The above equation results into equation 5.8, which represents a non-linear relationship between $X$ and $P$.

### 5.2.3 *Support Vector Machines*

Support vector machines (SVMs) are supervised machine learning models which can be used for both classification and regression. They learn a hyperplane which can be used in classification and regression tasks. The SVM classifier constructs an optimal hyperplane by maximizing the margin between two classes. The vectors that are used to define the hyperplane are referred to as support vectors. The following function $f(x)$ is used to define the hyperplane:

$$f(x) = \beta_0 + \beta^T x \tag{5.14}$$

where $\beta$ is the weight vector and $\beta_0$ is the bias. A number of different hyperplanes can be obtained by scaling the values of $\beta$ and $\beta_0$. Out of all the possibilities of hyperplane representations, the following one is chosen:

$$|\beta_0 + \beta^T x| = 1 \tag{5.15}$$

where $x$ represents the training examples closest to the hyperplane. The distance $d$, between the hyperplane $(\beta, \beta_0)$ and the point $x$ is given as follows:

$$d = \frac{|\beta_0 + \beta^T x|}{||\beta||} \tag{5.16}$$

where $||\beta|| = \beta^T \beta$. This particular representation is called as canonical hyperplane [40]. Substituting equation 5.15 into equation 5.16, we obtain the distance to the support vectors as follows:

$$d = \frac{1}{||\beta||} \tag{5.17}$$

The margin, $M$ is given as twice the distance to the closest examples:

$$M = \frac{2}{||\beta||} \tag{5.18}$$

The maximization of the margin $M$ is achieved by minimizing the function $L(\beta)$ subject to certain constraints as follows:

$$\min_{\beta,\beta_0} L(\beta) = \frac{1}{2}||\beta||^2 \quad \text{subject to } y_i(\beta^T x_i + \beta_0) \geq 1 \quad \forall i \tag{5.19}$$

where $y_i$ represents the class labels from the training set. The requirement for the SVM hyperplane to classify correctly the training examples, $x_i$ is modelled by the constraints in the equation 5.19.

## 5.3 SUPERVISED MACHINE LEARNING MODELS FOR IDENTIFYING TRANSPOSITIONS

In this section, we discuss the implementation of machine learning classifiers for identifying transpositions. We also evaluate their performance on the multilingual corpus of directives and their corresponding NIMs. The directive articles and NIM provisions are first passed through the NLP pre-processing pipeline as discussed in Section 2.2.2. We utilize TF-IDF vectors for feature extraction. The dataset was divided into 80% training and 20% test set. We utilized the Mutltinomial Naive Bayes classifier as the baseline model. Figure 5.2 presents the results of the Multinomial Naive Bayes classifier to identify both similar ("True") and not similar ("False") provisions. The overall precision and recall (represented by Average) for both classes is computed as

$$weighted\_precision = \frac{P_T * |T| + P_F * |F|}{|T| + |F|} \tag{5.20}$$

$$weighted\_recall = \frac{R_T * |T| + R_F * |F|}{|T| + |F|} \tag{5.21}$$

where, $P_T$ and $P_F$ are the precision values for class True and False, and $|T|$ and $|F|$ are the number of instances in True and False class. The weighted recall is also computed in a similar way as per equation 5.21. We observe that the English Directive-NIM corpus achieves the highest precision, recall and F-score. The results indicate that Naive Bayes classifier is quite effective in differentiating both True and False class labels across all the three legislations.

We further evaluated logistic regression, support vector machines (SVM), multinomial Naive Bayes and an ensemble classifier over 10-folds cross-validation using TF-IDF features. The ensemble classifier is a voting classifier which is used to combine conceptually different

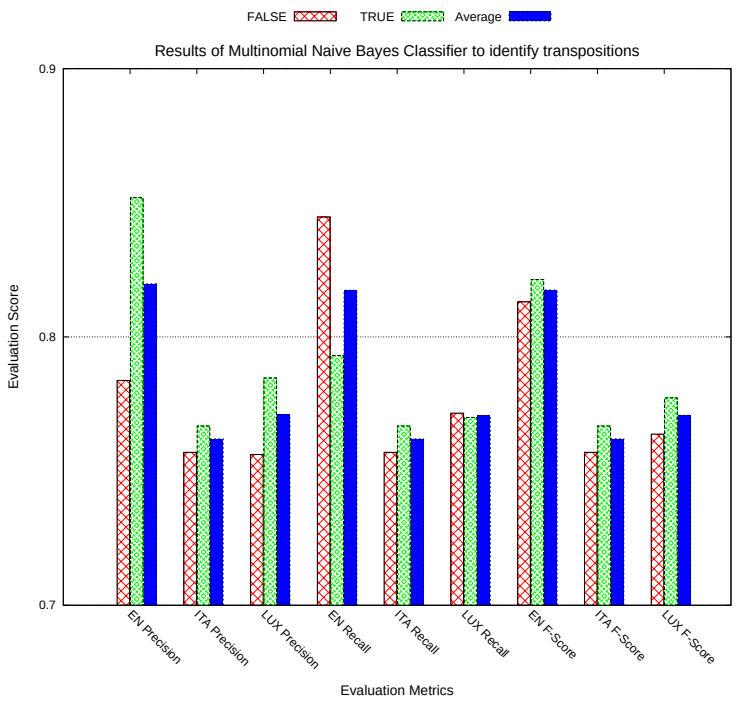


Figure 5.2: Results of Multinomial Naive Bayes to identify transpositions

machine learning classifiers [98]. A majority vote is used to decide the predicted class label. Figure 5.3 presents the results (weighted average values of precision, recall and F-score over both class labels) of different classifiers on the multilingual corpus. The results indicate that SVM classifier has the best performance in Italian and English legislation. This result is consistent with previous findings where SVM has been shown to outperform other classifiers for text classification ([54]). In case of Luxembourg legislation, the ensemble classifier outperforms other classifiers. The F-score values (for Luxembourg corpus) of logistic regression and SVM are comparable and we observe the benefit of using the ensemble classifier in this case.

We utilized the SVM classifier to experiment with different textual features due to its overall good performance over the multilingual corpus. We used latent semantic analysis, LSA and latent dirichlet allocation, LDA vectors as features for the classifier. A feature union of LSA and LDA features was also used. The feature vectors from LSA and LDA transforms are extracted individually and are then concatenated into a single transform. Figure 5.4 presents the results of the SVM classifier with different features for 10-folds cross-validation. The results indicate that TF-IDF + SVM outperforms LSA+SVM, LDA+SVM and (LSA+LDA) Feature Union + SVM. This also corroborates the results of the unsupervised methods where TF-IDF Cosine had the best performance.

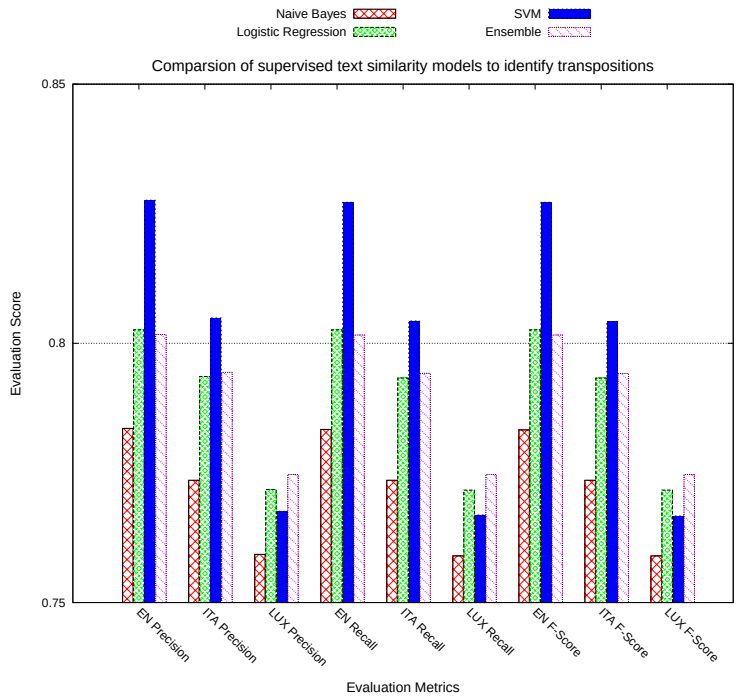


Figure 5.3: Comparison of different machine learning classifiers over 10-folds cross-validation

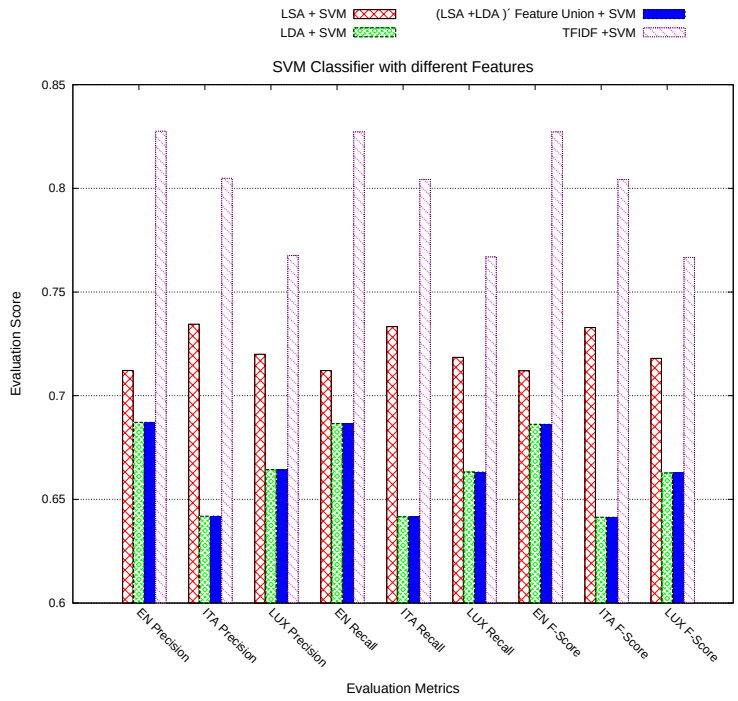


Figure 5.4: The performance of SVM classifier with different features

## 5.4 SUMMARY

In this chapter, we modeled text similarity as a supervised classification task. We studied three different machine learning classifiers and utilized them to identify transpositions over a multilingual corpus of 43 directives and their corresponding NIMs. The classifiers used different textual features based on vector transforms such as TF-IDF, LSA and LDA. The results over 10 folds cross validation indicate that TF-IDF + SVM classifier had the best performance over the multilingual corpus. In the next chapter, we discuss the related work.

# 6 RELATED WORK

The major motivation for this chapter is to present research works which have made an attempt to automate different legal tasks by developing information retrieval systems mainly based on text similarity, machine learning and concept-based techniques. These three areas are chosen as they are relevant to the scope of this thesis. Chapters 2 and 3 are based on text similarity techniques to identify transpositions. Chapter 5 discusses the machine learning techniques for automated identification of national implementing measures (NIMs). Chapter 4 presented a concept recognition system and concept-based text similarity techniques for identification of transpositions.

The prevalent methodology in legal practice relies on the skills and capabilities of legal professionals. The works presented in this chapter automate or semi-automate certain legal tasks by retrieving relevant information. Section 6.1 presents the related work on the use of text similarity techniques in the legal domain. Section 6.2 presents a review of machine learning techniques for legal information retrieval and prediction tasks. Section 6.3 discusses different concept and ontology-based methods for legal information retrieval. Section 6.4 presents a summary of this chapter.

## 6.1 TEXT SIMILARITY TECHNIQUES

In this section, we discuss the state-of-the-art methods for text similarity in the legal domain. We identified different areas where text similarity techniques have been utilized in legal information retrieval systems. The following subsections present the relevant research works in different areas such as retrieval of similar cases and judgments, retrieval of similar patents, legal question answering, legal statutes and provisions retrieval, automated conflict and dispute resolution, contracts compliance check and trademark retrieval.

### 6.1.1 *Retrieval of Similar Cases and Judgments*

Mandal et al. [77] utilized different text similarity measures to identify similar court cases from the Indian Supreme Court. The legal case documents were utilized for text similarity by selecting four different representations: whole document, document summary, paragraphs and the reason for citation (the text surrounding the citations to other cases). They implemented four models for document similarity : TF-IDF, word2vec, latent dirichlet allocation (LDA) and doc2vec (also

known as paragraph vectors). The results demonstrate that doc2vec outperforms other models in case of whole documents. This is because doc2vec is the only model in their implementation which captures the word order to some extent [65]. In case of paragraphs, both word2vec and doc2vec had similar performance and outperformed other methods. The document vectors were computed as the weighted average of the word embeddings (each embedding being weighted by the TF-IDF score of the word). In case of document summary, the results are highly dependent on the performance of the summarization algorithm. But LDA achieved a better result than other other methods, closely followed by TF-IDF. The overall results indicate that the doc2vec similarity measure over the entire document had the highest semantic correlation with the legal expert opinion. This was demonstrated by a higher pearson correlation coefficient of 0.69 in case of whole documents as compared to a correlation coefficient of 0.59 in case of paragraphs.

Kumar et al. proposed a method based on paragraph-similarity to improve the performance of link-based citation networks for finding similar judgments in common law system [59]. The authors identified different components of a legal judgment which might be useful to identify similar judgments. These include name of judgment, name of judges, judgment identifier, acts referred, headnote (summary of judgment) and citation to other judgments. They observed that most of the judgments contain very few citations and therefore other textual information from judgments must be utilized to identify similar judgments. The paragraph similarity system is based on the hypothesis that each paragraph is represented by a legal concept. Also a paragraph of a judgment may refer to a specific paragraph of another judgment (which represents the legal concept of that judgment). Therefore, two different judgments are considered to be similar if they have similar paragraphs. A paragraph link is introduced between two judgments $J1$ and $J2$ if a paragraph in $J1$ is similar to a paragraph in $J2$. Each judgment was segmented into its constituent paragraphs by using regular expressions. The pre-processing methods applied to paragraphs include: conversion to lowercase, stopword removal and stemming (Porter's algorithm [99]). The paragraph vectors were then computed by utilizing the TF-IDF weights for each term in the paragraph. The cosine similarity values were computed between a particular paragraph vector of a judgment $J1$ with all the paragraph vectors of judgment $J2$. A paragraph link is introduced between two paragraphs if the cosine similarity values are greater than a threshold value. The similarity between two judgments is computed by using the bibliography coupling method. Two judgments $J1$ and $J2$ are considered to be similar when the number of paragraph links are greater than or equal to the threshold number of paragraphs. The discussed methodology was applied a dataset of 3866 judgments from the Indian Supreme Court from 1970

to 1993. A threshold of 0.3 was selected for cosine similarity values of paragraphs. The threshold for the bibliography coupling method was chosen as 3. This implies that two judgments are considered to be similar if they have 3 or more similar paragraphs. An experimentation with different threshold values for paragraph similarity could have provided better results. In the evaluation, the authors presented the number of similar paragraph links between different judgments. The model was evaluated for 50 judgments with a gold standard prepared by legal domain experts. The authors present the similarity scores between different judgments but do not provide any accuracy measures. The computation of standard information retrieval metrics such as precision and recall could have provided more insights into the performance of the system.

The authors in [58] presented a similarity analysis of legal judgments from the Supreme Court of India. The objective of the system was to retrieve the most similar judgments for a given input judgment under a common law system. The authors identified different sections of the judgment such as act, headnote, citation and case citation. They developed four different similarity measures : all-term cosine similarity, legal-term cosine similarity, bibliographic coupling similarity and co-citation similarity. In the all-term cosine similarity, all the terms from the judgment were utilized to create a vector space model on the basis of TF-IDF score. In the legal-term cosine similarity, only terms that occurred in a legal dictionary were considered to create vectors. The bibliographic coupling similarity between two judgments is equal to the number of common out-citations. Out-citations are the case-citations in a judgment "J" which refer to other judgments. In-citations for a particular judgment "J" are the case-citations from other judgments which refer to the judgment "J". The co-citation similarity score between two judgments is equal to the number of common in-citations. Experiments were carried out on a dataset of 2430 judgments from the Supreme Court of India. A gold standard mapping was created for 20 judgments by domain experts. The results of the similarity analysis indicate that the all-term cosine and co-citation similarity techniques were not useful for finding similar judgments. The legal-term cosine similarity and bibliographic coupling similarity had a better performance than other two methods. The presence of legal terms and citations to the other judgments were the most important features among similar judgments. The target users of the system are lawyers who would benefit by the availability of the relevant judgments for a new case.

