# A Text Similarity Approach for Automated Transposition Detection of European Union Directives

Rohan NANDA [a,1], Luigi DI CARO [a] and Guido BOELLA [a]
[a] *Department of Computer Science, University of Turin, Italy*

**Abstract.** This paper investigates the application of text similarity techniques to automatically detect the transposition of European Union (EU) directives into the national law. Currently, the European Commission (EC) resorts to time consuming and expensive manual methods like conformity checking studies and legal analysis for identifying national transposition measures. We utilize both lexical and semantic similarity techniques and supplement them with knowledge from EuroVoc to identify transpositions. We then evaluate our approach by comparing the results with the correlation tables (gold standard). Our results indicate that both similarity techniques proved to be effective in detecting transpositions. Such systems could be used to identify the transposed provisions by both EC and legal professionals.

**Keywords.** transposition, text similarity, EU legislation

## 1. Introduction

The effective transposition of European Union (EU) directives 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 directives. The European Commission (EC) is responsible for monitoring the national implementations to ensure their compliance with EU law. The transposition measures adopted by Member States in national legislation to achieve the objectives of the directive are known as national implementing measures (NIMs) [4]. The Commission monitors the NIMs (communicated by the Member States) to ensure that Member States have taken appropriate measures to achieve the objectives of the directive. The steps taken by the Commission to monitor NIMs include Conformity Checking and Correlation tables [7]. The Commission outsources the monitoring of NIMs to subcontractors and legal consulting firms [1]. The conformity check studies carried out by a team of competent legal experts, comprise legal analysis and concordance tables. The concordance tables identify the specific provisions of NIMs which implement a particular article of the directive. Correlation tables are prepared by the Member States to ensure that the directive is completely transposed. They identify the specific provisions of NIMs for each article of a directive in a tabular format. Correlation tables are generally not available to public as they are sent by Mem-

[1] Corresponding Author

ber States to the Commission as part of a confidential bilateral exchange. There is no agreed format or compulsory content for correlation tables [7].

These legal measures undertaken by the Commission to monitor NIMs are time-consuming and expensive [3]. 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. Therefore, there is 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 the national level.

This paper presents the first work in automated transposition detection of EU directives. The objective is to identify the specific provisions of NIMs which transpose a particular article of the directive. We study and compare the results from both lexical and semantic similarity techniques on five directives and their corresponding NIMs by evaluating them with a gold standard (correlation tables). We were restricted to study only five directives as we could find correlation tables for only certain NIMs in English (due to our lack of competency in other EU languages).

The rest of the paper is organized as follows. In the next section, we discuss the proposed approach for automated transposition detection of EU directives. Section 3 discusses the results and analysis. Section 4 presents the conclusion and future work.

## 2. Automated Transposition Detection of EU Directives

In this section, we describe our approach for automated transposition detection of EU directives (Figure 1). We utilized cosine similarity vector model (lexical similarity technique) to detect transposing provisions with similar words. Latent semantic analysis (semantic similarity technique) was chosen to detect transposing provisions with same semantics but different wordings. First of all, each group of directive and NIMs were stored in a format to adhere to the structure of their particular correlation table. This enabled us to compare our results with that of the correlation tables. From here on the term provision refers to both article (of Directive) and provision (of NIM). Preprocessing included removing punctuation, conversion to lowercase and tokenization. Further stop words were removed using NLTKs corpus of stopwords for English. NLTKs part-of-speech tagger (POS tagger) was used to filter out nouns, verbs and adjectives from the remaining set of tokens [2]. The tokens obtained after pre-processing were enriched with the knowledge from EuroVoc[2], a multilingual thesaurus of the European Union. The tokens in the corpus were enriched with synonym and near-synonym terms as per equivalence relationship of EuroVoc [8]. Afterwards, the set of new tokens are stemmed to reduce the inflectional forms of words. Each provision of the corpus is then represented in a bag-of-words format. It is a list of each token and its count in a particular provision. Further, we applied Term Frequency-Inverse Document Frequency (tf-idf) weighting scheme to all the provisions [10]. We implemented latent semantic analysis (LSA) by applying Singular Value Decomposition (SVD) to the tf-idf provision-token matrix. SVD decomposes the tf-idf matrix into separate matrices which capture the similarity between tokens and provisions across different dimensions in space [6].

