A Unifying Similarity Measure for Automated Identification of National Implementations of European Union Directives

Rohan Nanda
University of Turin
nanda@di.unito.it

Luigi Di Caro
University of Turin
dicaro@di.unito.it

Guido Boella
University of Turin
guido@di.unito.it

Hristo Konstantinov
APIS, Bulgaria
hristo.konstantinov@gmail.com

Tenyo Tyankov
APIS, Bulgaria
tencho@apis.bg

Daniel Traykov
APIS, Bulgaria
dtraykov@apis.bg

Hristo Hristov
APIS, Bulgaria
hristohv@apis.bg

Francesco Costamagna
University of Turin
francesco.costamagna@unito.it

Llio Humphreys
University of Turin
lliobh@gmail.com

Livio Robaldo
University of Luxembourg
livio.robaldo@uni.lu

Michele Romano
University of Zurich
micheleromano87@gmail.com

ABSTRACT

This paper presents a unifying text similarity measure (USM) for automated identification of national implementations of European Union (EU) directives. The proposed model retrieves the transposed provisions of national law at a fine-grained level for each article of the directive. 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 domain 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.

CCS CONCEPTS

•**Information systems** → **Information retrieval;**

KEYWORDS

European law, legal information retrieval, transposition

ACM Reference format:
Rohan Nanda, Luigi Di Caro, Guido Boella, Hristo Konstantinov, Tenyo Tyankov, Daniel Traykov, Hristo Hristov, Francesco Costamagna, Llio Humphreys, Livio Robaldo, and Michele Romano. 2017. A Unifying Similarity Measure for Automated Identification of National Implementations of European Union Directives. In *Proceedings of ICAIL '17, London, United Kingdom, June 12-16, 2017,* 10 pages.
DOI: 10.1145/3086512.3086527

1 INTRODUCTION

The effective application of European Union (EU) Law is mandatory for Member States and it is key to achieving EU policy objectives. Member States have the responsibility for ensuring correct and timely implementation of EU law. Among the three major EU legal instruments (directives, regulations and decisions) this paper studies the transposition of directives into national law. This is because directives are not directly applicable and Member States need to pass legislation to implement them into national law. 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 of implementation.

Each directive is associated with a deadline by which Member States must implement national transposition measures. These transposition measures are called national implementing measures (NIMs). Member States send the text of the NIMs to the European Commission (EC) . The Commission then examines them to ensure Member States have taken appropriate measures to achieve the objectives of the directive. The Commission uses conformity checking studies and correlation tables to monitor the transposition of directives in different Member States [8]. The conformity checking studies are carried out by subcontractors and legal firms and comprise legal analysis and concordance tables. Correlation tables identify the specific provisions of NIMs that transpose each article of a directive in a tabular format. They are prepared by Member States and sent to the Commission for review. The current transposition monitoring methods are time-consuming and

ICAIL '17, London, United Kingdom
© 2017 ACM. 978-1-4503-4891-1/17/06…$15.00
DOI: 10.1145/3086512.3086527

expensive, especially for cross-border and comparative legal research at European and national levels [7]. 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 to evaluate whether the obligations of the directive have been correctly transposed or not. In this paper, 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 domain 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). We provide two use cases where our system would assist legal practitioners 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.

The rest of the paper is organized as follows. In the next section, we discuss the related work. Section 3 describes the proposed model. Section 4 discusses the results and analysis. The paper concludes in Section 5.

2 RELATED WORK

In this section, we discuss state-of-the-art methods for short text similarity as we are interested in finding text similarity between precise and short legal texts (provisions in our case). In [14], the authors investigated the application of existing text similarity techniques to automatically identify the transposed NIM provisions. They utilized cosine similarity and latent semantic analysis (LSA) to identify transpositions in English legislation. Their results indicate that cosine similarity achieved better performance with a higher F-score. Humphreys et al. [10] also used cosine similarity for mapping recitals to provisions in EU legislation. Their research showed that the presumed similarity between recitals and provisions can be identified using text similarity systems. However, manual verification would be required to remove the invalid mappings suggested by the system. The authors in [13] investigated the application of knowledge-based and corpus-based measures of text similarity for automatic short answer grading. They demonstrated that both measures were effective for the task of short answer grading. The best performance was achieved by LSA. [17] utilized latent Dirichlet allocation (LDA) to compute similarity at sentence level. They observed that topic sparseness between texts leads to short distances (which implies high similarity scores). The proposed LDA-based semantic similarity model outperformed LSA when tested on the Microsoft Research Paraphrase Corpus.

