Cross-lingual Annotation Projection in Legal Texts

Andrea Galassi[1] and **Kasper Drazewski**[2] and **Marco Lippi**[3,4] and **Paolo Torroni**[1,4]
[1]Department of Computer Science and Engineering, University of Bologna, Italy
[2]Bureau Européen des Unions de Consommateurs, Brussels, Belgium
[3]DISMI, University of Modena and Reggio Emilia, Italy
[4]European University Institute, Florence, Italy
[1]a.galassi@unibo.it, paolo.torroni@unibo.it
[2]kasper.drazewski@eui.eu
[3]marco.lippi@unimore.it

Abstract

We study annotation projection in text classification problems where source documents are published in multiple languages and may not be an exact translation of one another. In particular, we focus on the detection of unfair clauses in privacy policies and terms of service. We present the first English-German parallel asymmetric corpus for the task at hand. We study and compare several language-agnostic sentence-level projection methods. Our results indicate that a combination of word embeddings and dynamic time warping performs best.

1 Introduction

The European Union considers cultural and linguistic diversity, and in particular multilinguality, as some of its fundamental principles. So much so, that all the regulations and laws given by the Parliament are published in all of EU's twenty-four official languages. These could be an immense resource towards transparency, egalitarianism, accountability and democracy, giving the EU citizen access to legislative and policy proposals in their own and also in other languages (Steinberger et al., 2014). However, to transform such resources into assets for the citizen, linguistic tools are needed, that can automatically analyze textual sources and yield actionable information (Lippi et al., 2019a). Unfortunately, producing the annotated corpora required to train such linguistic tools is, even for a single language, notoriously expensive. Moreover, the vast majority of linguistic resources and tools focus on English. Likewise, the workforce of professionals needed for annotating legal documents may not be readily available in each language. We thus investigate methods for automatically transferring the annotations made on legal documents in a language with significant linguistic resources and domain experts, such as English, onto the corresponding versions of the same documents in a target language, where such resources and expertise may be lacking. Our ultimate goal is to use automatically generated annotations for training linguistic tools for the target language without resorting to expert annotators in that language.This would leverage the creation of classifiers that can leverage the linguistic resources available in the target language, to analyze documents in that language.

We chose to focus on the detection of unfair clauses in online Terms of Service (ToS) and Privacy Policies (PP), which are usually published in multiple languages. However, the domain of interest of our study spans across many other types of legal texts. Other EU official documents published in multiple languages include, for instance, EU Parliament laws and regulations, documents of the EU Court of Justice, policy documents, documents for public consultations, and so on. Therefore, the potential import of effective methods for facilitating cross-lingual legal text analysis is considerable. Reasons for focusing on ToS and PP are the interest for such documents from a consumer protection perspective, especially since the recent adoption of the European General Data Protection Regulation (GDPR), as well as the availability of tools for the analysis of such documents. One such tool is CLAUDETTE (Lippi et al., 2019b), a web server for the automatic detection of potentially unfair clauses in ToS. At the time of writing, CLAUDETTE is only available for the English language. An effective method for cross-lingual

Proceedings of the 28th International Conference on Computational Linguistics, pages 915–926
Barcelona, Spain (Online), December 8-13, 2020

annotation projection could extend its scope to a variety of languages other than English, without having to resort to domain expertise in any of these languages.

To carry out the present study, we built the first English-German parallel corpus for the task at hand. It turned out that the corpus is asymmetric, meaning a one-to-one sentence-by-sentence correspondence between source and target document is not always guaranteed. This aspect makes the problem particularly challenging, since we need to ensure a correct behaviour when a sentence does not appear in one of the documents or it has been integrated in another sentence: a property of sentence alignment called *robustness* (Simard and Plamondon, 1998). For example, we found source-document sentences corresponding to multiple target-document sentences, as well as significant rephrasing, resulting in the same sentence being annotated differently in the two documents.

The solution we propose makes use of an automated translation of the source document into the target document, obtained via a third-party tool, of a sentence-wise dissimilarity metric, and of a method for finding an alignment between warped time series. We evaluate performance with several combinations of dissimilarity scores, with or without dynamic time warping. Our results indicate that the best performance is achieved by a combination of word embedding-based dissimilarity and dynamic time warping.

The main novelty of our study lies in the confluence of three aspects: the projection is performed at *sentence-level*; the multi-lingual corpus is *asymmetric*; and the proposed methods are *language-agnostic*.

Sections 2 and 3 formalize the problem and describe the architecture of our solution; 4 and 5 present corpus and experiments; 6 discusses related work in the areas of sentence alignment and annotation projection; 7 concludes.