### 6.1.2 *Retrieval of Similar Patents*

Indukuri et al. [53] proposed a technique to carry out claim similarity analysis between two patents by utilizing natural language processing

techniques. The information present in claims is of high importance while awarding new patents and also in cases of patent infringement. The claim similarity detection system is based on the computation of semantic and lexical similarity. A natural language processing pipeline was developed for processing the claim texts. The texts are tokenzied and then tagged with part-of-speech tags using the Stanford log linear tagger [114, 115]. Only nouns were extracted and other part-of-speech tags were ignored. The nouns were reduced to their inflectional forms by using the Lovins stemming algorithm [72]. The lexical similarity between two claim texts is computed by taking into account the number of common words and total number of words. The similarity measure is directly proportional to the number of common words and inversely proportional to the total number of words. A separate natural language processing pipeline was developed for computing the semantic similarity between two claim texts. After part-of-speech tagging, singular and plural nouns were extracted. WordNet was used to compute the similarity scores between a noun of one claim with all the nouns of another claim. The similarity score for the most similar noun was recorded and those word pairs were stored. Finally, the semantic similarity scores for all the word pairs were added to compute the similarity between two claim texts. The authors evaluated their approach on a corpus of 73 patent claim texts in four different categories. The results indicate that semantic similarity had slightly better performance than lexical similarity. The proposed system thus acts as a support tool for patent analysts and intellectual property attorneys in the process of claim analysis. However, the system can benefit by incorporating vector representation of texts based on distributional semantics.

The authors in [85] argue that patent text analysis methods based on the extraction of individual concepts do not completely capture the semantic information represented in the patent texts. The methods based on individual concepts also do not model the relationships between individual concepts. Therefore, they propose a method for measuring the textual similarity between two patents by using combined concepts. The first step in the text analysis of a patent text involves a language processing pipeline. This consists of a number of steps such as stemming, part-of-speech tagging, stopwords removal and synonym substitution. All the synonymous terms are normalized to one single form. The terms obtained after the application of pre-processing pipeline are mapped to more general terms by using an ontology. The resulting terms are classified into individual and combined concepts. The individual concepts are represented by single terms. Multiword expressions were represented as combined concepts. Combined concepts were also be created by combining individual concepts. The size of individual concepts played a key role in defining a combined concept. Large combined concepts achieved a high

precision and low recall while retrieving similar patents. The size of the window from which the combined concepts were extracted also influenced the patent similarity. A smaller window resulted in capturing the contextual relationships between the individual concepts but missed out the relationships between concepts far away from each other in the text. A larger window on the other hand captured the co-occurrence of individual concepts away from each other. The overlap between the concepts in the two patent texts is used to compute a similarity measure. The proposed methodology was applied in two patent management tasks.

Zhang et al. [125] proposed a hybrid similarity measure to analyze the patent portfolios. A patent portfolio consists of a set of patents with similar textual and semantic features. The hybrid similarity measure is a combination of a semantic similarity measure and a categorical similarity measure. The categorical similarity measure was developed by using international patent categories (IPC). Fuzzy set models were used to deal with the ambiguity of the vague classifications provided by IPC. A membership function based on expert knowledge was used to define the degree to which a particular patent portfolio represents a certain category from the IPC. Each patent portfolio was then modelled as a vector representation of the membership grades of the fuzzy model of IPC. A cosine similarity value was then computed by utilizing the vectors of two patent portfolios. This value represented the categorical similarity measure between two patent portfolios. For the semantic similarity measure, a tree-based semantic model was proposed. The important representative terms from the corpus were extracted via term clumping to construct a portfolio-term matrix [124]. Each patent portfolio was represented as a tree and the terms were modeled as leaves. The important terms in each patent portfolio were identified by applying a clustering algorithm. The important terms were grouped together in clusters and the terms with the highest prevalence value in each cluster were called as Level 1 leaf. The remaining terms in the cluster which are linked to the Level 1 leaf were called as Level 2 leaf. Each patent portfolio was represented by a tree. It consisted of a root, Level 1 leaves mapped to the root and Level 2 leaves mapped to Level 1 leaves. The semantic similarity was computed by comparing the branches of the trees of the patent portfolios. The proposed measure followed the tree traversal approach for computing the similarity. However, it utilized branch-based comparisons instead of node-based comparisons to identify important words in each patent portfolio. The authors observed that the categorical similarity measure performed better for the raw corpus (without pre-processing). The semantic similarity measure showed a better performance in case of a clean and pre-processed corpus. Therefore, the categorical measure was applied to the raw corpus while the semantic measure was used on the pre-processed corpus. The authors evaluated

the similarity measures on a dataset of 65 portfolios (made from 1632 patents). The results indicate that the tree-based semantic similarity measure improved over the lexical cosine and jaccard measures in terms of both precision and recall. The semantic similarity model also outperformed the categorical similarity measure. The semantic model had a better performance than the lexical model. However, a comparison with other sophisticated semantic models based on word embeddings or latent semantic indexing could have provided a better understanding about the performance of the proposed models.

The work presented in [24] utilized text similarity techniques to identify knowledge linkages between patents and their citations. Their research is based on the hypothesis that the text similarity value between a citing-cited patent pair is higher than that of a non-citing-cited pair. The text similarity techniques based on the vector space model (VSM) were used to identify similarities between patents and their citations. The pre-processing steps included stopwords removal and stemming. The WF-IDF weighting method was used in the vector space model. The term frequency was computed by using the sub-linear scaling method. The WF-IDF measure is the product of the weighted term frequency and inverse document frequency. Each patent was thus represented as a WF-IDF vector. The text similarity between two patent vectors was computed by the cosine similarity measure. The authors computed the similarity values between different components of patents such as title, abstract, description and claims. The results indicate that citing-cited pairs had higher similarity values than non-citing-cited pairs. However, this approach did not take into account the semantic meaning of words because a similarity score of 0 is returned when the vocabularies of two patent texts being compared do not share any words. Semantic similarity methods using a dictionary or thesaurus can overcome this limitation.

Moehrle et al. [86] provided a methodology to model patent informatics and retrieval as a business process. The major business processes included pre-processing, patent analysis and discovered knowledge. The pre-processing process prepared the patent documents for the next task of patent analysis. The patent documents were classified (to one or more domain), digitized and stored in a database. The patent documents were then converted to a XML format by segmenting different sections. Some indexing keywords were associated with each document. The patent analysis process retrieved a set of patents for a particular query. The process used different retrieval models such as boolean retrieval, extended boolean retrieval, ontology-based retrieval and latent semantic indexing. Then text similarity measures were computed to retrieve the most similar patents for a particular query. The process of discovered knowledge assisted in strategic patent decision making.

The research work presented in [96] used a methodology based on text similarity to develop patent lanes. Patent lanes represent the evolution of patent clusters over the course of time. A fixed number of patents were added to a basic set for creating patent lanes. The patents were selected by keyword-based or classification-based search. Text similarity measures between all the pairs of patents were computed to create a similarity matrix. Cluster analysis techniques were applied to the oldest patents to form initial clusters and outliers. These represented the starting lanes. The new patents were treated in a chronological order. The similarity values were computed between all the patent pairs to identify the most similar patent for a particular patent. A new patent would expand the patent lane if its similarity value with the connected patent was higher than a certain threshold. However, if the similarity value was lower than the threshold then a new patent lane was added. The last step involved the extraction of important terms (using TF-IDF) to represent the patents and the lanes. A Flex N-grams based similarity measure was used with complete linkage and double-single-sided (DSS) inclusion [95]. Flex N-grams contain atleast one gap of variable length. Therefore, the patterns can be of undefined length. The sections of the patent which were considered for similarity calculation included descriptions, claims, abstract and title. The proposed methodology was applied to develop patent lanes for the patents belonging to the field of carbon fiber reinforcements. The case study on the carbon fiber indicate the usability of the system for patent attorneys. However, the system suffers from the limitation of not capturing the concept similarity among texts when different terminologies are used to represent the same concept. Also topic models could have been utilized to develop patent lanes as they have shown to perform well in clustering documents over time [119].

### 6.1.3 *Legal Question Answering*

Kim et al. developed a text similarity system for a legal question answering task [57]. The goal of the task was to retrieve relevant Japanese civil law articles for a legal bar exam question. The authors used both unsupervised and supervised models to address this task. The unsupervised models consisted of a TF-IDF model and a topic model-based information retrieval system. The questions and articles present in the corpus were tokenized and processed using the Stanford CoreNLP tools [79]. The TF-IDF model was implemented in Lucene[1]. The topic-based similarity model utilized latent dirichlet allocation (LDA) to represent both questions and articles as a distribution of topics. The results indicate that the TF-IDF model had a better performance than the LDA model. This is expected because articles and questions are short texts. The application of latent dimensionality reduction models

1 https://lucene.apache.org/core/

such as LDA lead to loss of informative features which are relevant for similarity. A ranked support vector machine (SVM) model was implemented with TF-IDF, LDA and dependency pairs and lexical words as features. The best performance was achieved by the ranked SVM model with a combination of both LDA and TF-IDF scores as features.

The work in [20] presented a text similarity approach which combined both lexical and distributional sentence features. The system was used to retrieve the relevant articles from the Japanese civil code for an input bar exam question as discussed in [57]. For an input question, a ranked list of relevant articles was retrieved by using a n-gram based relevance analysis method. The pre-processing of articles was followed by n-gram generation. The n-gram set of an article was expanded by including the n-grams of the referenced article. The relevant articles for each question were ranked by using a score function which took into account the size of the set of n-grams from both question and article, their relative significance and the inverse document frequency for the terms in the article collection. The top two articles in the ranked list were evaluated for ambiguity if the similarity scores were very close. Sentence embeddings were computed for both question and article by using word2vec and term order probabilities. They were used to resolve the ambiguity. The ranked list was updated by computing a cosine similarity score between the sentence embedding vectors of question and articles. The results demonstrate that the approach was competitive with other state-of-the-art methods for the question answering task. We also observed that the addition of word embeddings and term order probabilities improved the performance of the system.

Heo et al. [47] present a content fusion based text similarity system for a legal question answering system. Both questions and articles are represented in a vector format using three different models: TF-IDF, Word2Vec and LSA. The dataset consisted of 659 questions and 1098 legal articles. The dataset was pre-processed by performing tokenization, lemmatization and stopword removal. The content of each article was segmented into three parts: title, body and example sentences. Each part along with the question was represented in three different vector representations. The cosine similarity values were then computed between the question and each part of the article. The three different similarity scores were combined by the weights computed by least square method (LSM) and linear discriminant analysis. An additional dataset of 618 laws and 140 judicial precedents was utilized to train the word embeddings by using the skip-gram model. The results indicate that TF-IDF and LSA models outperform word2vec in terms of mean average precision and recall. The small dataset used to train word embeddings did not capture enough semantic information to accurately model the legal articles. The authors also illustrate the

better performance of the fusion weighted methods over unweighted similarity measures.

### 6.1.4 *Legal Statutes and Provisions Retrieval*

The work in [70] presents an approach to identify relevant legal statutes for a user query. The objective of the statute retrieval system is to identify statutes associated with the legal problems of the user. The authors proposed a three-phase prediction model for the classification of judgments. A statute is considered as a label in the classification task. Therefore, each judgment can have multi-labels (as each judgment cites at least one statute). A pre-processing pipeline is developed for cleaning the text. Words with noun, adjective and verb part-of-speech tags were retained. The TF-IDF weighting measure was applied to generate a vector representation of judgments. Each judgment in TF-IDF vector format is linked to its statute labels. A support vector machine (SVM) classifier is built on the TF-IDF judgment vectors and their associated statute labels. The terms in the user query are mapped to judgment terms (all terms in the judgment training corpus) by using the Normalized Google Distance (NGD). The SVM predictions are then applied to the user query to produce a list of top $k$ statutes. In the next step, the most relevant statutes from the top $k$ statutes are selected by mapping the terms from the user query to the statute terms. Each statute is represented in TF-IDF format. The Normalized Google Distance is applied to transform the user query into statute terms. A cosine similarity measure is then computed between the user query and the top $k$ statutes to find the most similar statutes. In the last stage, an apriori association algorithm is applied to find association rules between the statutes. As a result the most relevant statutes are retrieved. The system was evaluated on a set of around 1500 criminal judgments from China. Fifty civil news stories were used as queries for the statute retrieval task. The system was evaluated in three different phases by computing the recall at each phase. The results indicate that the best performance was achieved after the third phase of mining association rules. The performance of the system was also compared with other state-of-the-art models such as cosine similarity, Pearson correlation coefficient and Spearman's correlation coefficient. The proposed three-phased prediction model achieved a better recall than other methods. The system can also benefit by the inclusion of semantic knowledge from legal ontologies.

Humphreys et al. [52] developed a system to map recitals to legal provisions in the European legislation. A gold standard mapping was developed to link the recitals in the preamble with the articles in the normative provisions. However, the authors did not include the mappings from recitals to sub-provisions. The recitals and legal provisions were modeled as TF-IDF vectors. The similarity system

utilized part-of-speech tags and terms in the substantive titles as features. A cosine similarity score was computed between the TF-IDF recitals and provisions vectors. The results indicate that the system achieved a high accuracy due to the presence of a large number of true negatives (unbalanced dataset). The system also achieved a high recall but with low precision. Such text similarity systems could be used to automatically identify all possible correspondences between recitals and provisions but they would need to be checked by a legal knowledge engineer.

Magerman et al. [75] investigated the application of text similarity techniques based on vector space models and latent semantic analysis (LSA) to map patents and scientific publications. The system was evaluated on a corpus of 467 documents (30 patents and 437 publications). The pre-processing pipeline comprised stop-words removal, stemming, term reduction and weighting. The results indicate that the TF-IDF weighting scheme using vector space models achieved the best performance. The authors investigated the application of LSA with different singular value decomposition (SVD) ranks for approximation. They inferred that for their small dataset higher values of SVD rank have better performance than low rank values. They also noticed the application of LSA transform over TF-IDF weights degrades the performance of the TF-IDF model.

The work in [63] utilized word embeddings to extend traditional full text search models for legal corpus. The proposed method is useful to find both exact and semantically similar matches. The word embeddings were trained on a corpus of the German Civil Code (GCC). The word vectors for the individual words of the search query were accumulated to get a single dense vector for the search phrase. The corpus of word embeddings was then searched with a window size equal to the length of the search phrase vector. The window size was represented by $n$. The search was thus limited to phrases of length equal to the search phrase. The search window was then sequentially shifted by a length of $n/2$. The cosine similarity measure was then computed between the search phrase vector and all the accumulated vectors from the word embeddings corpus. The vectors were ranked on the basis of cosine similarity score to select the top $k$ matches. In the next step, the chosen matches were reconsidered and the window size was increased to three times the number of terms in the search phrase. The window was then moved term by term to get another vector. The accumulated vectors were then compared with the search phrase vector by computing the cosine similarity. The best match with the highest cosine similarity was selected. The proposed approach was evaluated on two datasets : the EU Data Protection Directive 94/46/EC (EU-DPD) and a corpus of ten German rental contracts. The results presented by authors included only some examples of search phrases and the phrases retrieved by the model. They did not utilize

annotated data or a gold standard mapping from legal practitioners to thoroughly evaluate the relevance of the retrieved phrases to the search query. The performance of the proposed model was quite slow due to the exhaustive searching based on the window search approach. This approach will be very computationally expensive when a large corpus of several thousands documents is considered for searching.

Rosso et al. [101] developed a passage retrieval system based on text similarity techniques for treaties, patents and contracts. The passage retrieval system was based on a density n-gram similarity measure. The query was provided as an input to the search engine. The search engine retrieved the snippets relevant to the query from a document corpus. Then n-grams were extracted from the query and all the retrieved snippets. A similarity measure was then computed by comparing the n-grams of the query with the n-grams of the snippets. Each snippet was linked to a passage in the document. The similarity values were then used to rank the retrieved passages for the input user query. The passage retrieval system achieved state-of-the-art performance in the treaty retrieval task based on the JRC-ACQUIS[2] dataset. However, it had a poor performance in the patent retrieval task (based on the European Patent Office dataset) as it did not take into account synonymous terms and query expansion methods.

### 6.1.5 *Automated Conflict and Dispute Resolution*

Mahfouz et al. [76] proposed a methodology for automatic extraction of implicit legal knowledge from Differing Site Condition (DSC) litigations for construction projects. An automated approach for conflict resolution in the construction industry would save both time and money. The corpus consisted of 600 cases from the Federal Court of New York from 1912 to 2009. A total of 15 important legal factors and concepts were considered for automatic extraction. These factors and concepts were identified by using statistical analysis. They were further evaluated and were found to be effective in predicting outcomes of DSC litigations. Each case was annotated with the important legal factors. Each legal factor was associated with a set of words which were implicitly linked to it. The annotated dataset was utilized to build machine learning models to identify the important legal factors in new DSC litigations. The corpus was converted into a bag-of-words representation. The capitalized words were converted into lowercase. The words were further reduced to their root forms. Each case was then represented as a vector of words in their root forms. Four different weighting schemes were applied to the bag-of-words-representation. These included term frequency (tf), logarithmic term frequency, augmented term frequency and term frequency inverse document frequency (TF-IDF).