The work in [12] investigated the application of both corpus and knowledge-based methods on short texts. The results show that both methods outperform lexical measures. They used the Microsoft Paraphrase corpus for evaluation. The pointwise mutual information measure achieved the best performance. In [2], a hybrid text similarity model for short texts was proposed based on WordNet (as a knowledge base) and a short corpus.The system had a comparable performance with state-of-the-art methods. The authors in [5] proposed a combined similarity measure by incorporating N-gram based similarity and concept based similarity (using WordNet). The resulting similarity was computed as a geometric mean of both similarity values. In [11], the authors proposed a hybrid similarity measure which combines the longest common subsequence string matching algorithm with a variant of pointwise mutual information algorithm. The proposed system achieved similar performance to another hybrid similarity measure (which combined corpus-based and knowledge-based measures). However, the proposed measure had lower time complexity as it did not use WordNet.

3 THE PROPOSED MODEL: A UNIFYING TEXT SIMILARITY MEASURE

In this section, we discuss the proposed model for automated identification of transposed NIM provisions of EU directives. Manual analysis of the 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. In such cases, an article is transposed by two or more provisions.

(1)-(3) motivated us to develop a specific model for automated identification of NIM provisions. We define a similarity measure for each observation and then propose a unifying similarity measure to take into account (1)-(3). The unifying measure 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.

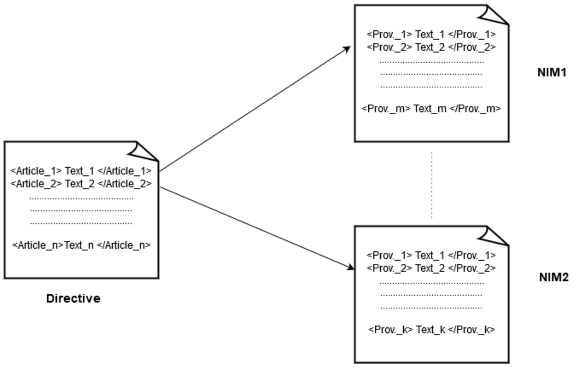

<Prov_1> Text_1 </Prov_1>
<Prov_2> Text_2 </Prov_2>
<Prov_m> Text_m </Prov_m>
NIM1
<Article_1> Text_1 </Article_1>
<Article_2> Text_2 </Article_2>
<Article_n>Text_n </Article_n>
Directive
<Prov_1> Text_1 </Prov_1>
<Prov_2> Text_2 </Prov_2>
<Prov_k> Text_k </Prov_k>
NIM2

Figure 1: Articles of a directive are compared with NIM provisions to retrieve the most semantically similar provisions

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 [10]. The cosine similarity between the vectors of Article *A* and provision *P* is computed as follows:

$$C(A,P) = \frac{A.P}{|A||P|} \tag{1}$$

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 (the implementation details are discussed in section 3.1).

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 [1]:

$$N(A,P) = \frac{sharedgrams}{totalgrams} \tag{2}$$

Here, *sharedgrams* is the number of 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 [1]:

$$N(A,P) = \frac{totalgrams^{warp} - unsharedgrams^{warp}}{totalgrams^{warp}} \tag{3}$$

where,

$$unsharedgrams = totalgrams - sharedgrams \tag{4}$$

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.3. 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 [18]. Then the intersection set, *I* of sorted tokens in *A* and *P* is computed as:

$$I = A \cap P \tag{5}$$

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.

$$A = I \cup R_A \tag{6}$$

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

$$P = I \cup R_P \tag{7}$$

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 [18]. 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 arthmetic mean of cosine similarity *CS(A,P)*, N-gram similarity *N(A,P)* and approximate similarity *AS(A,P)* as follows:

$$USM(A,P) = \frac{w_1 * CS(A,P) + w_2 * N(A,P) + w_3 * AS(A,P)}{w_1 + w_2 + w_3} \tag{8}$$

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 [9]. Each similarity measure is weighted in inverse proportion to its variance. The weight w_i for each similarity measure is thus given as:

$$w_i = \frac{1}{\sigma_i^2} \tag{9}$$

Here, σ_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 [6]. 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 transposition detection in English legislation in practice [14].

3.1 System Description

In this section, we describe the system implementation, including structuring, pre-processing and selection of features. Each Directive-NIM pair (in the same language) was stored in the same structure as their particular correlation table. This enabled us to evaluate our model with the correlation tables. This was carried out for each Directive-NIM pair because correlation tables have no standard way of structuring directives and NIMs. It is important to mention that while computing the similarity, both directive and NIM are in the same language (Figure 1).