2 Problem Definition

Our problem is transferring the knowledge provided by some experts and encoded in the form of document annotations in a source language, into a target language. Annotations consist of tags attached to sentences in a collection of documents. Our approach is based on the projection of annotations (Yarowsky et al., 2001) between the two languages. Henceforth we will use English (E) and German (G) as source and target, respectively.

The input is defined by three resources:

D_E) the annotated English version of the *source* document;

D_G) the corresponding German version of the same document, which is the *target* to be annotated;

D_G^t) the automated translation of D_G into English.[1]

The goal is to find a correspondence between the sentences $e_1, \ldots, e_n$ in D_E and the sentences $g_1, \ldots, g_m$ in D_G via the automatically translated sentences $g_1^t, \ldots, g_n^t$. In this way, the original annotations (i.e., the labels) $\ell_1, \ldots, \ell_n$ associated with the sentences in D_E can be transferred (projected) from the English document into a sequence of corresponding labels $\ell'_1, \ldots, \ell'_m$ in the target. All the correspondences are thus evaluated among pairs of English sentences.

We shall remark that our setting is *asymmetrical*, therefore owing to the different way the content of documents may be rendered in different languages, n may differ from m.

Although our experimental results are limited to the English/German language pair, the approach itself is *language-agnostic*. In particular, it does not rely on any language-specific features. For the same reason, we consider automatic translation as a service, and not as a variable in the experimental study.

3 Projection Architecture

Our methodology for annotation projection is two-fold. The first step is the computation of a set of matches between each sentence of the translated target document, D_G^t, and one or more sentences of the source document D_E. Each sentence of D_G^t is thus labeled using the labels of the corresponding sentences of D_E, if any. The second step amounts to transferring the labels from D_G^t to the target

[1] The choice to translate into English is arbitrary. One could equally choose to rely on translating D_E into German.

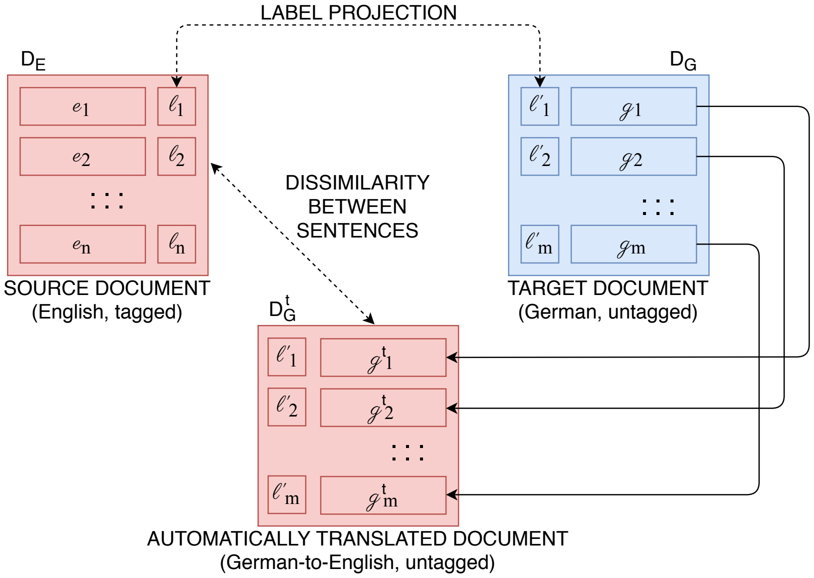

Figure 1: Projection between a source document D_E in English and a target D_G in German. Automated translation is used to obtain a translated version of D_G called D_G^t, so that labels attached to sentences in D_E can be mapped onto those in D_G.

document D_G, which is straightforward, since by construction there is a perfect match between D_G and D_G^t. For the computation of matches, we considered four different procedures, all amenable to performing projection into German or into English.

The simplest procedure, which we take as a baseline, P_t, is illustrated in Figure 1. It consists in computing the dissimilarity between any two sentences from the two documents, by simply considering them as strings of symbols, and then matching each non-annotated sentence of one document with the least dissimilar sentence in the other one. A similar procedure, P_e, requires creating a neural embedding for each sentence and then computing the dissimilarities between embeddings. These procedures create a relation between source and target sentences: each target sentence is associated with one source sentence, whereas each source sentence can be associated with any number of target sentences (possibly none).

In the case of online ToS and PP written in different languages, it is not surprising to observe a high degree of parallelism between the source and the target documents. Therefore, it is fair to assume that in most cases the sentences will appear in the source document in roughly the same order as their matching counterparts appear in in the target document.[2] To exploit this property, we design two additional procedures relying on Dynamic Time Warping (Kruskall, 1983; Salvador and Chan, 2007) to compute the match: P_{t+DTW} for textual input data, and P_{e+DTW} for embedded input sentences.