2 https://ec.europa.eu/jrc/en/language-technologies/jrc-acquis

Each case was then represented as a vector with the frequency of words (from different weighting measures). The authors developed three machine learning models (Naive Bayes, Decision Trees and projective adaptive resonance theory (PART)) for each of the four weighting schemes. The accuracy of the models was evaluated over 10-folds cross validation. The results indicate that the performance of the Naive Bayes classifier was most influenced by different weighting schemes. However, the performance of Decision Trees and PART models were not influenced by the different weighting measures. The TF-IDF weighting measure had the best performance for all three classifiers. The Naive Bayes classifier had the best performance among the three classifiers. The authors also compared the results with a previous implementation of support vector machines (SVM) classifier. The Naive Bayes classifier outperformed the SVM classifier. The better performance of Naive Bayes can be attributed to the fact that the presence of a large number of terms (in case-term matrix) resulted in a better estimation of the conditional probability. An investigation of other weighting measures such as latent semantic analysis and latent dirichlet allocation might be interesting for future work because the objective of the research is to find latent and implicit legal knowledge.

The research work in [36] presents a text similarity system for retrieving similar cases for alternative dispute resolution in construction accidents. The availability of similar historical construction cases can provide alternative dispute resolution and avoid expensive litigations. A database of construction cases was developed using the Westlaw[3] legal research service. The database consisted of 360 construction accident cases in Hong Kong from 1990 to 2011. The pre-processing consists of a number of steps to achieve a bag-of-words representation of cases. Tokenization was carried out to segment cases into a collection of words and phrases. The maximum length of phrases is set to 3 to retain important terms. A combination of stemming and lemmatization was applied to the tokenized terms. The next step involved the extraction of important terms by setting threshold values for term length and minimum term frequency. The extracted terms from all the document collections are stored and indexed in a dictionary. Each construction case is represented using the vector space model (VSM). The authors identify that the most relevant and distinguished terms from the case documents should be used to create vectors. Therefore, TF-IDF weighting scheme was used to weight the terms. The dimensionality reduction transform was applied using latent semantic analysis (LSA) on the TF-IDF vectors. A new case is processed through the same pipeline to get a vector representation. The cosine similarity and euclidean distance scores are computed to retrieve the most similar cases for a given new case. The authors evaluate the system by retrieving the top-$k$ relevant cases. They observed that the number of

3 http://www.westlawinternational.com/

relevant $k$ (retrieved cases) varies for each type of case. The rare types of disputed cases tend to have lower values of $k$ as compared to more common kind of cases. It was also observed that the increase in $k$ did not lead to a major improvement in recall. This is because the most relevant cases have the top-most rank in the retrieved list.

### 6.1.6 *Contracts Compliance Check and Trademark Retrieval*

The work in [105] presents an approach which uses text similarity techniques to check compliance of contract templates. Most information technology (IT) businesses use standardized contract templates for a particular service. However, the major challenge in standardization is to ensure that the actual contracts are compliant with the standard templates. The text similarity techniques are used to develop a system which identifies the top candidate templates for a particular contract on which a structural and semantic analysis can be performed. The task of compliance checking is quite challenging because the actual contract can be compliant with a template even in the absence of a symmetric alignment between the headings or a high magnitude of similarity between the texts. There may also be several hundred contract templates for a particular service in large organizations. Therefore, a front end filter is developed to identify the best templates for which tree matching is performed. The contracts stored as images were converted into text format by using optical character recognition (OCR). However, the errors (such as spelling errors) arising due to this conversion were not resolved. The system was developed using Apache Lucene[4]. The techniques used to compare contracts and templates included latent semantic analysis and cosine similarity. The contract and template text files are used to create separate Lucene indices. The terms from the template and contract index are extracted using Lucene's API to create the term space which is used to convert template and contract documents into vectors. The authors use IDF smoothing and term boosting for term weighting. Very high IDF values were reduced by using a smoothing function. Term boosting was achieved by doubling the weight of domain-specific terms (terms appearing in document titles and section headings). A term-document matrix was created and a cosine similarity score was computed between the vectors representing the contracts and templates. The latent semantic analysis (LSA) transform was applied on the term document matrix to create a lower dimensional concept space. The cosine similarity was computed between the LSA vectors of contracts and templates. The retrieved templates for each contract are ranked based on the similarity values. The system was evaluated by retrieving the templates above a certain threshold and also the top-k templates (for $k = 1, 3, 5$). The results indicate that cosine similarity

4 http://lucene.apache.org/

achieves a higher recall in top-k retrieval. In case of threshold-based retrieval, LSA had a better recall than cosine similarity. It was also observed that IDF smoothing and term boosting achieved a better performance as compared to term frequencies.

Anuar et al. [6] developed a conceptual model based on semantic similarity for identifying similar trademarks. The visual, conceptual and phonetic similarity between two trademarks is one of the key factors in infringement litigations. The conceptual model utilizes a database with European trademark infringement cases from 1999 to 2012. The authors divided the extracted disputed trademarks from the database into four types of similarity: exact match, synonyms/antonyms, lexical relations and foreign mark (cross-lingual synonyms). The exact matches account for 50% of the dataset and the remaining categories combined together account for the remaining 50%. The main objective of the conceptual model is to identify the remaining 50% disputed trademarks which cannot be found by exact string matches. The proposed conceptual model consists of an indexing and a retrieval module. In the indexing module, the trademarks are first classified into English and non-English. The non-English trademarks are translated into English by utilizing multi-lingual dictionaries. Then pre-processing is applied to get tokens for each trademark in the database. The feature set for a trademark consists of the token and synsets extracted from external knowledge bases (dictionary or thesaurus). The best synsets are chosen by disambiguating the tokens with respect to the synsets definition. The retrieval module takes the trademark text as input and generates the tokens in a similar fashion to the indexing module. The synonyms and antonyms of the query tokens are retrieved from the knowledge base and stored as additional features. The system compares the features of the query trademark with the feature set of each entry in the database to categorize and rank the retrieved results. The features are compared for exact matches, synonym/antonyms and lexical relations. The authors presented an example to illustrate their approach but a discussion about the evaluation of the proposed model was planned for future work. A comparison of the proposed conceptual model with other similar approaches for trademark similarity may provide insights about the benefits and limitations of their model.

Table 6.1 presents the different legal-tech use cases for text similarity techniques identified in this section. We identified six major use cases where text similarity techniques have been used for information retrieval. We observe that most of the works utilized unsupervised text similarity techniques based on lexical and semantic features. There were few works which used supervised text similarity techniques due to the presence of a labeled dataset.

Table 6.1: Text Similarity Related Work

| **Legal-Tech Use Cases** | **Text Similarity Techniques** |
|---|---|
| Retrieval of similar cases and judgments | TF-IDF Cosine, Word2vec, LDA and paragraph vector [77], Bibliographic coupling similarity based on number of paragraph links [59], Legal-term cosine similarity, citation-based similarity[58] |
| Retrieval of similar patents | Syntactic and Semantic Similarity based on common words [53], Concept-based patent similarity [85], Hybrid similarity measure as a combination of semantic and categorical similarity [125], Flex N-grams based similarity [96] |
| Legal Question Answering | TF-IDF, Word2vec and LSA [47], N-gram and word embeddings based similarity measure [20], TF-IDF, LDA and SVM for supervised text similarity) [57] |
| Legal Statutes and Provisions retrieval | TF-IDF Cosine [52], Word embeddings based similarity [63], N-gram based similarity [101], TF-IDF, SVM, Normalized Google Distance [70] |
| Automated Conflict and dispute resolution | Naive Bayes and SVM with TF-IDF, TF, logarithmic and augmented term frequency [76], TF-IDF and LSA[36] |
| Other (Contracts Compliance check, trademark retrieval) | IDF smoothing, term boosting and LSA[105], Knowledge-based similarity using synonyms [6] |

## 6.2 MACHINE LEARNING FOR LEGAL INFORMATION RETRIEVAL

This section presents a literature review of the machine learning methods used in legal information retrieval. The following subsections discuss various state-of-the-art research works in different areas such as prediction of court decisions, classification of legal norms and acts, extraction of semantic relations and contract elements, patent litigation likelihood and readability of legislative sentences.

### 6.2.1 *Prediction of Court Decisions*

The authors in [3] developed a machine learning system to predict the judicial decisions of the European Court of Human Rights (ECHR). The system utilized textual features from different subsections of the case such as "relevant applicable law", "facts", "circumstances", "Law" and "full case" to predict whether there has been a violation of an article of the convention of human rights. A dataset of 584 cases was compiled from articles 3, 6 and 8 of the Convention. The authors utilized N-grams and topics as features for the binary classifier. The top-2000 most frequent N-grams (for N $\epsilon$ $\{1, 2, 3, 4\}$) were utilized from the dataset. Topics were created for each article in the dataset by clustering semantically similar N-grams together. This provides a concise representation of the semantic space by reducing the dimensionality of the feature space. A support vector machine (SVM) classifier was trained using the textual features to predict if there is a violation or non-violation for a particular case (with respect to the Article of the Convention). The results indicate that N-gram features from the "circumstances" subsection achieve a better performance as compared to other subsections. The topics features developed by clustering similar N-grams achieve the highest accuracy from all the feature set. Topics capture the overall gist from the N-grams of different subsections and thus are able to be a good predictor. The authors also infer that the information contained in the "circumstance" subsection is a key predictor in determining if the case is a violation or not.

Another work in [55] also used a machine learning based approach to predict the outcomes of cases of the Supreme Court of the United States. The problem of predicting outcomes is modelled as a binary classification task. The objective is to predict whether the Supreme Court will affirm or reverse the decision of a lower court. The predictions of the individual judges is used to forecast the overall decision of the Supreme Court. The authors utilized the Supreme Court Database to build their machine learning prediction model [110]. The database consists of around sixty years of high quality formatted data on the behaviour and outcomes of the Supreme Court. The authors utilized extremely randomized trees (ERT) model with court-level, justice-level,

case-based and historical features. ERTs are ensemble models based on random forests which leverage randomness from both features and data to make robust predictions with less variance as compared to classification and regression trees. The proposed model consists of more than ninety input variables as features. The model was trained on the cases from 1946-1953. The predictions were made from the term of 1953 to the end of the 2012-13 term. The model was developed for two types of predictions: vote prediction for individual justices and overall prediction of the Supreme Court. The system is evaluated using 10 folds cross-validation. The model correctly predicts 70.9 % of individual justice votes and 69.7 % of overall case outcomes of the Supreme Court. The results point out that the Court reverses a majority of the cases it accepts. The machine learning model is able to identify most of these reversal decisions correctly. It was also observed that prediction of affirm decisions was more difficult than reverse decisions. In a large number of cases, the model incorrectly predicted that the court will reverse the decision. These false positives are mainly due to the asymmetry in the reversal/affirm rates of the Supreme Court. The majority of the predictive power of the model is attributed to a number of behavioral trends including historical trends, current supreme court trends, individual supreme court justice trends.

Dunn et al. [34] presented a machine learning based approach to predict the decisions of the asylum court in the United States. Their research was motivated by findings of disparities in the asylum judgments. The objective of the proposed system is to predict the outcome of an asylum application based on input features: judge (unique identifiers for each judge, gender and work history), attorney (whether applicant is represented by an attorney or not), nationality of the applicant, location (local courthouse), base city (regional immigration court assigned to the applicant), language, case type (affirmative or defensive) and notice to appear. A random forest model was built on the data obtained from the Executive Office for Immigration Review (EOIR). The model uses features available at the time an applicant receives a notice to appear. Therefore, the model makes an early prediction by using only the information and features available at the time when case opens. The dataset consists of around 600,000 cases. The dataset was divided into 80% train and 20% test set. An incremental evaluation was carried out with different features set to evaluate the predictive power of each feature. Both accuracy and area under the receiver operating characteristic curve (AUC) were computed. The results point out that just a single feature, Judge ID was correctly able to predict the outcomes with an accuracy of 71%. The addition of applicant's nationality feature increases the accuracy to 76%. The highest accuracy is achieved when all the engineered features are utilized for prediction. It was also observed that the highly predictable judges held fewer hearings before making the decision. This may indicate

snap or predetermined judgments. On the other hand, the judges who held more hearings were found to be less predictable by the model. Such a predictive system could be useful for asylum seekers and law clinics (working in asylum cases) to estimate their chance of a grant.

Wongchaisuwat et al. developed a machine learning based model to forecast the litigation likelihood for patents [120]. The model also forecasts the expected time to litigation. These forecasts are highly useful for corporations working in patent litigation as they lead to time and cost savings. The patents belonging to three technology classes (Wireless Network, Telecommunication and Advertising) were added to the dataset. The litigation information about each patent was added from the LexMachina[5] dataset. The authors utilized both textual and non-textual features to build their litigation likelihood prediction model. The textual information from the claims section of the patents was used to build a term-document matrix with TF-IDF weights. The non-textual features included both numerical and categorical features. Some of the numerical features include number of claims, number of words in claims, number of inventors, number of backward references and average PageRank scores of backward references.

The categorical features were collected from the website of the US Securities and Exchange Commission and included market share price, earnings per share and revenue of the patent's assignee [120]. A support vector machine (SVM) classifier was built on the dataset by utilizing both textual and non-textual features. The authors also proposed a cluster and ensemble based method which groups litigated and non-litigated cases and applies a ensemble classification model for prediction. The results of the litigation likelihood model represent the difficulty of the prediction task. The highest F-score for the best performing patent type was 0.19. Therefore, further feature engineering is required to improve the accuracy of the system. The addition of more hand-crafted features may also be beneficial. The model for predicting the time to litigation was developed by adding a class variable with a time period. However, the F-score values for this model were not presented in the paper.

### 6.2.2 *Classification of Legal Norms and Acts*

The work in [126] presents a machine learning model for classifying legal texts from different jurisdictions of US states. The dataset consists of legal acts in the domain of public health system from eight different US states. The model is trained incrementally on a dataset of source states and is used to make predictions on two target states. The predictions on the target dataset improved by a considerable margin with the addition of data from more states. The generic nature of the framework allows the use of different classifiers for this task. Also

5 https://lexmachina.com

more sophisticated features such as word embedding-based vectors and topic models can be added for a better semantic representation of legal texts for classification.

The authors in [73] developed an automatic system for classifying sentences in Dutch law. A classification of different types of sentences was developed by utilizing the patterns found in each type. Some of the classification types included in this study are definitions and type extensions, deeming provisions, norms, application provisions, penalization, value assignments and lifecycle. A total of fourteen different types of provisions were considered for classification. These types were identified by engineering 81 patterns from twenty Dutch laws. The patterns mainly consisted of verb phrases and keywords. A program which reads a structured legal text (in MetaLex[6] format) and classifies its sentences according to the engineered patterns was developed. The classification program was evaluated on a dataset of fifteen different Dutch laws consisting of new laws, amendments and Royal Decrees. The manual annotations of sentence types on these laws was used to evaluate the classification program. The classifier had a very good performance with an overall accuracy of 94%. The authors also presented the number of found and missed sentences for each type. We also observe that the largest number of sentences belong to the norm category. Some sentence types were not identified correctly due to the presence of auxiliary patterns. This problem can be addressed by developing priority rules or patterns. The inclusion of F-score, precision and recall metrics in the evaluation could be considered for future work.

Waltl et al. [117] developed a system to classify legal norms by using active machine learning for German statutory texts. The system uses a combination of rule-based methods and active machine learning to reduce the time and effort for labeling norms. A rule-based system is used to automatically generate a corpus of labeled and unlabeled norms. A set of labeled norms are then used to train a classifier. The classifier then predicts the labels of the unlabeled norms. According to a query strategy, the unlabeled norms are chosen by the machine learning classifier to distinguish more efficiently between different types of norms. The instances labeled by the classifier are then checked by a domain expert and added to the training set. The corpus consisted of 504 sentences. Each sentence was assigned to a norm type by a legal expert. The authors considered eight norm types in their experiments. These include objection, statutory rights, statutory duties, legal consequence, procedure, continuation, reference and definition. 75% of the dataset was used for incremental training and the remaining 25% as testing set. Tokens and part-of-speech tags were used as features for training classifiers. Naive Bayes, Logistic regression and Multi-Layer perceptron classifiers were used for both supervised learning

6 http://www.metalex.eu/

and active learning. The results indicate that active machine learning had a better performance than supervised learning in case of Naive Bayes and Logistic Regression. The performance of the system could have been improved by utilizing well known textual features such as TF-IDF.