From here on, the term provision refers to both article (of Directive) and provision (of NIM). The next step is pre-processing of the data. This consists of a number of steps to remove noise from the text. The punctuation was removed and the text was converted to lower case. Then tokenization was carried out to extract single words from the text. The stopwords were removed using NLTK's corpus of stopwords for English, French and Italian depending on the language of directive and NIM being considered. The outcome of feature selection suggested that keeping all parts of speech in the text yielded the best results. This is because provisions are precise texts. If we consider only certain parts of speech like nouns and verbs then the system loses some important features which are present in other parts of speech. After pre-processing, each provision in the Directive-NIM corpus is represented in a bag-of-words format. It is a list of each token and its frequency in a particular provision. Then we applied the Term Frequency-Inverse Document Frequency (tf-idf) weighting scheme to all the provisions [16]. 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 the previous section. The approximate similarity was also computed on the Directive-NIM corpus obtained after pre-processing. Further this corpus was tokenized and then the approximate similarity was computed as discussed in the previous section. 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.

4 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. The extended English language corpus (four additional directives) was used to thoroughly evaluate the performance of USM on additional directives. Results are discussed in subsection 4.1, 4.2, and 4.3 below. 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.

The EUR-Lex portal provides a list of NIMs which are adopted by the Member States and communicated to the Commission for a particular directive. The NIMs for each directive were identified as per the information from EUR-Lex. However, in some cases our correlation domain experts discovered that some NIMs mentioned in EUR-Lex were outdated and also preceded the date of entry into force of the correspondent directive. A possible reason for the presence of these extra NIMs on EUR-Lex is that they probably represent the entire national normative framework of the discipline mentioned by the directive. For this reason, usually only one national implementation measure truly corresponded to each directive. Other NIMs mentioned on the EUR-Lex website were not included in the official correlation tables (prepared by the Member States) for Ireland and the UK, and as such were not used in our experiments.

We observed from the correlation tables that there were some cases when a particular article is transposed by multiple NIM provisions. Thus, there was a need to consider the cases where the transpositions identified by the system are a subset of actual transpositions (as per the correlation table). Therefore, we carried out two evaluations : strict and lenient. In strict evaluation, only exact matches of the results of our system with the correlation tables is considered as a true positive (TP). In lenient evaluation, a partial match with the correlation table is also considered as a true positive. Lenient evaluation is probably more appropriate because recall is more important than precision in our task. It is important to identify as many transpositions as possible, even if they don't match the exact set of transpositions in the correlation tables. We evaluate our model for both strict and lenient evaluation by computing Precision, Recall and F-score (the harmonic mean of precision and recall). Accuracy was not considered as a reliable measure of evaluation because we have very different number of true positives and true negatives resulting in an unbalanced dataset. Accuracy is not a fair metric of evaluation in such cases. We experimented with different threshold values by incrementing the threshold from low to high values in the range of 0 to 1 (as the range of the similarity measure is between 0 and 1). The threshold which yielded the most number of true positives was chosen. This is because the objective of the system is to identify and retrieve as much transpositions as possible. Then precision, recall and F-score were computed for this threshold value. In case of equal number of maximum true positives, the threshold value which provides the maximum F-score was chosen.

Table 1: 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)

4.1 Results on the Multilingual corpus

Table 1 displays the directives and NIMs considered in the multilingual corpus. Figures 2 and 3 show the results of automated identification of NIM provisions by the proposed model for strict and lenient evaluation respectively. 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 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 average precision, recall and F-score measures across all directives (Figure 4). The average of the evaluation metrics across all directives indicate that the Luxembourg Directive-NIM pairs consistently achieved better recall and F-score than their Italian and English counterparts in both strict and lenient evaluation (Figure 4). 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 5, 6 and 7 respectively).

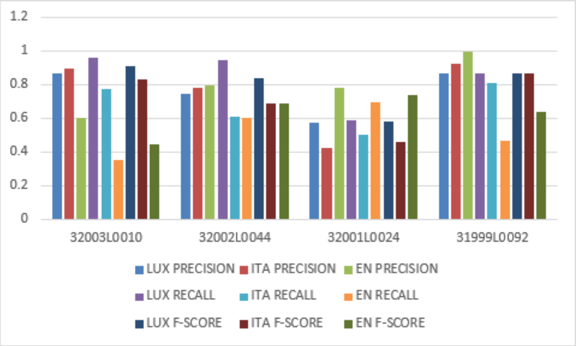


Figure 2: Results of strict evaluation of automated identification of NIM provisions by USM on the multilingual corpus