[2] http://eurovoc.europa.eu/drupal/?q=abouteurovoc

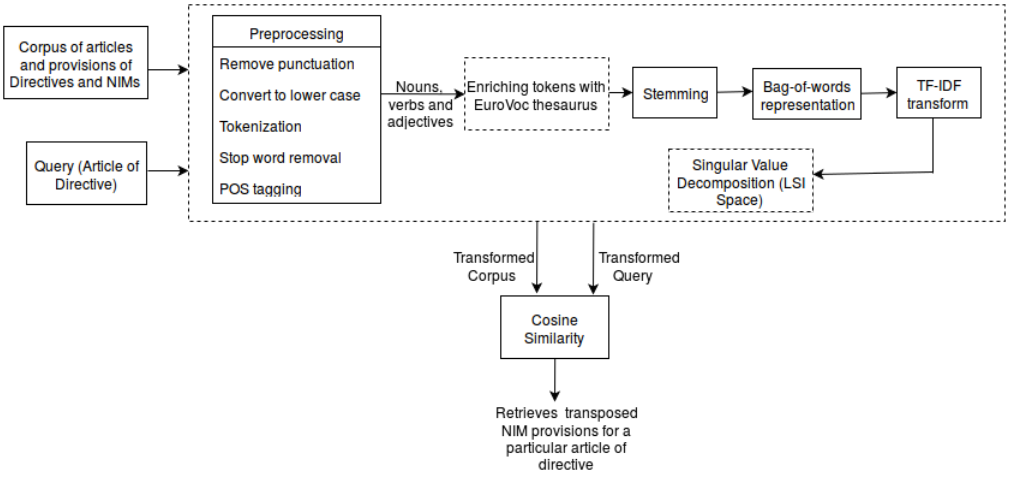


**Figure 1.** System architecture for automated transposition detection

The query (specific article of directive) is also transformed through the above steps. Since we wanted to evaluate the influence of adding knowledge from EuroVoc and also compare the performance of cosine similarity (CS) and LSA, we carried out the evaluation into four cases : (i) Cosine similarity (CS), (ii) Cosine similarity with EuroVoc, (iii) Latent semantic analysis (LSA), (iv) Latent semantic analysis with EuroVoc. It is important to note that dotted block of EuroVoc in Figure 1 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 as the cosine of the angle between the query vector and each provision vector in the reduced-dimensional space.

## 3. Results and Analysis

In this section, we study the results of transposition detection of five directives using the techniques discussed in the previous section. Table 1 represents the directives and NIMs under consideration. Directive1, Directive2, Directive3 and Directive4 are each transposed by one NIM. Directive5 is transposed by four NIMs. We observed that there were many cases where a particular article of a directive is transposed by multiple provisions of a NIM. Therefore, we also considered the cases where the provisions retrieved by our system are a subset of the transposed provisions as per the correlation tables. These are referred to as partial matches. We evaluate our system for both exact and partial matches. The implementation was carried out in Python and utilized NLTK and Gensim libraries [9][2].

We evaluate our system by computing the metrics: Precision, Recall and F-score (harmonic mean of precision and recall) for both exact and partial matches (partial matches are considered correct while computing precision and recall). We did not con-