The work in [12] presents a system for multi-label classification of Italian legislative texts. Each legislative document in the corpus is associated with multiple labels from the EuroVoc[7] thesaurus. The corpus consisted of 23000 documents with about 7000 categories from EuroVoc. The transformation of multi-label data format into mono-label data format was carried out by segmenting a document with multiple labels. The segmentation is carried out in the vector representation of document. Each document with multiple labels was represented as a combination of individual vectors with single labels. The segmentation of the document vector into individual vectors with single labels is achieved by computing the pointwise mutual information (PMI) score over the co-occurrence matrix of terms and labels. A SVM classifier was built on the mono-label dataset by utilizing vector representation of text as feature. The probability distribution of different classes within a document predicted by the SVM classifier are used to assign multiple labels to each document. The system showed a competitive performance along with other state-of-the-art methods which used EuroVoc for classification. The system achieved a high precision among the first ranked categories but a lower recall due to the high number of categories associated to each document.

The work in [104] presents the application of interactive machine learning methodology as an extension to a legal information retrieval (IR) system for statutory analysis. The legal IR system retrieves a list of relevant statutes for an input query prepared by a legal researcher. The list of retrieved statutes is analyzed by the lawyer to identify a set of relevant provisions. These set of provisions are used to resolve the input legal query. The objective of the interactive machine learning approach is to filter out the irrelevant provisions retrieved by the legal IR system. Thus, the final output would consist of provisions most relevant to the legal issue represented by the input query. A machine learning classifier takes the retrieved provisions as input and classifies them as relevant or not relevant. A legal expert provides feedback on the output of the classifier. The expert can correct the classifier if necessary. The classifier learns from the human input and improves its prediction. A support vector machine (SVM) classifier with liner kernel is used for classification. It utilizes unigram textual features for classifying statutes as relevant or not relevant. The system was evaluated on a corpus of statutory texts from Kansas and Alaska. The performance of the machine learning classifier was evaluated for two experiments : i) without human interaction and ii) with human

7 http://eurovoc.europa.eu/

interaction. The results indicate that classifier shows a better performance when trained with human input. It was also observed that the accuracy of the classifier improved with the number of inputs from the expert.

### 6.2.3 *Extraction of Semantic Relations and Contract Elements*

Boella et al. developed a model to extract semantic relations from legislative text by using syntactic dependencies and machine learning [11]. Their approach is based on the assumption that a semantic tag is associated with a limited number of syntactic contexts. The task is modelled as a classification problem where each term in the corpus is linked to a semantic label given its syntactic dependencies. The syntactic dependencies are extracted by using the TULE parser [68]. The dependencies are then transformed into generic representations in triples format. A positive label is assigned for each noun in the corpus if it has been annotated with a semantic tag. All the nouns that are not linked to a semantic tag in the annotated data are labelled as negative. All the labelled instances are then vectorized as per the vector space model (VSM). The authors chose three semantic labels for this task: active role, passive role and involved object. A support vector machine (SVM) classifier is built using the vectorized instances as features. The system was evaluated on a corpus consisting of around 150 legal documents containing 2253 annotated semantic tags for nouns. The results indicate that the system achieves a very high F-score for classifying the active roles. However, the system had a high false negative rate in the classification of passive roles. In case of involved objects also, the system could not classify many instances correctly due to the wide coverage of this semantic label. This approach could be highly useful for ontology learning as it provides a decent support for automated identification of important semantic relations. Future work may comprise addition of other relations and improvement in feature engineering.

The work in [23] presents an approach based on natural language processing and machine learning to automate the extraction of key elements from legal contracts. The authors utilized an annotated dataset of around 3,500 contracts with eleven contract element types. The different contract element types included contract title, contracting parties, contract dates (start, effective, termination), contract period, value, governing law, jurisdiction, legislation references and clause headings. Each contract element type is linked to an extraction zone (a particular type of clause or region in the contract text where an element usually occurs). Therefore, the first task in contract element extraction involves the identification of different extraction zones by using a set of regular expressions. This also results in obtaining a well defined contract structure. The contract element extraction system

would look for a particular element only in its associated extracted zone. The contract element types occurring in a particular extraction zone were marked as positive instances. All other tokens which do not represent a contract element type in an extraction zone are labelled as negative instances. A number of linguistic hand-crafted features (some of them include uppercase tokens, lowercase tokens, numeric tokens, token length, numeric tokens, part-of-speech tags) were added to each contract element type. The authors utilize a sliding window logistic regression classifier for contract element extraction [122]. The classifier reads the tokens from each extraction zone and classifies them as positive or negative. A sliding window of size 5-6 tokens is used. Two versions of the model were built. The first one used hand-crafted features and the second one utilized word embeddings trained on an unlabelled corpus of 750,000 English contracts as feature vectors for the sliding window.

The authors also built another variation of the model by using support vector machines (SVMs) [23]. Lastly, a version of both SVM and Logistic Regression with the concatenation of word embeddings and hand-crafted features was developed. It can be observed from the results that the versions of SVM and Logistic Regression with concatenated features outperform other models. The model is able to achieve a decent F-score for each contract element type. We also observe that the recall achieved by the model using hand-crafted features is lower than that of word embeddings. This is expected due to the low coverage of hand-crafted features. The contract element extraction system may serve as a strong backbone for extraction of fine-grained semantic knowledge for legal analysis.

### 6.2.4 *Prediction of the Readability of Legislative Sentences*

Curtotti et al. developed a machine learning based system to predict the readability of legislative sentences [29]. They created a small corpus by randomly collecting 890 sentences from the United States Legal Code and the United States Code of Federal Regulations. In addition, 251 non-legal sentences were randomly added from the Brown corpus and a corpus of graded readers to calibrate the model. A gold standard corpus was creating by annotating the difficulty of each sentence as easy or hard. The difficulty was assigned by computing a score which combined different user responses. A baseline model was developed by using SVM classifier and features such as sentence length, average word length, type to token ratio, readability metrics and proportion of verbal phrase chunks. The overall accuracy was 72.7% over 10 folds cross validation. The authors extended the model by adding other features such as character and token-level N-grams and parsed features from dependency and context-free grammar. With the addition of the new features the accuracy of the classifier increased

to 76.7%. The dependency grammar based features were not found to be effective in improving the predictability of the model. Some interesting correlations are derived between different features and the readability score. However, they are limited by the small size and random nature of the dataset.

Table 6.2 presents the major use cases where machine learning techniques have been utilized to develop legal information retrieval systems. Majority of the works utilize supervised machine learning classifiers such as support vector machines (SVM), Naive Bayes, Logistic Regression and Decision Trees. We also list the various textual and non-textual features utilized by the machine learning models. It is interesting to note that machine learning models have been used in both prediction and extraction tasks.

## 6.3 CONCEPT-BASED INFORMATION RETRIEVAL

This section presents an overview of state-of-the-art concept-based methods used in legal information retrieval. Concept-based approaches have been used in information retrieval to incorporate semantic information from external knowledge bases such as ontologies, dictionaries and thesauruses. The following sub-sections present research works from relevant areas such as concept-based legal information retrieval, ontology learning from legal texts and named entity recognition in legal texts.

### 6.3.1 *Concept-based Legal Information Retrieval*

The authors in [44] developed an ontological framework to improve the keyword search-based information retrieval system of e-Government. The system consists of a search client, a search server and an ontology server. The system utilized legal ontologies built for the real estate transactions of the Spanish government. The concept instances from the legal ontology are associated with documents. The system takes a user query and links it to the Ontology server to retrieve the relevant concepts. The legal documents satisfying the query are retrieved by the system. The search server is based on Lucene.

In [2], a multilingual and a multi-level ontology, called European Legal Taxonomy Syllabus (ELTS) was developed for European law. It is a lightweight ontology and a knowledge base of legal concepts linked by semantic relations derived from the linguistic patterns of legal concepts in different jurisdictions within the European Union. ELTS was not developed as an axiomatic formalized ontology so that it can be flexible to take into consideration the conflicts and inconsistencies in the law. This is because the formalization of conflicting regulations has still many open questions. The ELTS ontology schema has a specific ontology for a particular jurisdiction. Specific relations

Table 6.2: Machine Learning techniques for Legal Information Retrieval

| **Legal-Tech Use Cases** | **ML Techniques** | **Features** |
|---|---|---|
| Prediction of Court Decisions | Prediction of judicial decisions of European Court of Human Rights (ECHR) by using SVM classifier [3] | Topic and N-gram features from sections of case |
| | Prediction of outcomes of the Supreme Court of USA by using extremely randomized trees [55] | Historical, case-based, court-level and justice-level features |
| | Prediction of the decisions of the asylum court of USA by using random forest classifier [34] | Judge identity, location of the court, nationality of the applicant etc. |
| | Forecast the litigation likelihood of patents by using SVM Classifier [120] | Textual (TF-IDF weights from claim section) and non-textual (number of claims, number of words in claims, number of inventors, market share price, revenue of patent assignee etc.) |
| Classification of Legal Norms and Acts | Classification of legal texts from different jurisdictions of US states by using decision trees [126] | TF-IDF weighted bag-of-words |
| | Automatic classification of norms in Dutch Law using a pattern-based classifier [73] | Patterns (verb phrases and keywords) extracted from Dutch Law |
| | Classification of legal norms in German statutory texts using rule-based methods and active machine learning (Naive Bayes, Logistic Regression and Multilayer Perceptron) [117] | Tokens and part-of-speech (POS) tags |
| | Multi-label classification of Italian legislative texts using SVM classifier [12] | Vector representation of texts with pointwise mutual information |
| Extraction of Semantic Relations and Contract Elements | Semantic relation extraction from Italian legislative texts using SVM classifier and syntactic dependencies [11] | Vectorized instances of dependencies |
| | Extraction of contract elements by using sliding window logistic regression, SVM classifier [23] | Handcrafted: Uppercase tokens, lowercase tokens, numeric tokens, POS tags and word embeddings |
| Prediction of the Readability of Legislative Sentences | Prediction of the readability of legislative sentences of US Legal Code and US Code of Federal Regulations using the SVM classifier [29] | sentence length, average word length, type to token ratio, N-grams etc. |

are used to model the relationship between concepts of different jurisdictions. These relations are different from the relations which model the relationship between concepts from the same jurisdiction. The ELTS ontology focusses on addressing the possible misalignments between the terms and concepts at European and national level. Such misalignments particularly happen while transposing directives into national legislation. For instance, the concepts at both European and national level are same but represented by different terms. In such cases, it is essential to align different terms under the same concept. ELTS associates each concept with domains, source and descriptions. Also each concept is mapped to one or more terms (a term can be a single word or a phrase). The source for each concept is linked to the particular document from which it was obtained. The documents are stored and maintained in a database. The ELTS ontology includes ontological relations such as "purpose", which link a concept to its source legal principle, and "concerns", which relates a concept to other similar (or related) concepts. These relations exist within an ontology (for a particular jurisdiction). Other relations such as "implementation" link the concepts between European and national level ontologies.

The ELTS ontology also takes into account the temporal dimension of legal concepts. Legal concepts can evolve over time due to the introduction of new legislation. The temporal dimension is modelled by adding a new ontological relation, "replaced by". This relation adds a date field indicating the substitution date to be added to the new concept. When a concept is replaced by a new one, the relations of the old concept are gathered and assigned to the new concept. The ELTS ontology also combines specific concepts into more broad and abstract concepts to represent complex entities. This is quite useful in cases when several EU directives (with specific concepts) are transposed into a single national implementing measure. The NIM merges all the concepts provided by the individual directives in a specific domain to define an abstract concept. These concepts are modelled by using an ontological relation called "grouped" which merges the context specific conceptualizations. The ontology is based on the open-source software from the Gene Ontology project [7].

The work in [13] presents Eunomos, a legal information retrieval and knowledge management system for Italian national law. The legal texts are downloaded from the national portals by using web spiders. The system consists of a database of about 70,000 Italian legislative documents which are converted into legislative XML format using a parser. Cross-references are extracted from the legal texts to build a network of texts citing each other. The three major components of the system include : a legal document management system, a legal knowledge management system and an external tier. The legal document management system consists of a database of legal texts in XML format, a database of the network of references represented

by a uniform resource name (URN), and a database with entries of provisions classified into different types. The knowledge management system consists of a database of terms with their associated concepts. It also consists of the relations connecting the different concepts. This database is connected to the document management system to map concepts to legal texts and provisions. The external tier consists of a database of user profiles with logins and areas of interest. The users are alerted when their legislation of interest is updated. Eunomos consists of a rule-based classification to classify normative provisions present in legal XMLs. The rule-based system utilizes linguistic features and priority rules to resolve ambiguities to assign the correct type to each provision. The rules basically develop characteristic patterns to identify a provision type. The rule-based classification system was evaluated on a corpus of 2306 provisions manually annotated by legal experts. The system achieved a very good performance in terms of precision and recall.

Eunomos also comprises a text similarity system to retrieve the most similar legislative texts for a given legislation in the database. The legislative documents are pre-processed and converted into term frequency-inverse document frequency (TF-IDF) vectors. The cosine similarity measure is then computed between the query document and the documents in the database. The system selected the number of relevant documents for each query document by using the categories associated with already labelled documents. The Eunomos system implements a document classification system to classify new and existing legislative texts into appropriate categories. A total of 15 categories were defined and a training set of 156 legislative documents was annotated. The system utilized bag-of-words TF-IDF and lemmatized nouns (extracted by TULE parser [68]) as features to build a Support vector machines (SVM) classifier for document classification. The model achieved a high accuracy of 0.9272 with 10-folds cross-validation. The system has found use in many areas such as financial, legal and public sector.

The authors in [103] proposed a methodology for semi-automatic transformation of a traditional web information retrieval system into a semantically aware retrieval system. The first step involves creation of an adequate semantic ontology. The chosen knowledge domain is that of legal documents belonging to the Portuguese Attorney General's Office. The documents were available in XML format. The Darpa Agent Markup Language (DAML + OIL) was chosen as the semantic language to represent the knowledge. Both structural and semantic objects were utilized in the process of creating ontology. The structural objects comprised document and classification classes of the ontology. The classification represents the domain subjects for each document. The semantic objects comprises classifying a set of selected verbs and nominal expressions for populating the ontology. The verbs and nomi-

nal expressions were extracted using a parser. Their frequency over the entire corpus was computed. The top verbs and nominal expressions were verified by legal experts and a subset of them were chosen. Afterwards, the verbs and nominal expressions were characterized into different concepts by normalizing the extracted verbs, subjects and direct objects. The concepts were assigned by using WordNet [83]. The discussed methodology was then utilized to annotated documents with semantic information from the ontology. The DAML + OIL code was added to the XML documents. Further, instances of the corresponding action of the verbs (present in the documents) with subject and direct object were added. Finally, a semantic inference engine was developed to address questions about the semantic representation of documents. The system utilized Prolog as the query language to translate both the ontology and the documents from DAML + OIL notation to Prolog rules. The authors show that the Prolog inference engine is able to answer queries relevant to the semantic content of the documents.

Casellas et al. [22] developed a legal information retrieval system for: i) frequently asked questions (FAQ) search and ii) case law search and browse system for the Spanish judges. The FAQ search system offers the functionality of searching a database of stored question-answer pairs through a natural language interface. The input query is given by the user in natural language in the form of a question. The system finds the best match between the input query and the stored question, so the stored answer can be retrieved for the input question. The system has three searching stages namely: i) Ontology domain detection ii) keyword and synonym detection iii) Ontology concept graph path matching. In addition, the system also utilizes cache memory to provide quick access to recently used data. The ontology domain detection stage involves determining the sub-domain of the target FAQ set by analyzing the input user question. The database of stored question-answers is manually classified into different legal sub-domains. A set of text analysis tools are applied to identify the sub-domain of the input question. The system then searches the candidate questions only in the sub-domain of the input question to reduce the search time. In the next stage, keyword matching is performed between the tokenized input question and FAQ's (in the sub-domain) tokens in the database. Along with exact matches, morphological and synonym matching is also done to filter out irrelevant candidate FAQs. A score is computed between the input question and the candidate FAQs. The score is computed on the basis of the number of matching words (including exact, synonyms and morphological matches). The candidate FAQs with scores below a certain threshold are removed from the candidate list of FAQs. The next stage is called ontology concept graph path matching. The input question is parsed to identify certain grammatical patterns which are then searched for in the

Ontology of Professional Judicial Knowledge (OPJK) [21]. The system then utilizes semantic distance algorithm to match the ontology graph path of the input question with the graph paths of the remaining candidate FAQs. The computation of semantic distance is based on the construction of several paths between the nodes of ontology. A variant of Dijkstra algorithm is utilized to find the shortest distance path between two nodes. The semantic distance is an indicator of semantic similarity. The candidate FAQ with the shortest distance is retrieved as the most relevant FAQ to the user query.