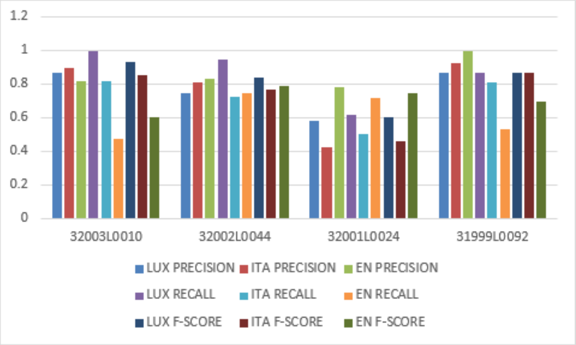


Figure 3: Results of lenient evaluation of automated identification of NIM provisions by USM on the multilingual corpus

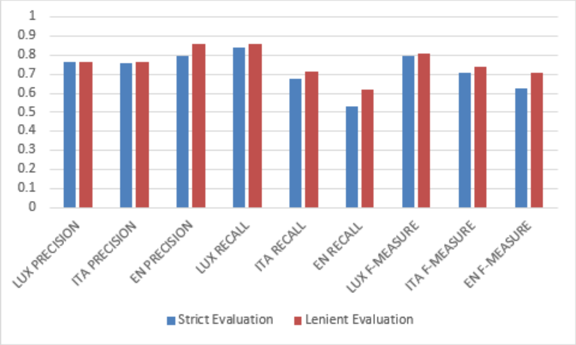
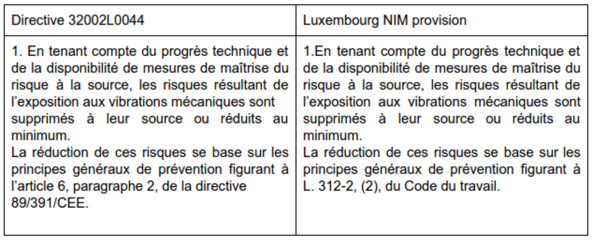


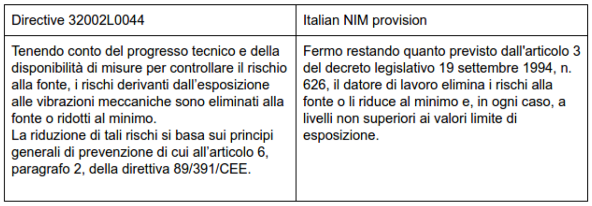
Figure 4: Average precision, recall and F-score for USM across all directives in the multilingual corpus

In the case of Luxembourg (Figure 5), 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 6), 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 (referred 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 7), both the directive and NIM provision convey the same



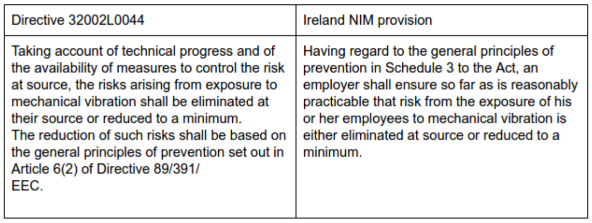
Directive 32002L0044	Luxembourg NIM provision
1. En tenant compte du progrès technique et de la disponibilité de mesures de maîtrise du risque à la source, les risques résultant de l'exposition aux vibrations mécaniques sont supprimés à leur source ou réduits au minimum. La réduction de ces risques se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

Figure 5: Article 5.1 of Directive CELEX 32002L004 and corresponding NIM provision 5.1 of Luxembourg



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. 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 6: 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. The reduction of such risks shall be based on the general principles of prevention set out in Article 6(2) of Directive 89/391/ 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.

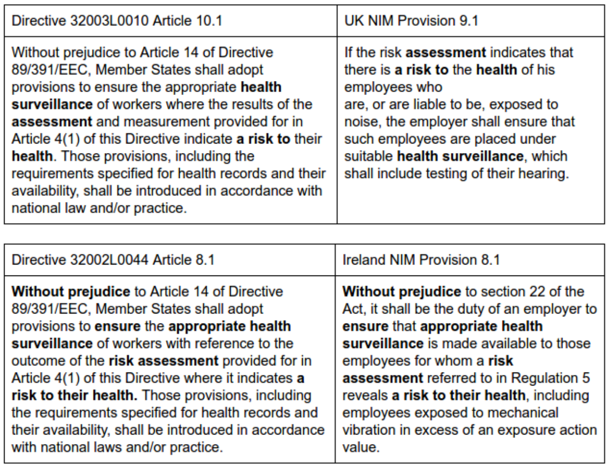
Figure 7: Article 5.1 of Directive CELEX 32002L004 and corresponding NIM provision 6.1 of Ireland