**Table 1.** Directives and NIMs under consideration

| Directive-NIM group | Directives (CELEX number) | NIMs (Country and Number) |
|---|---|---|
| (Directive1, NIM1) | 32011L0085 | Ireland (Statutory Instrument No. 508/2013) |
| (Directive2, NIM2) | 32001L0024 | Ireland (Statutory Instrument No. 198/2004) |
| (Directive3, NIM3) | 31999L0092 | United Kingdom (Statutory Instrument No. 2776 of 7/11/2002) |
| (Directive4, NIM4) | 32003L0010 | United Kingdom (Statutory Instrument No. 1643 of 28/06/2005) |
| (Directive5, NIM5, NIM6, NIM7, NIM8) | 31998L0024 | United Kingdom (Statutory Instrument No. 2677 of 24/10/2002)<br>United Kingdom (Statutory Instrument No. 2676 of 24/10/2002)<br>United Kingdom (Statutory Instrument No. 2675 of 24/10/2002)<br>United Kingdom (Statutory Instrument No. 2776 of 07/11/2002) |

sider accuracy as we have very different number of true positives and true negatives resulting in an unbalanced dataset. We model and evaluate the system by considering the four cases for both partial and exact matches as mentioned in the previous section. Figure 2 shows the results of the transposition detection of all five directives. Appropriate threshold levels for transposition detection for both CS and LSA were determined through experimentation on the dataset.

The results in Figure 2 indicate no clear winner in terms of performance. However, we do make a few interesting observations. In terms of F-Score, CS achieves the best performance across all 5 directives. The performance of LSA was similar to CS in Directive1 and Directive2. However, it was outperformed by CS 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 [5]. 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 CS systems (as recall depends on false negatives). The same is observed through the graphs of Figure 2. In Directive3, Directive 4 and Directive 5 the recall of LSA is always lower than CS due to these higher false negatives. In Directive1 and Directive2 CS 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 CS in Directive1, Directive2, Directive3 and Directive5. However, the precision values of CS are quite close to LSA. In majority of the cases, LSA achieves higher precision (except in Directive4). While CS always achieves higher recall (except Directive1 and Directive2, where they have same recall). In terms of F-score, CS outperforms LSA (except Directive1 and Directive2, where they have same F-score). Overall in terms of all three metrics CS has the best performance due to higher recall and F-score and decent precision in all the directives.

We also observe from the results that the addition of knowledge from EuroVoc does not improve the performance of both CS and LSA. 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 CS and LSA results.