Dini et al. [32] presented a hybrid approach utilizing both lexical and legal conceptual representations for cross-language information retrieval. They developed a multilingual legal WordNet for both lawyers and laymen users. The multilingual WordNet has concepts in six European languages (Italian, Dutch, Portuguese, German, Czech and English). A tool was developed to extract legal definitions from the European Union law sources. A combination of structure-dependent and language-dependent techniques were used for automatic extraction of definitions from the directives. The group of definitions present mostly in article 2 of the directives are first identified. The group of definitions is then divided into individual definitions by means of paragraph division. However, not all definitions can be individually identified through this method because some definitions are covered in multiple paragraphs. The definitions obtained through paragraph division are stored into a term-definition pair. The terms are separated from the definition by utilizing punctuation (quotes, double quotes, commas, dash) and linguistic regular expressions ("means", "shall mean", etc.). The extracted definitions were added to the WordNet. The definitions for the domain of consumer protection law were also extracted from the national implementing measures (NIMs) and other relevant national acts. Two different kinds of concepts were represented in the system : lexical and legislative concepts. The lexical concepts are represented by the terms and the lexical meanings associated with them. The legal concepts are represented by the terms and the legal definitions associated with them. An interlingual index is defined to link the legal WordNet concepts through equivalence relations. Semantic relations such as "implemented as" were added to link EU concepts to the national implementations. A legal document index was developed to map the legal concepts to their sources in the legislative documents. A semi-automatic alignment technique was utilized to link the legal terms in the European directives in different languages. As a result a list of conceptually equivalent terms across different languages is obtained. In case of absence of equivalence relations, analogous hierarchical structures are used to find semantic relations between terms in two different languages.

The work in [113] presents a legal claim identification method based on a novel information retrieval method with hierarchically labelled

data. The system extracts the litigation claims from Intellectual Property (IP) pleading documents and also identifies the relevant entities inside each claim. The IP documents have two layers annotations. The first layer consists of claim text regions (text segments which consist of claims). The second layer consists of entities inside the claims such as patent numbers, claim numbers and claim types. The system first identifies the claim region within the document. Then it identifies the entities inside each claim region. For identifying the claim regions, sentences are used as atomic elements to define the granularity. However, words were used as atomic elements to identify entities. The first approach, called as Top-Down CRF utilizes two independent conditional random fields (CRFs) called as Claim CRF and Entity CRF. The Claim CRF model works at the document level and identifies (or predicts) the claim sequence for the entire document. It works at a sentence-level granularity as claims span over different sentences and paragraphs. The Entity CRF model works at the sentence level and considers word as the smallest atomic unit. The predicted claim sequences (beginning and inside of claims) are used as an input to the Entity CRF model to output the predicted entity sequence.

In the Top-Down approach the Claim CRF is ignorant of the entities while generating the claim sequence [113]. Therefore, the second approach called as Bottom-up CRF, is based on the hypothesis that the performance of Claim CRF could improve when entities are already tagged inside the claim texts. In the bottom-up approach, Entity CRF is first used to generate the predicted entity tag sequence. Then the Claim CRF is conditioned on the tagged entities and claim texts to produce the predicted tagged claim sequence. The authors also develop a joint hierarchical CRF which models both claim and entity layers jointly. The benefit of using this model is that the performance of Claim CRF can improve if it knows which types of entities are more likely to occur inside and outside the claims. Similarly, the performance of Entity CRF can improve if it knows a particular entity is inside or outside the claim. A variant of pseudo-likelihood is used to build this model because exact learning of joint probability is practically infeasible. A semi-supervised Bottom-up CRF was also developed where first the Entity CRF is trained only on the claim texts and is used to produce predicted entity labels on the training set. The labelled entries generated by the CRF are augmented with the labelled entities from outside the claims to generate a semi-supervised labelled dataset. The Claim CRF is then trained on this semi-supervised labelled data to generated the predicted claim sequences. The results indicate that the semi-supervised Bottom-up approach has the best performance (in terms of precision, recall and F-score) to identify claims as compared to other models. This is because it has entities from both inside and outside the claims to recognize the correlation between claims and entities (inside and outside) better than other models.

The work in [106] presents a query expansion method using lexical ontologies and user feedback to improve Boolean information retrieval. The authors developed a lexical ontology with 5500 terms, concepts and relations among them. The words in the query are searched in the ontology knowledge base to find synonymous relations. Weights are assigned to matching terms according to their relevance. The weights are added together to be used in boolean information retrieval. Linguistic pre-processing is quite limited and restricted to just truncation of terms. The system was evaluated on a corpus of State aid law using the lexical ontology. The results indicate that the query expansion method (with synonyms from the lexical ontology) improves over the standard boolean information retrieval method. The major advantage of such query expansion methods is the increase in the coverage of retrieved documents. However, such models also result in a lower precision. The system can also utilize more sophistical information retrieval methods based on vector space models or latent semantic analysis.

### 6.3.2 *Ontology Learning from Legal Texts*

The work presented in [62] utilizes natural language processing (NLP) techniques to identify legal ontology components such as, concepts and relations. Such tasks are necessary to reduce the manual effort of creating ontologies by hand. Such techniques are quite useful for legal concepts because their definition may vary depending on the source or context (and NLP techniques have shown good performance in such tasks). The system utilized 57 Codes from the French law. A syntactical analyzer was utilized to identify terms for the ontology such as noun phrases, verb phrases and adjective phrases. The terms which refer to legal concepts were chosen and only nouns and noun phrases were considered. A list of 300,000 terms was thus constructed. The authors investigated with different statistical indexing methods such as term frequency, inverse document frequency, TF-IDF (term frequency-inverse document frequency) and entropy to distinguish between legal and non-legal terms. They concluded that the above mentioned statistical indices were not useful to distinguish between legal and non-legal terms. Therefore, the authors separated a list of empty terms (common terms in the corpus such as title, chapter, book and general provisions). The remaining terms were referred to as fundamental legal terms and consisted of 16,681 entries. Text analysis tools were used to identify terms and the syntactical dependencies among them [62]. A set of coordination relations were identified with terms containing conjunctive phrases, "and" or "or". Furhter, statistical analysis was performed on the list of legal terms to identify relations among them. The statistical analysis is based on the hypothesis that semantically similar terms occur in similar contexts. A mutual information measure

is defined which takes into account the frequency of the term and its context words to quantify the dependency between the term and its context. This measure is used for weighting the context words and defining a vector to associate each term with the context words. The cosine similarity between two terms is computed by utilizing the vectors of the context words from both terms. The terms and relations identified using the above techniques are utilized to construct a graph called as "ontological resource. Semantic relations such as subsumption are inferred from the identified relations. For instance, a legal subsumption relation is inferred between a concept and its legal qualification. A general subsumption relation is inferred between a concept and a more abstract form of the concept. Other semantic relations inferred for ontology learning include: i) relations linking a concept to its components ii) relations linking concepts with different senses. One of the drawbacks of this work is that the use of statistical text analysis and weighting methods may miss out some semantic relations which could be useful for the ontology. Also some manual effort is required to infer the key semantic relations from the identified set of relations.

The work in [66] presents a methodology to automatically extract ontological knowledge from Italian legislative documents. A text-to-knowledge (T2K) system consisting of tools for natural language processing (NLP), statistical analysis and machine learning was developed [31]. The T2K system consists of a linguistic tool for Italian legal texts which carries out tokenization, morphological analysis and extracts chunks and dependency relations [37]. The next step involves the extraction of terms from the obtained chunks. Term extraction is the first and most important part of ontology learning because terms represent concepts which are modelled in the ontology [66]. Stopwords are removed and single terms are identified by frequency counts on the chunks. To identify multi-word terms, chunk patterns encoding syntactic templates for complex nominal terms are utilized. These include adjectival, prepositional and compounded modifications. The obtained list of complex patterns is ranked by using the log-likelihood ratio (a measure of relevance that quantifies the co-occurrences of the constituents of the complex terms). The system also takes into account the term variants to improve the indexing and retrieval. The types of term variants included in the system are: orthographic variants, inflectional variants, structural variants, variants including modifiers and variants combining different types of variation. The obtained terms are conceptually organized by using the dependency relations. The dependency-annotated text from the terms is used to extract the best verb-object and verb-subject pairs. A distributionally-based algorithm computes the measure of semantic similarity between between two terms by taking into account the the overlap between the best verbs for each term. A ranked list of similar terms is constructed for a particular

target term to identify the most relevant terms. The discussed methodology for ontology learning was applied on two legal corpora from Italian legislative texts: Consumer Law Corpus and Environmental Corpus.

The authors in [81] propose a methodology for learning ontology from legal texts by utilizing both the structure and content of the documents. The legal texts are converted from pdf to a well-formed XML format with articles, paragraphs and sentences. Pre-processing techniques (based on natural language processing) are applied on the texts to remove noise and highlight the important words. The dictionaries and grammars from the NOOJ software tool were applied to identify syntactic, lexical and morphological patterns [109]. This results in a list of candidate terms which are lemmatized by using WordNet. Further, statistical filtering is applied using TF-IDF weighting and weights assigned to terms based on their occurrence in the title of the legislative texts. The outcome is a list of domain terms (very important terms which may refer to domain concepts). The authors intend to utilize relational concept analysis (RCA) methodology for hierarchical extraction of content elements. RCA not only extracts the concepts from data (in form of attributes and relations) but also infers relations between related concepts. The presented methodology has been only implemented to the extent of identification of domain terms. Further work needs to be done to evaluate the performance of the proposed methodology for ontology learning.

The construction of both taxonomic and non-taxonomic relationships is one of the key challenges in ontology learning. An approach which addresses this challenge in domain-specific ontologies is discussed in [121]. The taxonomic relationships are acquired using WordNet while the non-taxonomic relationships are learned by using domain-specific texts. The system uses a rapid domain ontology development environment which takes a set of domain terms as input [60]. The input domain terms are mapped to WordNet by spell check. This results in a hierarchical structure of nodes from domain terms to the root of WordNet. The unnecessary nodes (parent-child and sibling relationships) are trimmed. The trimmed model is further refined by removing sub-trees which include other nodes along with best spell-matched nodes. Only the paths with just best spell-matched nodes are kept. A domain expert user is also consulted to construct the conceptual hierarchy by removing the sub-tress with other nodes. The non-taxonomic relationship learning module considers the word co-occurrences based on high-dimensional vectors. A set of high frequency 4-grams were extracted from the corpus to capture a bigger context of terms as compared to words. A co-occurrence matrix is constructed among the 4-grams and a context vector is computed for each 4-gram. The vector representation of words is computed as the sum of context vectors which occur at the position of the words within the

texts. The concept vectors are computed as the sum of the word vectors contained in the best matched synset (from WordNet) of domain terms. A pair of similar concepts are grouped together by computing a cosine similarity measure. The concepts with similarity score above a certain threshold are considered to have a non-taxonomic relationship between them. The information from the taxonomy relationship module along with the concept co-occurrence information is utilized by the rapid domain ontology development environment to determine the concept pairs hierarchically close to each other. The user input is also taken into account to remove unnecessary concept pairs. The proposed ontology learning was applied in the legal domain to learn both taxonomic and non-taxonomic relationships for contracts for the international sales of goods (CISG). In case of taxonomic relationship, around half of the concepts are extracted from WordNet. The precision of the system was around 30 percent due to cocnept drift. The low precision values are also because the system just checks for syntactical features. The semantic matches such as synonymous terms were not taken into account. The results of the non-taxonomic case study indicate that most of the extracted concept pairs were advisable for the legal domain.

### 6.3.3 *Named Entity Recognition (NER) in Legal Texts*

NER systems identify text spans of entity mentions. These mentions are generally assigned to Person, Organization and Location names. In named entity linking, the mentions are linked to entities in a knowledge base on the basis of contextual similarity. In [18], the authors developed a legal named entity recognizer and linker by aligning YAGO[8] (WordNet-and Wikipedia-based ontology) and the LKIF [49] ontology. The alignment was carried out manually by mapping a concept node in LKIF to its equivalent in YAGO. They utilized different models like support vector machines (SVM), Stanford Named Entity Recognizer (NER) [38], and neural networks and evaluated the system on a small sample of judgements from the European Court of Human Rights (ECHR). Their results indicate that LKIF level of generalization is not suitable for named entity recognition and classification as their system was unable to distinguish between the classes defined in LKIF. However, their NER system achieved a much better performance while distinguishing YAGO classes (even with a larger number of classes).

The authors in [33] developed a named entity recognition and classification system to recognize entities like judges, attorneys, companies, courts and jurisdictions in US case law, depositions, pleadings and other trial documents. They utilized dictionary lookup, contextual pattern rules and statistical models for identifying named entities. Each legal document was pre-processed by applying tokenization,

8 http://www.yago-knowledge.org/

zoning and line-blocking. The zoning method used different textual features to identify the caption parts within the legal documents. The line-blocking method was used to identify structural blocks of text by using a rule-based system. The authors utilized different rule-based and statistical taggers to identify court names, jurisdictions, document titles, judge names and document types. A named entity resolution system was developed by using a SVM classifier. The system was evaluated on manually and automatically acquired datasets of case law.

Current NER systems are based on conditional random fields (CRF) [61], which allow to train a unique model for the classification and recognition of named entities. In [38], the authors developed a CRF which used Gibbs sampling instead of the standard Viterbi algorithm. They demonstrated that the use of Gibbs sampling allowed the system to distinguish between mentions of organization or person on the basis of context, thus enforcing label consistency. The authors in [19], developed a NER system using AdaBoost. The system uses a window, along with a set of features (part-of-speech tags and dictionary of words) to capture the local context of a word. Ronan et al. [26] proposed a unified neural network model along with a CRF for NER and other natural language processing (NLP) tasks like part-of-speech tagging, chunking and semantic role labeling. A neural network (typically a long short-term memory) [48] generates a matrix of size num_words $\times$ num_tags, which contains the score for each tag. This matrix is passed as input for the CRF. The main advantage of these models is their capability to capture important features from the word embedding, thus improving the performance of the CRF model.

Table 6.3 presents an overview of different legal-tech use cases identified in this section. These include: ontology learning from legal texts, concept-based legal information retrieval and named entity recognition in legal texts. Ontology learning methods utilize natural language processing (NLP) techniques for extracting components at different levels of granularity: terms, concepts and relations [16]. The ontology learning methods have been differentiated on the basis of granularity and NLP techniques. The concept-based information retrieval methods derive concepts from an ontology or a knowledge base and link them to legal documents or provisions. We identified different domains where these techniques have been utilized. Named entity recognition systems also use an ontology or a knowledge base to associate mentions to relevant entities. A labeled dataset is created by manual annotations. Machine learning methods are then used to identify entity mentions.

Table 6.3: Concept and Ontology Based Information Retrieval in the Legal Domain

| **Legal-Tech Use Cases** | **Techniques** | **Granularity** |
|---|---|---|
| Ontology Learning from Legal Texts | Dictionaries and grammars are used to identify syntactic, lexical and morphological patterns. Domain terms identified by statistical filtering through TF-IDF. [81] | Terms |
| | NLP techniques: tokenization, morphological analysis and dependency relations to extract terms. Clustering of semantically related terms into concepts by using semantic similarity [66] | Terms, Synonyms, Concepts and Concept hierarchies |
| | NLP techniques such as syntactical analyzer to identify terms, statistical indexing methods (TF-IDF) to separate legal and non-legal terms. Text Similarity measure to identify semantic relations among terms. [62] | Terms, Synonyms, Concepts, Concept hierarchies and Relations |
| **Legal-Tech Use Cases** | **Techniques** | **Domain** |
| Concept-based Legal Information Retrieval | Legal concepts are linked to legislative documents for information retrieval [2] [13]. | Information Retrieval system for European and National Law |
| | Concept instances from legal ontology are associated with documents for enhancing keyword-based search for document retrieval [44]. | e-Government (real estate transactions of the Spanish government) |
| | Development of a multilingual legal WordNet by extracting legal definitions from European legislative documents. Equivalence relations were defined using an interlingual index and legal concepts were mapped to legislative documents [32]. | Cross-language information retrieval for legal concepts in European law |
| | Ontology domain detection, keyword and synonym detection, and ontology concept graph path matching to retrieve relevant questions for a query [22]. | Retrieval of Frequently Asked Questions (FAQs) and Case Law Search system for Spanish Judges |
| Named Entity Recognition (NER) in Legal Texts | Legal Named Entity Recognizer (NER) and Classifier by aligning YAGO and LKIF ontologies. NER was implemented by different classifiers such as SVM, CRF and neural networks [18] [49]. | Judgments of European Court of Human Rights and Wikipedia [18] |
| | Legal NER using dictionary lookup, contextual rules and statistical patterns [33]. | US Case Law [33] |

## 6.4 SUMMARY

In the literature review, we have discussed different text similarity, machine learning and concept-based methods for legal information retrieval. We identified the major legal-tech use cases, domains and feature set for the different approaches. The text similarity techniques have been mainly used for retrieval of similar judgments and patents, legal question answering, retrieval of legal statutes and provisions, automated conflict and dispute resolution, compliance check for contracts and trademark retrieval. Both lexical and semantic unsupervised text similarity techniques were used in majority of these applications. There were few works which used word embeddings and supervised machine learning models such as legal statutes retrieval and automated conflict resolution respectively. Machine learning techniques were mainly used for prediction of court cases, litigation likelihood and readability of legal texts. They were also used for classification of legal norms and extraction of key elements and relations from legislation and contracts. The prediction tasks used both textual and non-textual features whereas the extraction task utilized only textual features. We also presented concept and ontology based methods for legal information retrieval. The ontology learning tasks utilize natural language processing and statistical analysis to identify terms, concepts and relations. The automated construction of legal ontology is dependent on the distinction of legal and non-legal terms. The concept-based legal information retrieval systems utilize legal ontologies and knowledge bases to associate concepts to legal texts and provisions. The added semantic knowledge from the ontologies enhances the retrieval quality of traditional information retrieval systems. Named entity recognition systems identify legal entities such as courts, judges, contracts and acts in legal texts. They are built by training machine learning models over a labeled corpus.