3.1 Dissimilarity between sentences

Sentence dissimilarity can be measured in many ways (Gomaa and Fahmy, 2013). A first family of methods consists in variants of the the so-called classic *edit-distance*, whereby dissimilarity is computed as the number of operations (insertion, deletion, substitution of symbols) needed to transform one string into another. For this work, we considered the Hamming (1980), Levenshtein (1966), Damerau-Levenshtein (Damerau, 1964), and Needleman-Wunsch (Needleman and Wunsch, 1970) dissimilarities. With the exception of the Hamming distance, all these dissimilarities yield a heavy computational footprint. Moreover, they must be computed for each sentence pair $\{a, b\}$, independently of what may be known of other pairs. In other words, there is no easy reuse of partial results, which adds to the computational footprint. This aspect considerably limits the application of these metrics in our setting, where each sentence needs comparing with potentially many other sentences. Other methods are based on *token-based* scores. Of those, we consider the Jaccard (1912) distance, based on the number of words in

[2] A detailed analysis on a sample of our data (800 sentences) showed that sentence inversion is indeed infrequent in our corpus. In fact, it didn't occur even once in the data we analyzed.

common. Finally, there are methods based on *compression*. Among them, we selected the Normalized Compression Dissimilarity (NCD), which amounts to compressing the two sentences, first individually and then together, and then comparing the size of the compressed data.[3] Jaccard and NCD both require computing intermediate results (values or representations) which are independent of one of the sentences. Storing such results yields a significant reduction in the computational footprint. All of the above methods can be applied to sentences encoded as strings of symbols, as in P_t and P_{t+DTW}. In the sequel we will describe dissimilarity measures used with sentence embeddings, as in P_e and P_{e+DTW}.

3.2 ELMo embeddings

Word embeddings are a way to associate a numerical vector to each word in a corpus, typically computed through sub-symbolic techniques. Usually, these embeddings are learned through a computationally demanding training process based on a very large corpus. Such learned representations embed many different aspects of the entity they represent, that can be used as features for other tasks. Aditionally, pre-trained embeddings yield a lightweight computational footprint, which makes them particularly suitable when the available computational resources are limited. Moreover, pre-training is especially effective when working with small corpora, which cannot by themselves encode a rich language model.

ELMo (Peters et al., 2018) embeddings are produced by exploiting a bi-directional language model, which consists in a neural network trained to predict a word in a sentence by looking at the surrounding words in the sentence. Thus, the representation provided by an ELMo word embedding is able to capture the context of each word. Given all the embeddings of the words in a sentence, a compact representation of that sentence can be obtained, for instance, as their average.

To measure the dissimilarity between embeddings we considered the cosine dissimilarity (Kenter and de Rijke, 2015) and the Bray-Curtis dissimilarity (Bray and Curtis, 1957). The cosine dissimilarity between two numerical vectors is obtained by subtracting the cosine of the angle between the two vectors from 1. The Bray-Curtis dissimilarity between two numerical vectors a and b is a normalized version of the Manhattan distance, as it is computed as the sum over the absolute differences between elements a_j and b_j, divided by the sum over the elements computed for each vector, separately.

3.3 Dynamic Time Warping

Dynamic Time Warping (DTW) and its more efficient approximation Fast DTW (FTDW) enable finding an alignment and evaluating a numerical dissimilarity between time series that may have been *warped* (i.e., stretched or shrunk) along the time axis (Salvador and Chan, 2007). DTW measures the dissimilarity between pairs of elements of the two series to create a matrix. Each element of the matrix represents a matching between these elements, and its value represents their dissimilarity, or *cost*, of the matching. The algorithm computes the cheapest path from one end to the other of the cost matrix. The alignment between the two series is given by the cells in the path, while the dissimilarity measure is the cost of the path. DTW guarantees to find an optimal alignment with quadratic complexity with respect to the length of the time series.[4] FDTW is a popular, linear-complexity approximation of DTW, which does not guarantee optimality, but increases the chance of finding a good match, by iteratively computing constrained approximate paths. Instead of computing the full cost matrix from the beginning, FDTW first computes coarse versions of the time series by aggregating adjacent elements. This procedure is iterated to obtain many versions of the same time series at different resolutions. DTW is then applied to the smallest-resolution time series, to compute a coarse cost map and a path across it. Then, a constrained DTW is applied to the pair of time series with a larger resolution, using the path already computed to guide the creation of the cost map: only the costs of the cells which correspond to the neighbourhood of the previous path are computed, by imposing constraints on the path for this level. Finally, a constrained DTW is applied to the full-resolution time series, providing the final path and the dissimilarity score.