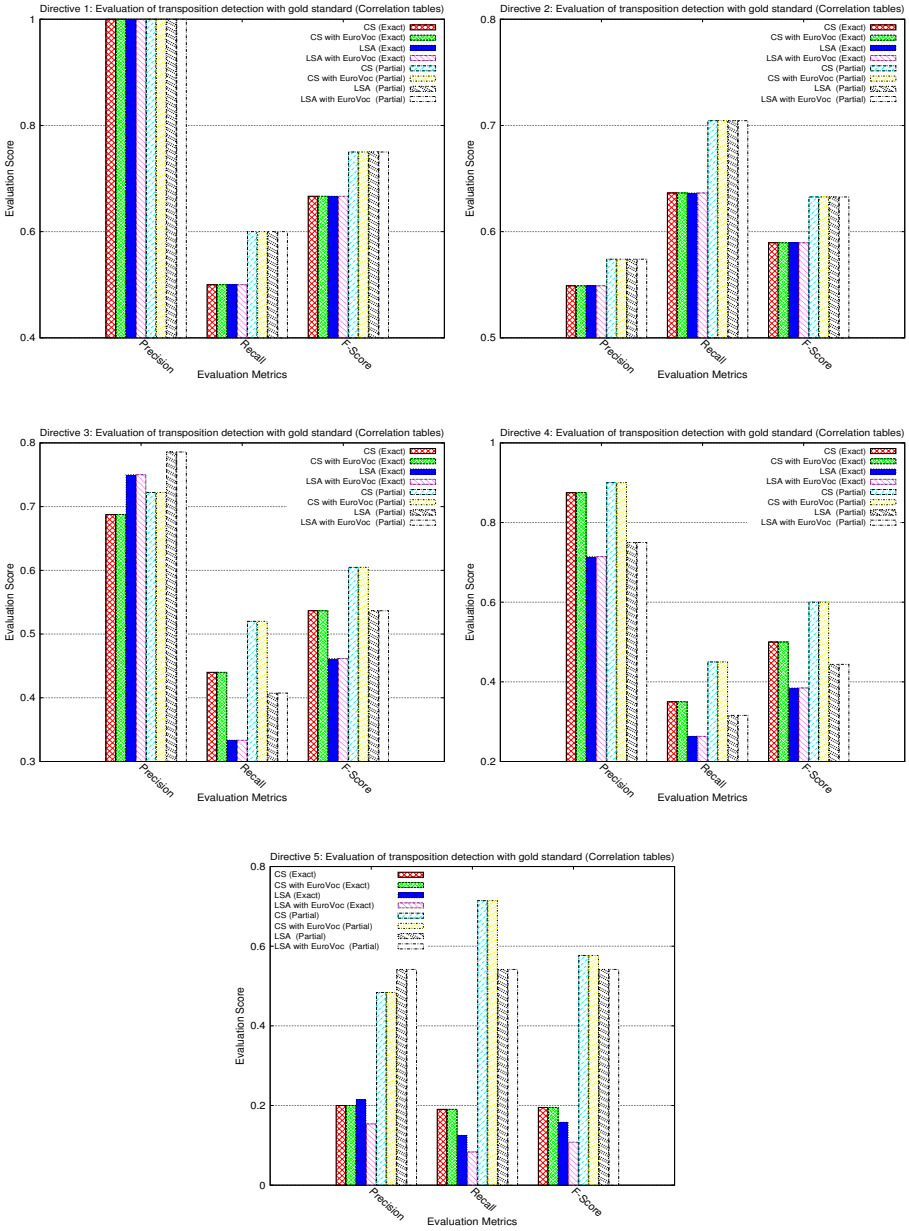


**Figure 2.** Evaluation of transposition detection with gold standard (Correlation tables)

## 4. Conclusion and Future Work

This paper presented the first work in automated transposition detection of EU directives by the application of text similarity approaches. We identified the need for a technological approach for monitoring NIMs. We investigated the application of both lexical (co-

sine similarity) and semantic (latent semantic analysis) similarity techniques in transposition detection. External knowledge from EuroVoc thesaurus was also used to supplement both similarity techniques. We evaluated our approach by comparing the results with the correlation tables. Our results indicate that both cosine similarity and latent semantic analysis were effective in detecting transposition. The overall performance of cosine similarity was superior to LSA in terms of F-score. Our initial experiments indicate that such systems can be useful for legal information retrieval to assist the Commission and legal professionals. Our future work will comprise using both n-gram models and quality phrase extraction to improve upon our current work. We also intend to study the transposition detection for a particular directive in different Member States. This would help us to characterize and compare how different Member States transpose the same directive with respect to their legal or domestic policy. We are also interested in developing a statistical language-independent model for transposition detection of directives across several Member States.

## Acknowledgements

Research presented in this paper is conducted as a PhD research at the University of Turin, within the Erasmus Mundus Joint International Doctoral (Ph.D.) programme in Law, Science and Technology. Luigi Di Caro has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skodowska-Curie grant agreement No 690974 for the project "MIREL: MIning and REasoning with Legal texts".

## References

[1] Milieu, Conformity checking. Electronic, accessed 8 September 2016, Retrieved from http://www.milieu.be/index.php?page=conformity-checking.

[2] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python*. " O'Reilly Media, Inc.", 2009.

[3] Giuseppe Ciavarini Azzi. The slow march of european legislation: The implementation of directives. *European integration after Amsterdam: Institutional dynamics and prospects for democracy*, pages 52–67, 2000.

[4] European Commission. Monitoring the application of Union law, 2014 Annual Report

[5] Georgina Cosma and Mike Joy. An approach to source-code plagiarism detection and investigation using latent semantic analysis. *IEEE transactions on computers*, 61(3):379–394, 2012.

[6] Susan T Dumais. Latent semantic analysis. *Annual review of information science and technology*, 38(1):188–230, 2004.

[7] Mariolina Eliantonio Marta Ballesteros, Rostane Mehdi and Damir Petrovic. Tools for ensuring implementation and application of eu law and evaluation of their effectiveness, July 2013.

[8] 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. *Semantic Interoperability in the European Digital Library*, page 111, 2008.

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

[10] Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21, 1972.