We observed that the state-of-the-art information retrieval methods are quite successful in automating or semi-automating certain manual legal tasks. This is also demonstrated by the good performance of different natural language processing (NLP) and machine learning techniques discussed in this chapter. We identify some of the major research challenges in the development of legal information retrieval systems. The lack of legal datasets and gold standard corpus for machine learning and natural language processing applications is one of the biggest challenge for the development of data-driven legal information retrieval models. There are only a very few legal datasets available for research purposes. This is because development of legal datasets is time-consuming and expensive as it requires expertise of legal annotators. Some of the freely available legal datasets include the legal question answering dataset released by the Competition on

Legal Information Extraction/Entailment[9], the United States Supreme Court Database[10], the Patent Litigations dataset from the US Patent and Trademark Office[11], Google Patents Public Data[12] and the Case Law Access Project[13]. There is also a lack of semantic annotation and relation extraction tools for the legal domain. Most of the current works in natural language processing and machine learning rely on generic state-of-the-art tools. However, effective legal information retrieval systems must utilize domain oriented tools for semantic analysis of the legal text. The existing legal ontologies and vocabularies can be utilized for the development of such legal annotation tools. This was also demonstrated in [90] and [18].

The research work presented in this thesis has contributed to the development of a legal gold standard corpus of European directive articles and NIM provisions. It has also led to the development of a domain-specific concept recognition system for European and national legislation. Furthermore, in this thesis work, we have implemented text similarity systems by utilizing unsupervised, supervised (machine learning) and concept-based approaches. The unsupervised approaches consist of lexical and semantic models such as TF-IDF cosine, latent semantic analysis (LSA), latent dirichlet allocaion (LDA), unifying similarity measure (USM), provision vectors based on Word2vec and Fasttext and paragraph vectors. The supervised approaches consisted of machine learning classifiers such as Naive Bayes, Logistic Regression and Support Vector Machines (SVMs) with textual features (TF-IDF, LSA and LDA). In the concept-based approach we utilized word sense disambiguation and dictionary lookup to develop a text similarity system to identify transpositions. Therefore, this thesis work presents a comprehensive overview of the different text similarity approaches to automate the identification of national implementing measures (NIMs).

9 http://webdocs.cs.ualberta.ca/~miyoung2/COLIEE2017/
10 http://scdb.wustl.edu/
11 https://www.kaggle.com/uspto/patent-litigations
12 https://www.kaggle.com/bigquery/patents
13 https://case.law/about/

# 7 CONCLUSION AND FUTURE WORK

In this chapter, we present the conclusion and future work for this thesis work. The objective of the thesis was to develop legal information retrieval systems based on text similarity techniques for automated identification of national implementing measures of European Union directives. The identification of transpositions was carried out by retrieving the most semantically similar NIM provisions for a particular directive article.

## 7.1 OVERALL RESULTS SUMMARY

This section presents a summary of the overall results of the unsupervised text similarity models on the multilingual corpus of 43 directives and their corresponding NIMs. The supervised models were developed by training a subset of the gold standard for the multilingual corpus of 43 directives and their corresponding NIMs. Hence they were tested only on a subset of the entire corpus. Therefore, their performance cannot be compared directly with the unsupervised models which were evaluated on the entire multilingual corpus of 43 directives and their corresponding NIMs. The evaluation of the supervised text similarity models is discussed in Chapter 5.

Figure 7.1 presents the macro-average precision, recall and F-score of the best performing unsupervised text similarity measures over the multilingual corpus of 43 directives and their corresponding NIMs. The results show that TF-IDF Cosine had the highest macro-average F-Score for Ireland, Luxembourg and Italian legislation. The performance (F-Score) of the Babelfy-based similarity measure is competitive with the TF-IDF Cosine measure in Luxembourg and Italian Directive-NIM corpus. The Babelfy-based similarity measure has the second best F-score in the Luxembourg and Italian Directive-NIM corpus. LSA and USM_chars similarity measures also had a good performance but were outperformed by TF-IDF Cosine in all the three legislations. We also observe that the lexical and semantic methods of text similarity have a much better performance than the word and paragraph embedding models. This indicates that a majority of transpositions can be identified by highlighting important features using lexical weighting schemes (TF-IDF, USM) and semantic features (Babelfy concepts and latent dimensions of LSA). The word and paragraph embeddings models have shown to perform well when trained on a large corpus [82]. The multilingual legal corpus of directives and NIMs used to train word embeddings consisted of only 27365, 14365 and 16233

documents in English (English Directives + Irish legislation), French (French Directives + Luxembourg legislation) and Italian (Italian Directives + Italian legislation) respectively. This is because word and paragraph embeddings are prediction-based models and require sufficient training data to fit the large number of parameters. In case of a small corpus, word embedding models have been shown to be outperformed by count-based methods such as LSA [4]. This is also evident in our results. A more in-depth discussion about the performance comparison of different word embedding models is presented in section 3.7.

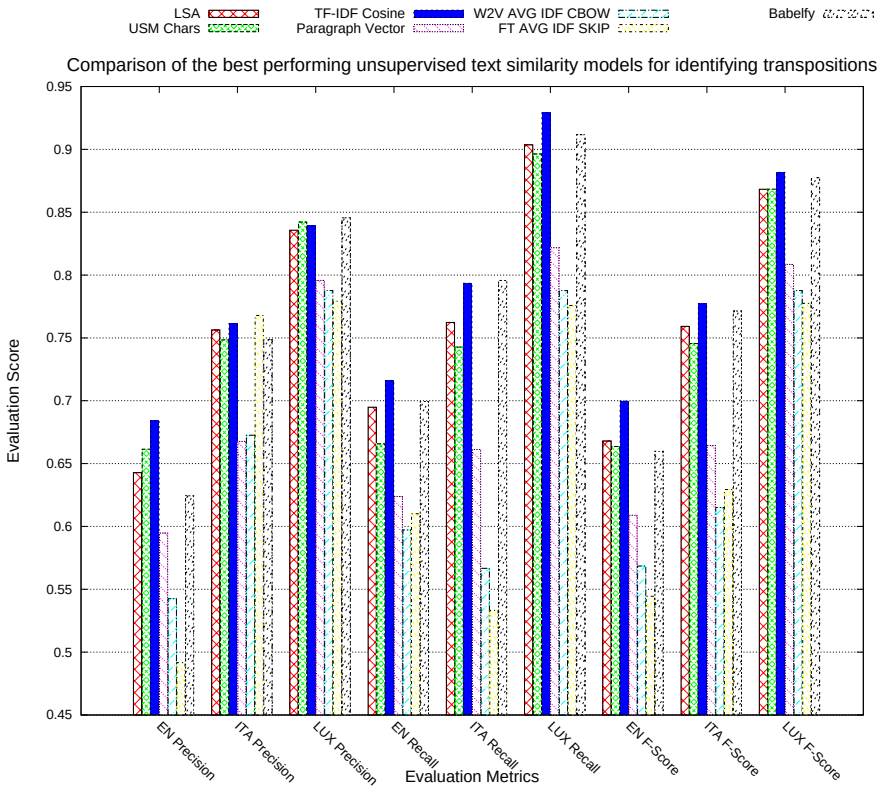


Figure 7.1: Macro-average scores for the best performing unsupervised text similarity models

The analysis of the results indicates two separate types of transpositions. We observe that some directive articles are generic while the others are specific. However, the national transposition provision in both cases are quite specific. We present an example to illustrate this. Table 7.1 presents an example of a generic directive article being transposed by a very specific national implementing provision. The directive article instructs the Member States to ensure that the explosives available on the market comply with the requirements of the directive. But it does not explicitly mention the requirements. The

NIM provision transposes the article by explicitly mentioning the requirements from different provisions. Such kind of transpositions were not identified by text similarity systems because there is a very low magnitude of semantic correlation between the two texts. This brings us back to the hypothesis discussed in Section 1.2 of Chapter 1. Text similarity techniques are only able to identify transpositions which share a certain magnitude of semantic similarity. Our results indicate that a majority of transpositions can be identified by utilizing text similarity techniques. The future research work can focus on development of other natural language processing models using cross-references and other derived features from legislations to identify these more complex cases of transpositions which are not identified by text similarity models.

Table 7.2 presents an example of transposition where both the directive article and NIM provision are quite specific. The NIM provision and directive article share a high magnitude of semantic similarity. Such transpositions are identified by the text similarity system. It is also important to note that many complex cases of transposition where directive articles and NIM provisions share only a mild degree of similarity were also identified by text similarity techniques ( Figure 2.13 in Section 2.2.4.3, Table 3.2 in Section 3.7 and Table 2.5 in Section 2.3.3).

Table 7.1: Article 4 of Directive (CELEX Number: 32014L0028) and its implementing NIM provision 4 from Ireland legislation (CELEX Number: 72014L0028IRL_239853)

| Article 4 of Directive | Provision 4 of Ireland NIM |
|---|---|
| Member States shall take the necessary measures to ensure that explosives may be made available on the market only if they comply with the requirements of this Directive. | A person shall not make available on the market any explosives unless the explosives— (a) satisfy the essential safety requirements set out in Schedule 1, (b) have been the subject of a conformity assessment procedure set out in paragraph (a) or (b) of Regulation 12; (c) have been submitted to a notified body for a conformity assessment under Regulation 19, (d) have passed the conformity assessment referred to in paragraph (a) or (b) of Regulation 12, (e) have an EU declaration of conformity drawn up in respect of them in accordance with Regulation 13, (f) have affixed to them the CE marking in accordance with Regulation 14 and Article 30 of the EC Regulation, (g) bear a unique identification in accordance with the Regulations of 2009, and (h) when properly stored and used for their intended purpose, do not endanger the health and safety of persons. |

Table 7.2: Article 12.3 of Directive (CELEX Number: 32002L0092) and its implementing NIM provision 19.7 from Ireland legislation (CELEX Number: 72002L0092IRL_34868)

| Article 12.3 of Directive | Provision 19.7 of Ireland NIM |
|---|---|
| Prior to the conclusion of any specific contract, the insurance intermediary shall at least specify, in particular on the basis of information provided by the customer, the demands and the needs of that customer as well as the underlying reasons for any advice given to the customer on a given insurance product. These details shall be modulated according to the complexity of the insurance contract being proposed. | Before entering into a contract with a customer relating to the provision of a particular insurance product, an insurance intermediary shall, on the basis of information provided by the customer, specify to the customer both the customer's demands and needs and the reasons underlying any advice given to the customer. The intermediary shall, so far as necessary, modify the information according to the complexity of the insurance contract being proposed. |

## 7.2 CONCLUSION

This thesis work presented a legal information retrieval system based on semantic textual similarity methods for automated identification of national implementations of European Union directives. We studied the control exercised by the European Commission for monitoring the transposition of directives and identified the need for automating this task. We identified two use cases (single jurisdiction and cross-border legal research) where our system would assist lawyers and Commission officials by automatically identifying transpositions at a fine-grained provision level. We developed and evaluated different text similarity models using both unsupervised and supervised approaches. The unsupervised methods included a comprehensive set of lexical, semantic, knowledge-based, embeddings-based and concept-based text similarity techniques. The text similarity systems utilized several features such as tokens, N-grams, topic models, word embeddings, paragraph embeddings and concept identifiers from external knowledge bases. A multilingual corpus of 43 directives and their corresponding NIMs from Ireland, Luxembourg and Italy was prepared. A gold standard mapping between the directive articles and NIM provisions (prepared by two legal researchers) was used to evaluate the text similarity systems. A thorough evaluation of different text similarity techniques for identification of transposing provisions over the multilingual corpus of 43 directives and their corresponding NIMs was carried out. The results indicate that the lexical and semantic techniques such as TF-IDF Cosine, LSA and USM_chars had a better performance than text similarity models based on word and paragraph

embeddings (Wor2vec, fastText and paragraph vectors). However, the word and paragraph embedding models were able to identify some complex cases of transposition (refer to Section 3.7) which were missed by lexical and semantic techniques.

We also developed a concept recognition system for European directives and national legislation. The system was based on conditional random fields (CRFs) and utilized IATE dictionary to identify concepts in both European and national legislative texts. We demonstrated that such system can be used to align legal terminology at European and national level. The concept recognition system thus formed the basis for a concept-based similarity measure to identify transpositions of European directives. We utilized both word-sense disambiguation (from Babelfy) and dictionary lookup from IATE to develop a concept-based text similarity measure for automated identification of national implementations of European directives. The results over the multilingual corpus indicate that the performance of the Babelfy-based text similarity system was competitive with the best performing TF-IDF Cosine measure. The gold standard mapping between directive articles and NIM provisions was used to build supervised text similarity models based on machine learning classifiers. We utilized Naive Bayes, Logistic Regression and Support Vector Machines (SVMs) classifiers with textual features based on vector transform such as TF-IDF, LSA and LDA. The results indicate that the SVM classifier with TF-IDF features had the best overall performance for the multilingual corpus.

The thorough evaluation of various text similarity approaches over three different legislations demonstrate the usability of the system to identify transpositions. The best performing unsupervised similarity measure, TF-IDF Cosine had macro-average F-Score of 0.8817, 0.7771 and 0.6997 for the 43 directive-NIM corpus of Luxembourg, Italian and English corpus respectively. The system also achieved high recall values of 0.929, 0.7934 and 0.7159 in the same order. Thus, the system is able to identify a high percentage of the actual transpositions. The knowledge engineer or system user would have to look for only a few number of transpositions which were missed by the system. The decent precision values (0.839, 0.7615 and 0.6842 for Luxembourg, Italian and English corpus respectively) achieved by the system result in efficiency gains. This means that most transpositions identified by the system need not be cross-checked by legal experts. Therefore, only a little manual effort would be required by legal knowledge engineers to remove false positive transpositions. The major advantage of this system is its ability to provide a head start to the legal expert involved in the transposition monitoring process by providing all the identified transpositions. The legal expert may use the system's output to validate the identified transpositions. The presence of such a system allows the legal expert to focus their expertise to identify ambiguous and complex cases of transpositions which are not captured by the

semantic features of the text similarity approach. Thus, the system has the potential to be effectively used as a legal support tool to aid the manual work of identifying transpositions for cross-border legal research for both the European Commission (EC) and legal professionals.

## 7.3 FUTURE WORK

In this section, we highlight the following future research directions:

**Integration of a provision type check classifier:** The identification of normative provision types has been investigated with different approaches such as semantic role labeling [51], machine learning [39], rule-based methods [73] and active learning [117]. The future research work can focus on the development of an appropriate provision type classifier for European directives and national legislation. The provision type identifier can be integrated with the text similarity system for the identification of transpositions. After a directive article and a NIM provision are deemed similar by the text similarity system, they can be checked for similar provision types by the provision type classifier. The presence of this additional check may help reduce the number of false positives and improve the precision. However, such a system is based on the assumption that the transposing NIM provisions may have the same provision type as that of the directive article. In the future work, it would be interesting to investigate this.