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 (lenient evaluation) 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 8 shows the transposition of two very similar articles from these two directives. We observe that the content of both articles is almost the same. 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 8 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 4 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 languages. This demonstrates that our model could be scalable for identifying transpositions in an automated way in different legal systems. It has the potential to be effectively used as a legal support tool for identifying transpositions in an automated way for cross-border legal research for both the European Commission (EC) and legal professionals. The high precision values (average of 0.767, 0.765 and 0.858 for French, Italian and English Directive-NIM pairs respectively, in lenient evaluation) achieved by the automated system result in efficiency gains. This means in practical terms most transpositions identified by the system need not be cross-checked by legal experts. Due to the high average precision values, only a little manual effort would be required by legal knowledge engineers to remove false positive transpositions. The decent recall values (average of 0.857, 0.713 and 0.617 for French, Italian and English Directive-NIM pairs respectively, in lenient evaluation) suggest that the system is able to identify most of the transpositions for each directive. The legal experts may only need



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

to identify a few transpositions manually for the false negative cases, resulting in considerable efficiency gains. Thus, our system could be a useful and efficient support tool to aid the manual work of identifying transpositions. The results show that further work is required to achieve a higher recall (especially for Italian and English legislation) to aid the manual process of identifying transpositions. Nevertheless, with the current system we can be sure that the identified transpositions were correct to a greater degree of certainty, as illustrated by the high precision values.

4.2 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. 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. Figures 9 and 10 show the comparison of USM with other state-of-the-art methods for strict and lenient evaluation respectively.

4.2.1 Italian Legislation Results. In the case of Italian Directive-NIM pairs, USM outperforms other methods in terms of F-score in all four directives in both strict and lenient evaluation. 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 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 average precision, recall and F-score values across all directives for different similarity measures. The results are shown in Figures 12 and 13. 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 average F-score for USM was 0.710 and 0.736 for strict and partial evaluation respectively. USM was also able to retrieve a greater number of transpositions than other methods as it achieved a higher recall.

4.2.2 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 in strict evaluation. 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 (Figures 12 and 13) 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).

4.2.3 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 in both strict and lenient evaluation. For CELEX 32003L0010, both USM and Euclidean similarity achieve the best F-score in lenient evaluation. However, in strict evaluation, Euclidean similarity achieves a better F-score than USM. 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 lenient evaluation. In terms of the average comparison of evaluation metrics (Figures 12, 13), USM achieved the highest F-score in both strict and lenient evaluation. In terms of recall it was minutely outperformed by LSA in strict evaluation. But in lenient evaluation USM achieved the best recall. USM also achieved the best performance in precision (tied with Euclidean in strict evaluation). We observed from the results that USM achieved the highest recall in all three cases of Luxembourg, Italian and English legislation (in lenient evaluation). In strict evaluation, LSA achieved minutely higher recall than USM for English legislation only. 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.