**Development and Population of Legal Ontologies:** The identification of the transposed NIM provisions for directive articles results in the alignment of European and national legislation at a fine-grained provision level. The aligned European and national legal provisions can be used to develop and populate legal ontologies to align concepts from European and national law. The legal taxonomy syllabus, an existing ontology which consists of concepts from European and national law can be enriched and populated with new concepts from the aligned legislation [2].

# ACADEMIC ACTIVITIES AND PUBLICATIONS DURING THE PHD

**Academic Activities**: During the Law, Science and Technology (LAST-JD) PhD program I attended several courses in the domain of Artificial Intelligence and Law and Legal Informatics at different consortium partners: University of Bologna, University of Turin, Autonomous University of Barcelona and University of Luxembourg. Some of the courses included Philosophy of Law, Semantic Web and Legal Ontologies, Privacy Enhancing Technologies, Data Protection Law and Argumentation Course. I also attended the Grid'5000 Winter School 2016 at INRIA Grenoble, France. I also participated and presented my research topic in the IT Law and Legal Informatics Summer School 2018 at the University of Saarland, Germany.

**Publications**: The outcomes of my research were disseminated in JURIX 2016, JURIX 2017, ICAIL 2017 and Artificial Intelligence and Law journal (2018). The major publications (first author) are listed below:

- Rohan Nanda, Luigi Di Caro and Guido Boella. **A Text Similarity Approach for Automated Transposition Detection of European Union Directives**. In Proceedings of the 29th International Conference on Legal Knowledge and Information Systems (JURIX 2016), Pages: 143-148, Volume 294, December 2016, Sophia Antipolis, France. Link: http:/ebooks.iospress.nl/volumearticle/45748

- Rohan Nanda, Luigi Di Caro, Guido Boella, Hristo Konstantinov, Tenyo Tyankov, Daniel Traykov, Hristo Hristov, Francesco Costamagna, Llio Humphreys, Livio Robaldo and Michele Romano. **A Unifying Similarity Measure for Automated Identification of National Implementations of European Union Directives**. In Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law (ICAIL 2017), ACM, Pages:149-158, June 2017, London, United Kingdom. Link: https://dl.acm.org/citation.cfm?doid=3086512.3086527

- Rohan Nanda, Giovanni Siragusa, Luigi Di Caro, Martin Theobald, Guido Boella, Livio Robaldo and Francesco Costamagna. **Concept Recognition in European and National Law**. In Proceedings of the 30th International Conference on Legal Knowledge and Information Systems (JURIX 2017), Pages:193-198, Volume 302, December 2017, Luxembourg. Link: http://ebooks.iospress.nl/publication/48062

- Rohan Nanda, Giovanni Siragusa, Luigi Di Caro, Guido Boella, Lorenzo Grossio, Marco Gerbaudo and Francesco Costamagna. **Unsupervised and Supervised Text Similarity Systems for Automated Identification of National Implementing Measures of European Directives**. In Artificial Intelligence and Law, October 2018. Link: https://doi.org/10.1007/s10506-018-9236-y

- Rohan Nanda, Adebayo Kolawole John, Luigi Di Caro, Guido Boella and Livio Robaldo. **Legal Information Retrieval Using Topic Clustering and Neural Networks**. In Proceedings of COLIEE 2017. 4th Competition on Legal Information Extraction and Entailment at the International Conference on Articial Intelligence and Law (ICAIL) 2017. Link: https://doi.org/10.29007/psgx

# BIBLIOGRAPHY


[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. "TensorFlow: A System for Large-Scale Machine Learning." In: *OSDI*. Vol. 16. 2016, pp. 265–283.

[2] Gianmaria Ajani, Guido Boella, Luigi Di Caro, Livio Robaldo, Llio Humphreys, Sabrina Praduroux, Piercarlo Rossi, and Andrea Violato. "The european legal taxonomy syllabus: a multilingual, multi-level ontology framework to untangle the web of european legal terminology." In: *Applied Ontology* (2017).

[3] Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preoţiuc-Pietro, and Vasileios Lampos. "Predicting judicial decisions of the European Court of Human Rights: A natural language processing perspective." In: *PeerJ Computer Science* 2 (2016), e93.

[4] Edgar Altszyler, Mariano Sigman, Sidarta Ribeiro, and Diego Fernández Slezak. "Comparative study of LSA vs Word2vec embeddings in small corpora: a case study in dreams database." In: *arXiv preprint arXiv:1610.01520* (2016).

[5] R. Angell, G. Freund, and P. Willett. "Automatic spelling correction using a trigram similarity measure." In: *Information Processing & Management* 19.4 (1983).

[6] Fatahiyah Mohd Anuar, Rossitza Setchi, and Yu-Kun Lai. "A conceptual model of trademark retrieval based on conceptual similarity." In: *Procedia Computer Science* 22 (2013), pp. 450–459.

[7] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. "Gene Ontology: tool for the unification of biology." In: *Nature genetics* 25.1 (2000), p. 25.

[8] Sonia Bergamaschi and Laura Po. "Comparing LDA and LSA Topic Models for Content-Based Movie Recommendation Systems." In: *International Conference on Web Information Systems and Technologies*. Springer. 2014, pp. 247–263.

[9] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*. 1st. O'Reilly Media, Inc., 2009. ISBN: 0596516495, 9780596516499.

[10] David M Blei, Andrew Y Ng, and Michael I Jordan. "Latent dirichlet allocation." In: *Journal of machine Learning research* 3.Jan (2003), pp. 993–1022.

[11] Guido Boella, Luigi Di Caro, and Livio Robaldo. "Semantic relation extraction from legislative text using generalized syntactic dependencies and support vector machines." In: *International Workshop on Rules and Rule Markup Languages for the Semantic Web*. Springer. 2013, pp. 218–225.

[12] Guido Boella, Luigi Di Caro, Leonardo Lesmo, and Daniele Rispoli. "Multi-label Classification of Legislative Text into EuroVoc." In: *Legal Knowledge and Information Systems: JURIX 2012: the Twenty-fifth Annual Conference*. Vol. 250. IOS Press. 2012, p. 21.

[13] Guido Boella, Luigi Di Caro, Llio Humphreys, Livio Robaldo, Piercarlo Rossi, and Leendert van der Torre. "Eunomos, a legal document and knowledge management system for the web to provide relevant, reliable and up-to-date information on the law." In: *Artificial Intelligence and Law* 24.3 (2016), pp. 245–283.

[14] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. "Enriching word vectors with subword information." In: *arXiv preprint arXiv:1607.04606* (2016).

[15] Michael T. Brannick. *Logistic Regression*. http://faculty.cas.usf.edu/mbrannick/regression/Logistic.html. [Online; accessed 9-July-2018].

[16] Paul Buitelaar, Philipp Cimiano, and Bernardo Magnini. *Ontology learning from text: methods, evaluation and applications*. Vol. 123. IOS press, 2005.

[17] Laurent Candillier, Frank Meyer, and Françoise Fessant. "Designing Specific Weighted Similarity Measures to Improve Collaborative Filtering Systems." In: *ICDM*. 2008.

[18] Cristian Cardellino, Milagro Teruel, Laura Alonso Alemany, and Serena Villata. "A low-cost, high-coverage legal named entity recognizer, classifier and linker." In: *Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law*. ACM. 2017, pp. 9–18.

[19] Xavier Carreras, Lluis Marquez, and Lluís Padró. "Named entity extraction using adaboost." In: *proceedings of the 6th conference on Natural language learning-Volume 20*. Association for Computational Linguistics. 2002, pp. 1–4.

[20] Danilo S. Carvalho, Vu Tran, Khanh Van Tran, and Nguyen Le Minh. "Improving Legal Information Retrieval by Distributional Composition with Term Order Probabilities." In: *COLIEE 2017. 4th Competition on Legal Information Extraction and Entailment*. Ed. by Ken Satoh, Mi-Young Kim, Yoshinobu Kano, Randy Goebel, and Tiago Oliveira. Vol. 47. EPiC Series in Computing. EasyChair, 2017, pp. 43–56. DOI: 10.29007/2xzw. URL: https://easychair.org/publications/paper/7Pv.

[21] Pompeu Casanovas, Núria Casellas, Christoph Tempich, Denny Vrandečić, and Richard Benjamins. "OPJK and DILIGENT: ontology modeling in a distributed environment." In: *Artificial Intelligence and Law* 15.2 (2007), pp. 171–186. ISSN: 1572-8382. DOI: 10.1007/s10506-007-9036-2. URL: https://doi.org/10.1007/s10506-007-9036-2.

[22] Núria Casellas, Pompeu Casanovas, Joan-Josep Vallbé, Marta Poblet, Mercedes Blázquez, Jesús Contreras, José-Manuel López-Cobo, and V. Richard Benjamins. "Semantic Enhancement for Legal Information Retrieval: Iuriservice Performance." In: *Proceedings of the 11th International Conference on Artificial Intelligence and Law*. ICAIL '07. Stanford, California: ACM, 2007, pp. 49–57. ISBN: 978-1-59593-680-6. DOI: 10.1145/1276318.1276328. URL: http://doi.acm.org/10.1145/1276318.1276328.

[23] Ilias Chalkidis, Ion Androutsopoulos, and Achilleas Michos. "Extracting Contract Elements." In: *Proceedings of the 16th Edition of the International Conference on Articial Intelligence and Law*. ICAIL '17. London, United Kingdom: ACM, 2017, pp. 19–28. ISBN: 978-1-4503-4891-1. DOI: 10.1145/3086512.3086515. URL: http://doi.acm.org/10.1145/3086512.3086515.

[24] Lixin Chen. "Do patent citations indicate knowledge linkage? The evidence from text similarities between patents and their citations." In: *Journal of Informetrics* 11.1 (2017), pp. 63–79.

[25] G. Ciavarini Azzi. "The slow march of European legislation: The implementation of directives." In: *European integration after Amsterdam: Institutional dynamics and prospects for democracy* (2000).

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

[27] *Conformity Checking, Milieu*. English. Milieu. accessed 25 May 2016, Retrieved from http://www.milieu.be/index.php?page=conformity-checking. URL: http://www.milieu.be/index.php?page=conformity-checking.

[28] Georgina Cosma and Mike Joy. "An approach to source-code plagiarism detection and investigation using latent semantic analysis." In: *IEEE transactions on computers* 61.3 (2012), pp. 379–394.

[29] Michael Curtotti, Eric McCreath, Tom Bruce, Sara Frug, Wayne Weibel, and Nicolas Ceynowa. "Machine learning for readability of legislative sentences." In: *Proceedings of the 15th International Conference on Artificial Intelligence and Law*. ACM. 2015, pp. 53–62.

[30] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. "Indexing by latent semantic analysis." In: *Journal of the American society for information science* 41.6 (1990), p. 391.

[31] Felice Dell'Orletta, Alessandro Lenci, Simone Marchi, Simonetta Montemagni, and Vito Pirrelli. "Text-2-knowledge: una piattaforma linguistico-computazionale per l'estrazione di conoscenza da testi." In: *XL Congresso Internazionale di Studi della Società di Linguistica Italiana*. 2009, pp. 285–300.

[32] Luca Dini, Wim Peters, Doris Liebwald, Erich Schweighofer, Laurens Mommers, and Wim Voermans. "Cross-lingual legal information retrieval using a WordNet architecture." In: *Proceedings of the 10th international conference on Artificial intelligence and law*. ACM. 2005, pp. 163–167.

[33] Christopher Dozier, Ravikumar Kondadadi, Marc Light, Arun Vachher, Sriharsha Veeramachaneni, and Ramdev Wudali. "Named entity recognition and resolution in legal text." In: *Semantic Processing of Legal Texts*. Springer, 2010, pp. 27–43.

[34] Matt Dunn, Levent Sagun, Hale Şirin, and Daniel Chen. "Early Predictability of Asylum Court Decisions." In: *Proceedings of the 16th Edition of the International Conference on Articial Intelligence and Law*. ICAIL '17. London, United Kingdom: ACM, 2017, pp. 233–236. ISBN: 978-1-4503-4891-1. DOI: 10.1145/3086512.3086537. URL: http://doi.acm.org/10.1145/3086512.3086537.

[35] M. Eliantonio, M. Ballesteros, M. Rostane, and D. Petrovic. *Tools for ensuring implementation and application of EU Law and evaluation of their effectiveness*. Tech. rep. 2013. URL: http://www.europarl.europa.eu/RegData/etudes/etudes/join/2013/493014/IPOL-JURI_ET(2013)493014_EN.pdf.

[36] Hongqin Fan and Heng Li. "Retrieving similar cases for alternative dispute resolution in construction accidents using text mining techniques." In: *Automation in construction* 34 (2013), pp. 85–91.

[37] Stefano Federici, Simonetta Montemagni, and Vito Pirrelli. "Shallow parsing and text chunking: a view on underspecification in syntax." In: *Cognitive science research paper-university of Sussex CSRP* (1996), pp. 35–44.

[38] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. "Incorporating non-local information into information extraction systems by gibbs sampling." In: *Proceedings of the 43rd annual meeting on association for computational linguistics*. Association for Computational Linguistics. 2005, pp. 363–370.

[39] Enrico Francesconi and Andrea Passerini. "Automatic classification of provisions in legislative texts." In: *Artificial Intelligence and Law* 15.1 (2007), pp. 1–17.

[40] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*. Vol. 1. 10. Springer series in statistics New York, NY, USA: 2001.

[41] Nir Friedman, Dan Geiger, and Moises Goldszmidt. "Bayesian network classifiers." In: *Machine learning* 29.2-3 (1997), pp. 131–163.

[42] William A Gale, Kenneth W Church, and David Yarowsky. "A method for disambiguating word senses in a large corpus." In: *Computers and the Humanities* 26.5-6 (1992), pp. 415–439.

[43] Gene H Golub and Christian Reinsch. "Singular value decomposition and least squares solutions." In: *Numerische mathematik* 14.5 (1970), pp. 403–420.

[44] Asunción Gómez-Pérez, Fernando Ortiz-Rodríguez, and Boris Villazón-Terrazas. "Ontology-based legal information retrieval to improve the information access in e-government." In: *Proceedings of the 15th international conference on World Wide Web*. ACM. 2006, pp. 1007–1008.

[45] HM Government. *Transposition Guidance: How to implement European Directives effectively*. 2013. URL: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/682752/eu-transposition-guidance.pdf.

[46] J. Hartung, G. Knapp, and B. Sinha. *Statistical meta-analysis with applications*. Vol. 738. John Wiley & Sons, 2011.

[47] Seongwan Heo, Kihyun Hong, and Young-Yik Rhim. "Legal content fusion for legal information retrieval." In: *Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law*. ACM. 2017, pp. 277–281.

[48] Sepp Hochreiter and Jürgen Schmidhuber. "Long short-term memory." In: *Neural computation* 9.8 (1997), pp. 1735–1780.

[49] Rinke Hoekstra, Joost Breuker, Marcello Di Bello, Alexander Boer, et al. "The LKIF Core Ontology of Basic Legal Concepts." In: *LOAIT* 321 (2007), pp. 43–63.

[50] Liangjie Hong and Brian D Davison. "Empirical study of topic modeling in twitter." In: *Proceedings of the first workshop on social media analytics*. ACM. 2010, pp. 80–88.

[51] Llio Bryn Humphreys. "Populating Legal Ontologies using Information Extraction based on Semantic Role Labeling and Text Similarity." PhD thesis. University of Luxembourg, 2016.

[52] Llio Humphreys, Cristiana Santos, Luigi Di Caro, Guido Boella, Leon Van Der Torre, and Livio Robaldo. "Mapping Recitals to Normative Provisions in EU Legislation to Assist Legal Interpretation." In: *JURIX*. 2015, pp. 41–49.

[53] Kishore Vama Indukuri, Anurag Anil Ambekar, and Ashish Sureka. "Similarity analysis of patent claims using natural language processing techniques." In: *Conference on Computational Intelligence and Multimedia Applications, 2007. International Conference on*. Vol. 4. IEEE. 2007, pp. 169–175.

[54] Thorsten Joachims. "Text categorization with support vector machines: Learning with many relevant features." In: *European conference on machine learning*. Springer. 1998, pp. 137–142.

[55] Daniel Martin Katz, II Bommarito, J Michael, and Josh Blackman. "Predicting the behavior of the supreme court of the united states: A general approach." In: *arXiv preprint arXiv:1407.6333* (2014).

[56] Tom Kenter and Maarten De Rijke. "Short text similarity with word embeddings." In: *Proceedings of the 24th ACM international on conference on information and knowledge management*. ACM. 2015, pp. 1411–1420.

[57] Mi-Young Kim, Ying Xu, and Randy Goebel. "Legal question answering using ranking svm and syntactic/semantic similarity." In: *JSAI International Symposium on Artificial Intelligence*. Springer. 2014, pp. 244–258.