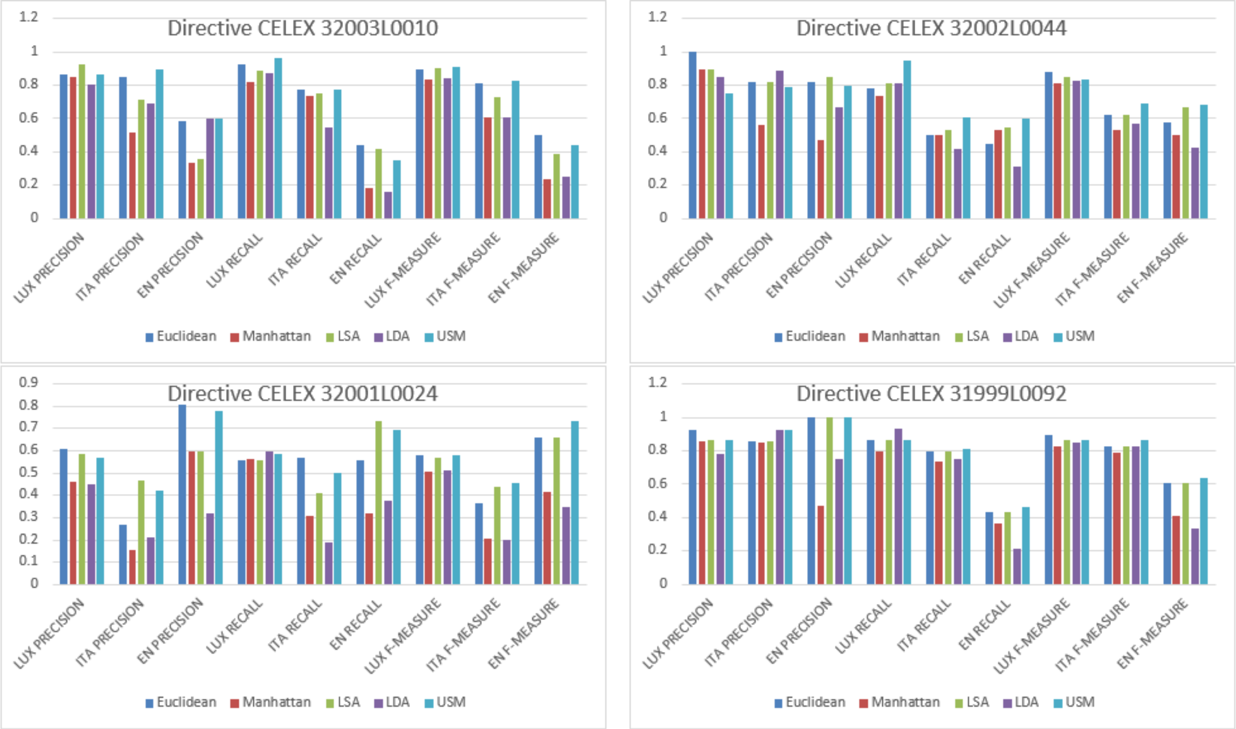


Figure 9: Comparison of the Unifying Similarity Measure (USM) with Euclidean, Manhattan, Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA) similarity measures for strict evaluation on the multilingual corpus

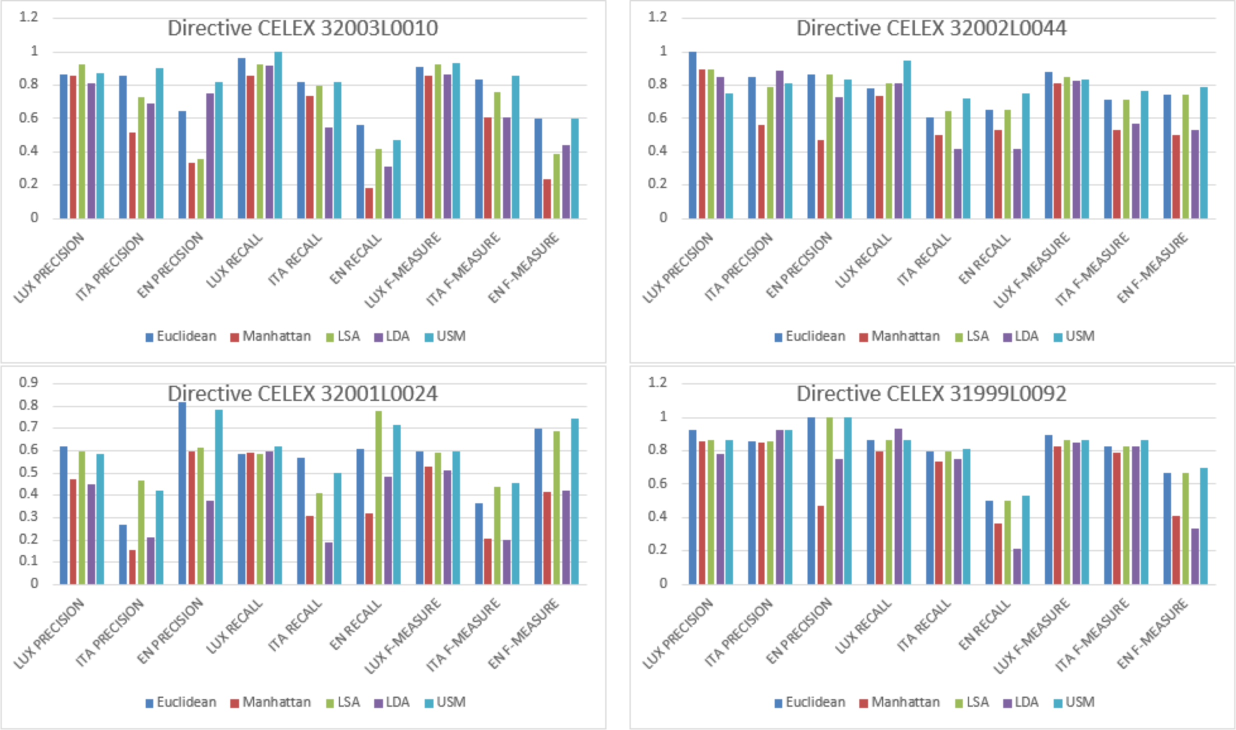
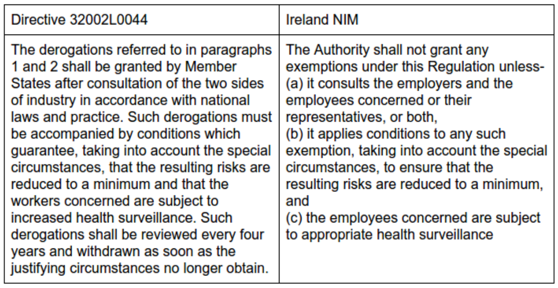


Figure 10: Comparison of the Unifying Similarity Measure (USM) with Euclidean, Manhattan, Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA) similarity measures for lenient evaluation on the multilingual corpus

We illustrate one example in Figure 11 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 Luxembourg and Italian legislation where LSA achieved lower recall than USM (Figure 12). The performance of LDA was poorer in terms of recall as compared to LSA, USM and Euclidean similarity. 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 12. In case of Luxembourg legislation, the recall of LDA was high enough as the wordings are more similar. 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.



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- (a) it consults the employers and the employees concerned or their representatives, or both, (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 (c) the employees concerned are subject to appropriate health surveillance

Figure 11: Article 10.3 from dir. CELEX 32002L0044 and corresponding NIM provision 10.3 of Ireland identified by USM

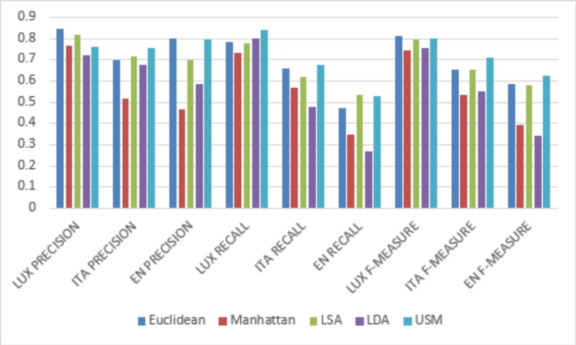


Figure 12: Strict evaluation comparison of USM with state-of-the-art similarity measures for average precision, recall and F-score across all directives in the multilingual corpus

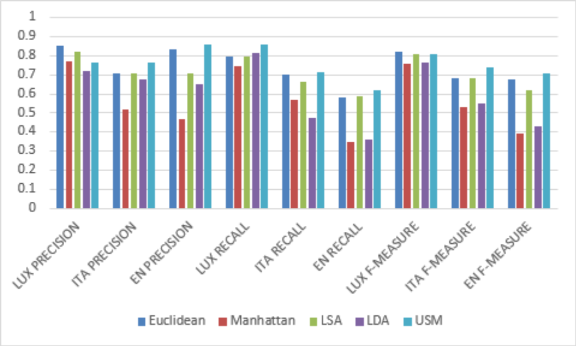


Figure 13: Lenient evaluation comparison of USM with state-of-the-art similarity measures for average precision, recall and F-score across all directives in the multilingual corpus

4.3 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[1]. The NIMs were taken from the legislation of Ireland. Table 2 shows the results of automated identification of NIM provisions for both strict and lenient evaluation. We observe that USM clearly outperforms other state-of-the-art text similarity measures in terms of F-score and recall. The average F-score, recall and precision values were 0.712, 0.693 and 0.738 across all four directives for lenient evaluation. Thus, USM model

[1] the corresponding list of NIMs from Ireland in order of the directives mentioned in Table 2 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: Comparison of USM with state-of-the-art text similarity methods on the extended English corpus

Strict Evaluation															
	Precision					Recall					F-Score				
Directives	Euclidean	Manhattan	LSA	LDA	USM	Euclidean	Manhattan	LSA	LDA	USM	Euclidean	Manhattan	LSA	LDA	USM
31998L0024	0.9523	1	0.9473	0.5925	0.913	0.606	0.147	0.5454	0.5925	0.6363	0.7407	0.2564	0.6923	0.5925	0.75
32000L0054	0.5937	0.7	0.6666	0.64	0.6764	0.6333	0.1707	0.5294	0.4848	0.7187	0.6129	0.2745	0.5901	0.5517	0.6969
32003L0122	0.7272	0.6428	0.8	0.75	0.6923	0.5714	0.6428	0.5333	0.4285	0.6923	0.64	0.6428	0.64	0.5454	0.6923
32006L0025	0.6666	0.5384	0.8	0.4166	0.6667	0.5	0.4117	0.4705	0.3846	0.6667	0.5714	0.4666	0.5925	0.4	0.6667
Lenient Evaluation															
	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

achieved encouraging results over the multilingual and English language corpus. We intend to carry out a more extensive multilingual testing of the USM approach in the future work.