[58] Sushanta Kumar, P Krishna Reddy, V Balakista Reddy, and Aditya Singh. "Similarity analysis of legal judgments." In: *Proceedings of the Fourth Annual ACM Bangalore Conference*. ACM. 2011, p. 17.

[59] Sushanta Kumar, P Krishna Reddy, V Balakista Reddy, and Malti Suri. "Finding similar legal judgements under common law system." In: *International Workshop on Databases in Networked Information Systems*. Springer. 2013, pp. 103–116.

[60] Masaki Kurematsu, Takamasa Iwade, Naomi Nakaya, and Takahira Yamaguchi. "DODDLE II: A domain ontology development environment using a MRD and text corpus." In: *IEICE transactions on information and systems* 87.4 (2004), pp. 908–916.

[61] John Lafferty, Andrew McCallum, and Fernando CN Pereira. "Conditional random fields: Probabilistic models for segmenting and labeling sequence data." In: (2001).

[62] Guiraude Lame. "Using NLP techniques to identify legal ontology components: concepts and relations." In: *Law and the Semantic Web*. Springer, 2005, pp. 169–184.

[63] Jörg Landthaler, Bernhard Waltl, Patrick Holl, and Florian Matthes. "Extending Full Text Search for Legal Document Collections Using Word Embeddings." In: *JURIX*. 2016, pp. 73–82.

[64] Pat Langley, Wayne Iba, Kevin Thompson, et al. "An analysis of Bayesian classifiers." In: *Aaai*. Vol. 90. 1992, pp. 223–228.

[65] Quoc Le and Tomas Mikolov. "Distributed representations of sentences and documents." In: *International Conference on Machine Learning*. 2014, pp. 1188–1196.

[66] Alessandro Lenci, Simonetta Montemagni, Vito Pirrelli, and Giulia Venturi. "Ontology learning from Italian legal texts." In: *Law, Ontologies and the Semantic Web* 188 (2009), pp. 75–94.

[67] Michael Lesk. "Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone." In: *Proceedings of the 5th Annual International Conference on Systems Documentation*. SIGDOC '86. Toronto, Ontario, Canada: ACM, 1986, pp. 24–26. ISBN: 0-89791-224-1. DOI: 10.1145/318723.318728. URL: http://doi.acm.org/10.1145/318723.318728.

[68] Leonardo Lesmo. "The turin university parser at evalita 2009." In: *Proceedings of EVALITA* 9 (2009).

[69] Wenhui Liao and Sriharsha Veeramachaneni. "A simple semi-supervised algorithm for named entity recognition." In: *Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing*. Association for Computational Linguistics. 2009, pp. 58–65.

[70] Yi-Hung Liu, Yen-Liang Chen, and Wu-Liang Ho. "Predicting associated statutes for legal problems." In: *Information Processing & Management* 51.1 (2015), pp. 194–211.

[71] Edward Loper and Steven Bird. "NLTK: The Natural Language Toolkit." In: *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*. ETMTNLP '02. Philadelphia, Pennsylvania: Association for Computational Linguistics, 2002, pp. 63–70. DOI: 10.3115/1118108.1118117. URL: https://doi.org/10.3115/1118108.1118117.

[72] Julie Beth Lovins. "Development of a stemming algorithm." In: *Mech. Translat. & Comp. Linguistics* 11.1-2 (1968), pp. 22–31.

[73] Emile de Maat and Radboud Winkels. "Automatic classification of sentences in dutch laws." In: *Legal Knowledge and Information Systems: JURIX 2008: the Twentieth First Annual Conference*. Vol. 21. IOS Press. 2008, p. 207.

[74] Laurens van der Maaten and Geoffrey Hinton. "Visualizing data using t-SNE." In: *Journal of machine learning research* 9.Nov (2008), pp. 2579–2605.

[75] Tom Magerman, Bart Van Looy, and Xiaoyan Song. "Exploring the feasibility and accuracy of Latent Semantic Analysis based text mining techniques to detect similarity between patent documents and scientific publications." In: *Scientometrics* 82.2 (2010), pp. 289–306. ISSN: 1588-2861. DOI: 10.1007/s11192-009-0046-6. URL: https://doi.org/10.1007/s11192-009-0046-6.

[76] Tarek Mahfouz, Amr Kandil, and Sukhrob Davlyatov. "Identification of latent legal knowledge in differing site condition (DSC) litigations." In: *Automation in Construction* 94 (2018), pp. 104–111.

[77] Arpan Mandal, Raktim Chaki, Sarbajit Saha, Kripabandhu Ghosh, Arindam Pal, and Saptarshi Ghosh. "Measuring Similarity Among Legal Court Case Documents." In: *Proceedings of the 10th Annual ACM India Compute Conference*. Compute '17. Bhopal, India: ACM, 2017, pp. 1–9. ISBN: 978-1-4503-5323-6. DOI: 10.1145/3140107.3140119. URL: http://doi.acm.org/10.1145/3140107.3140119.

[78] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press, 2008. ISBN: 0521865719, 9780521865715.

[79] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. "The Stanford CoreNLP Natural Language Processing Toolkit." In: *Association for Computational Linguistics (ACL) System Demonstrations*. 2014, pp. 55–60. URL: http://www.aclweb.org/anthology/P/P14/P14-5010.

[80] Mary L McHugh. "Interrater reliability: the kappa statistic." In: *Biochemia medica: Biochemia medica* 22.3 (2012), pp. 276–282.

[81] Imen Bouaziz Mezghanni and Faiez Gargouri. "Learning of legal ontology supporting the user queries satisfaction." In: *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2014 IEEE/WIC/ACM International Joint Conferences on*. Vol. 1. IEEE. 2014, pp. 414–418.

[82] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. "Efficient estimation of word representations in vector space." In: *arXiv preprint arXiv:1301.3781* (2013).

[83] George A Miller. "WordNet: a lexical database for English." In: *Communications of the ACM* 38.11 (1995), pp. 39–41.

[84] Vaughne Miller. "Legislating for Brexit: Statuotry Instruments Implementing EU Law." In: *Briefing Paper, House of Commons Library* Number 7867 (16 January 2017).

[85] Martin G Moehrle and Jan M Gerken. "Measuring textual patent similarity on the basis of combined concepts: design decisions and their consequences." In: *Scientometrics* 91.3 (2012), pp. 805–826.

[86] Martin G. Moehrle, Lothar Walter, Isumo Bergmann, Sebastian Bobe, and Svenja Skrzipale. "Patinformatics as a business process: A guideline through patent research tasks and tools." In: *World Patent Information* 32.4 (2010), pp. 291 –299. ISSN: 0172-2190. DOI: https://doi.org/10.1016/j.wpi.2009.11.003. URL: http://www.sciencedirect.com/science/article/pii/S0172219009001380.

[87] Andrea Moro, Alessandro Raganato, and Roberto Navigli. "Entity linking meets word sense disambiguation: a unified approach." In: *Transactions of the Association for Computational Linguistics* 2 (2014), pp. 231–244.

[88] Rohan Nanda, Luigi Di Caro, and Guido Boella. "A Text Similarity Approach for Automated Transposition Detection of European Union Directives." In: *Legal Knowledge and Information Systems - JURIX 2016: The Twenty-Ninth Annual Conference*. 2016, pp. 143–148. DOI: 10.3233/978-1-61499-726-9-143. URL: https://doi.org/10.3233/978-1-61499-726-9-143.

[89] Rohan Nanda et al. "A unifying similarity measure for automated identification of national implementations of european union directives." In: *Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law, ICAIL 2017, London, United Kingdom, June 12-16, 2017*. 2017, pp. 149–158. DOI: 10.1145/3086512.3086527. URL: http://doi.acm.org/10.1145/3086512.3086527.

[90] Rohan Nanda, Giovanni Siragusa, Luigi Di Caro, Martin Theobald, Guido Boella, Livio Robaldo, and Francesco Costamagna. "Concept Recognition in European and National Law." In: *Legal Knowledge and Information Systems - JURIX 2017: The Thirtieth Annual Conference, Luxembourg, 13-15 December 2017*. 2017, pp. 193–198. DOI: 10.3233/978-1-61499-838-9-193. URL: https://doi.org/10.3233/978-1-61499-838-9-193.

[91] Rohan Nanda, Kolawole John Adebayo, Luigi Di Caro, Guido Boella, and Livio Robaldo. "Legal Information Retrieval Using Topic Clustering and Neural Networks." In: *COLIEE 2017. 4th Competition on Legal Information Extraction and Entailment, held in conjunction with the 16th International Conference on Artificial Intelligence and Law (ICAIL 2017) in King's College London, UK*. 2017,

pp. 68–78. URL: http://www.easychair.org/publications/paper/347228.

[92] Rohan Nanda, Giovanni Siragusa, Luigi Di Caro, Guido Boella, Lorenzo Grossio, Marco Gerbaudo, and Francesco Costamagna. "Unsupervised and supervised text similarity systems for automated identification of national implementing measures of European directives." In: *Artificial Intelligence and Law* (2018), pp. 1–27.

[93] Roberto Navigli. "Word sense disambiguation: A survey." In: *ACM computing surveys (CSUR)* 41.2 (2009), p. 10.

[94] Roberto Navigli and Simone Paolo Ponzetto. "BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network." In: *Artificial Intelligence* 193 (2012), pp. 217–250.

[95] H. Niemann and M. G. Moehrle. "Car2X-Communication mirrored by business method patents: What documented inventions can tell us about the future." In: *2013 Proceedings of PICMET '13: Technology Management in the IT-Driven Services (PICMET)*. 2013, pp. 976–984.

[96] Helen Niemann, Martin G. Moehrle, and Jonas Frischkorn. "Use of a new patent text-mining and visualization method for identifying patenting patterns over time: Concept, method and test application." In: *Technological Forecasting and Social Change* 115 (2017), pp. 210 –220. ISSN: 0040-1625. DOI: https://doi.org/10.1016/j.techfore.2016.10.004. URL: http://www.sciencedirect.com/science/article/pii/S0040162516304048.

[97] Luis Polo Paredes, JM Rodrıguez, and Emilio Rubiera Azcona. "Promoting government controlled vocabularies for the Semantic Web: the EUROVOC thesaurus and the CPV product classification system." In: *Semantic Interoperability in the European Digital Library* (2008), p. 111.

[98] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python." In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[99] Martin F Porter. "An algorithm for suffix stripping." In: *Program* 14.3 (1980), pp. 130–137.

[100] Radim Řehůřek and Petr Sojka. "Software Framework for Topic Modelling with Large Corpora." English. In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. http://is.muni.cz/publication/884893/en. Valletta, Malta: ELRA, May 2010, pp. 45–50.

[101] Paolo Rosso, Santiago Correa, and Davide Buscaldi. "Passage retrieval in legal texts." In: *Journal of Logic and Algebraic Programming* 80.3-5 (2011), pp. 139–153.

[102] Stuart J Russell and Peter Norvig. *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited, 2016.

[103] José Saias and Paulo Quaresma. "Semantic enrichment of a web legal information retrieval system." In: *JURIX*. 2002, pp. 11–20.

[104] Jaromır Šavelka, Gaurav Trivedi, and Kevin D Ashley. "Applying an interactive machine learning approach to statutory analysis." In: *Proceedings of the 28th Annual Conference on Legal Knowledge and Information Systems (JURIX'15). IOS Press*. 2015.

[105] Asad Sayeed, Soumitra Sarkar, Yu Deng, Rafah Hosn, Ruchi Mahindru, and Nithya Rajamani. "Characteristics of document similarity measures for compliance analysis." In: *Proceedings of the 18th ACM conference on Information and knowledge management*. ACM. 2009, pp. 1207–1216.

[106] Erich Schweighofer, Anton Geist, et al. "Legal Query Expansion using Ontologies and Relevance Feedback." In: *LOAIT*. 2007, pp. 149–160.

[107] Fabrizio Sebastiani. "Machine learning in automated text categorization." In: *ACM computing surveys (CSUR)* 34.1 (2002), pp. 1–47.

[108] Wei Shen, Jianyong Wang, and Jiawei Han. "Entity linking with a knowledge base: Issues, techniques, and solutions." In: *IEEE Transactions on Knowledge and Data Engineering* 27.2 (2015), pp. 443–460.

[109] Max Silberztein. "NooJ: a linguistic annotation system for corpus processing." In: *Proceedings of HLT/EMNLP on Interactive Demonstrations*. Association for Computational Linguistics. 2005, pp. 10–11.

[110] Harold Spaeth, Lee Epstein, Ted Ruger, Keith Whittington, Jeffrey Segal, and Andrew D Martin. *Supreme Court Database Code Book*. 2014.

[111] Karen Sparck Jones. "A statistical interpretation of term specificity and its application in retrieval." In: *Journal of documentation* 28.1 (1972), pp. 11–21.

[112] Bernard Steunenberg and Mark Rhinard. "The transposition of European law in EU member states: between process and politics." In: *European Political Science Review* 2.3 (2010), pp. 495–520.

[113] Mihai Surdeanu, Ramesh Nallapati, and Christopher Manning. "Legal claim identification: Information extraction with hierarchically labeled data." In: *Workshop Programme*. 2010, p. 22.

[114] Kristina Toutanova and Christopher D Manning. "Enriching the knowledge sources used in a maximum entropy part-of-speech tagger." In: *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics-Volume 13*. Association for Computational Linguistics. 2000, pp. 63–70.

[115] Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. "Feature-rich part-of-speech tagging with a cyclic dependency network." In: *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*. Association for Computational Linguistics. 2003, pp. 173–180.

[116] Salvatore Trani, Diego Ceccarelli, Claudio Lucchese, Salvatore Orlando, and Raffaele Perego. "Dexter 2.0: an open source tool for semantically enriching data." In: *Proceedings of the 2014 International Conference on Posters & Demonstrations Track-Volume 1272*. CEUR-WS. org. 2014, pp. 417–420.

[117] Bernhard WALTL, Johannes MUHR, Ingo GLASER, Georg BONCZEK, Elena SCEPANKOVA, and Florian MATTHES. "Classifying Legal Norms with Active Machine Learning." In: *Legal Knowledge and Information Systems* (2017), p. 11.

[118] J. Wang, G. Li, and J. Fe. "Fast-join: An efficient method for fuzzy token matching based string similarity join." In: *Data Engineering (ICDE), 2011 IEEE 27th International Conference on*. IEEE. 2011.

[119] Xuerui Wang and Andrew McCallum. "Topics over Time: A non-Markov Continuous-time Model of Topical Trends." In: *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '06. Philadelphia, PA, USA: ACM, 2006, pp. 424–433. ISBN: 1-59593-339-5. DOI: 10.1145/1150402.1150450. URL: http://doi.acm.org/10.1145/1150402.1150450.

[120] Papis Wongchaisuwat, Diego Klabjan, and John O McGinnis. "Predicting litigation likelihood and time to litigation for patents." In: *Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law*. ACM. 2017, pp. 257–260.

[121] Takahira Yamaguchi. "Acquiring Conceptual Relationships from Domain-Specific Texts." In: *Workshop on Ontology Learning*. Vol. 38. 2001, pp. 69–113.

[122] Hsiang-Fu Yu, Fang-Lan Huang, and Chih-Jen Lin. "Dual coordinate descent methods for logistic regression and maximum entropy models." In: *Machine Learning* 85.1-2 (2011), pp. 41–75.

[123] Fabio Massimo Zanzotto, Marco Pennacchiotti, and Alessandro Moschitti. "A machine learning approach to textual entailment recognition." In: *Natural Language Engineering* 15.4 (2009), pp. 551–582.

[124] Yi Zhang, Alan L Porter, Zhengyin Hu, Ying Guo, and Nils C Newman. ""Term clumping" for technical intelligence: A case study on dye-sensitized solar cells." In: *Technological Forecasting and Social Change* 85 (2014), pp. 26–39.

[125] Yi Zhang, Lining Shang, Lu Huang, Alan L. Porter, Guangquan Zhang, Jie Lu, and Donghua Zhu. "A hybrid similarity measure method for patent portfolio analysis." In: *Journal of Informetrics* 10.4 (2016), pp. 1108 –1130. ISSN: 1751-1577. DOI: https://doi.org/10.1016/j.joi.2016.09.006. URL: http://www.sciencedirect.com/science/article/pii/S1751157715302169.

[126] Jaromír Šavelka and Kevin D. Ashley. "Transfer of Predictive Models for Classification of Statutory Texts in Multi-jurisdictional Settings." In: *Proceedings of the 15th International Conference on Artificial Intelligence and Law*. ICAIL '15. San Diego, California: ACM, 2015, pp. 216–220. ISBN: 978-1-4503-3522-5. DOI: 10.1145/2746090.2746109. URL: http://doi.acm.org/10.1145/2746090.2746109.