5 CONCLUSION AND FUTURE WORK

This paper presented a unifying text similarity measure (USM) for automatically identifying the NIM provisions of EU directives in the national law. USM benefited from the complementarity of three similarity measures to identify transposed provisions. We used our model to identity transpositions in a multilingual corpus of four directives and their corresponding NIMs in three different languages. We further tested the USM approach on an extended English language corpus of four additional directives. The model was evaluated by comparing the results with correlation tables. Our results indicate that the model achieved a higher recall and F-score than other state-of-the-art methods for text similarity in both the corpora. The average F-score values for French, Italian and English Directive-NIM pairs were 0.808, 0.736 and 0.708 respectively in the multilingual corpus. This shows that our model is able to identify transpositions in different legal systems with good performance. Further evaluation on the English language corpus demonstrated that USM consistently achieved a higher F-score and recall than other text similarity methods. In future work, we intend to investigate the evaluation of our model on a larger corpus of directives and NIMs for different legislations. A promising idea to achieve higher recall without crucially lowering the precision could be achieved by integrating the existing statistical system with the rule-based system proposed in [15]. We aim in our long-term research at devising such an hybrid approach and integrating it in our systems for legal informatics [3] [4].

ACKNOWLEDGMENTS

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. This work has been partially supported by 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". The authors would like to thank Sabrina Praduroux for her work on the correlation tables.

REFERENCES

[1] R. Angell, G. Freund, and P. Willett. 1983. Automatic spelling correction using a trigram similarity measure. *Information Processing & Management* 19, 4 (1983).

[2] I. Atoum and A. Otoom. 2016. Efficient Hybrid Semantic Text Similarity using Wordnet and a Corpus. *International Journal of Advanced Computer Science & Applications* 1, 7 (2016).

[3] G. Boella, L. Di Caro, M. Graziadei, L. Cupi, C. Salaroglio, L. Humphreys, H. Konstantinov, K. Marko, L. Robaldo, C. Ruffini, K. Simov, A. Violato, and V. Stroetmann. 2015. Linking Legal Open Data: Breaking the Accessibility and Language Barrier in European Legislation and Case Law. In *Proceedings of the 15th International Conference on Artificial Intelligence and Law*. ACM, 5.

[4] G. Boella, L. Di Caro, L. Humphreys, L. Robaldo, R. Rossi, and L. van der Torre. 2016. Eunomos, a legal document and knowledge management system for the web to provide relevant, reliable and up-to-date information on the Law. *Artificial Intelligence and Law* 24 (2016). Issue 3.

[5] D. Buscaldi, R. Tournier, N. Aussenac-Gilles, and J. Mothe. 2012. Irit: Textual similarity combining conceptual similarity with an n-gram comparison method. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics*. Association for Computational Linguistics.

[6] L. Candillier, F. Meyer, and F. Fessant. 2008. Designing specific weighted similarity measures to improve collaborative filtering systems. In *Industrial Conference on Data Mining*. Springer.

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

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

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

[10] L. Humphreys, C. Santos, L. Di Caro, G. Boella, L. van der Torre, and L. Robaldo. 2015. Mapping Recitals to Normative Provisions in EU Legislation to Assist Legal Interpretation.. In *Proc. of the 28th International Conference on Legal Knowledge and Information Systems.*

[11] A. Islam and D. Inkpen. 2008. Semantic text similarity using corpus-based word similarity and string similarity. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 2, 2 (2008).

[12] R. Mihalcea, C. Corley, C. Strapparava, and others. 2006. Corpus-based and knowledge-based measures of text semantic similarity. In *AAAI*, Vol. 6.

[13] M. Mohler and R. Mihalcea. 2009. Text-to-text semantic similarity for automatic short answer grading. In *Proc. of the 12th Conference of the European Chapter of the Association for Computational Linguistics.*

[14] R. Nanda, L. Di Caro, and G. Boella. 2016. A Text Similarity Approach for Automated Transposition Detection of European Union Directives. In *Proc. of the 29th International Conference on Legal Knowledge and Information Systems.*

[15] L. Robaldo, T. Caselli, I. Russo, and M. Grella. 2011. From Italian Text to TimeML Document via Dependency Parsing. In *Proc. of the 12th International Conference Computational Linguistics and Intelligent Text Processing.*

[16] K. Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation* 28, 1 (1972).

[17] R. Vasile, B. Nobal, and B. Rajendra. 2013. Similarity measures based on latent dirichlet allocation. In *International Conference on Intelligent Text Processing and Computational Linguistics.* Springer.

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