Both DTW and FDTW can thus be used to determine a many-to-many match between the elements of the two sequences, guaranteeing that each element of either sequence is matched with at least one

[3] The entropy of the sequence was used to measure compression.

[4] Complexity can be reduced by giving up optimality and imposing constraints on the alignments, for example by forbidding the elements of a sequence from having all the same match with a single element. We do not apply such approximations.

ToS	ENG	GER
Sentences	1,323	1,481
Tags	153	189
arbitration $\langle a \rangle$	6	10
unilateral change $\langle ch \rangle$	34	37
content removal $\langle cr \rangle$	18	23
jurisdiction $\langle j \rangle$	14	15
choice of law $\langle law \rangle$	15	15
limitation of liability $\langle ltd \rangle$	26	42
inclusion of consent to PP $\langle pinc \rangle$	2	2
unilateral termination $\langle ter \rangle$	28	32
contract by using $\langle use \rangle$	10	13

PP	ENG	GER
Sentences	768	855
Tags	266	351
data used for advertising $\langle ad \rangle$	25	30
data collected not form data subject $\langle basis \rangle$	45	55
data transferred to authorities $\langle cat \rangle$	97	127
data transferred to other users $\langle source \rangle$	29	37
data transferred to processors $\langle ta \rangle$	5	7
data transferred to controllers $\langle tc \rangle$	36	50
categories of data being collected $\langle tpr \rangle$	12	21
basis for processing $\langle tu \rangle$	17	24

Table 1: Composition of ToS and PP subsets.

element of the other sequence. It is worthwhile noticing that, in aligning the two series, these algorithms maintain the order of the series' elements. Given four elements a_1, a_2, b_1, b_2 belonging to time series A and B, such that $a_1 < a_2$ in A and $b_1 < b_2$ in B, if DTW determines (a_1, b_2) to be a match, then it can also determine (a_1, b_1) or (a_2, b_2) to be a match, but not (a_2, b_1) because that would be incompatible with (a_1, b_2). As a final remark, DTW only needs a measure of pairwise element dissimilarity, whereas FDTW also needs an aggregation function between adjacent elements, such as an average. Since such function is difficult to define for textual data, in P_{t+DTW} we will use the original DTW algorithm. Conversely, in P_{e+DTW} we will make use of FDTW, aggregating adjacent embeddings by computing their average. In the rest of this work we will not distinguish between the two algorithms, and we will address both as DTW.

4 Dataset

The English-German corpus created for this task includes ToS and PP: two types of documents employing very different languages and drafting techniques. Table 1 shows the distribution of unfair clauses in the corpus. The documents were sourced from the CLAUDETTE training corpus,[5] and the German versions were annotated by a legal expert fluent in English and German. The dataset described here, as well as the code used in the experiments, are publicly available.[6]

4.1 Subsets

The terms of service (ToS) set consists of 5 contracts used by online service providers: Box.com, Garmin, Grindr, Linkedin and MyHeritage. It includes 2,808 sentences and 342 tags identifying 27 classes, divided into 9 categories, as described by Lippi et al. (2019b). The data privacy set (PP) comprises privacy policies from Dropbox, Facebook, Supercell, Tumblr and Twitter. It includes 1,623 sentences and 617 tags identifying 21 classes, divided into 8 categories, as described by Contissa et al. (2018).

Annotations also indicate the degree of unfairness. For example, `ltd3` means high degree of unfairness on grounds of limitation of liability (e.g., in the case of damages caused by gross negligence), whereas `ltd1` indicates a fair clause (e.g. it does not exclude the provider's liability).

4.2 Discrepancies

We found that discrepancies between the English and the German documents usually pertain to the sentence structure and, aside from convenience of reading, appear to have no justification. Example 4.1 shows how the same information, contained in one clause in the English version (potentially unfair content removal and unilateral termination of contract), requires a manual split of the categories when transferred to the official German counterpart.

Example 4.1 (Box.com, line 40)**.**
$\langle cr2 \rangle \langle ter2 \rangle$ *We reserve the right to delete or disable Content alleged to violate copyright laws or these Terms and reserve the right to terminate the account(s) of violators.* $\langle /cr2 \rangle \langle /ter2 \rangle$

[5] `http://claudette.eui.eu`

[6] `https://bitbucket.org/a-galaxy/cross-lingual-annotation-projection-in-legal-texts`

$\langle cr2 \rangle$ *Wir behalten uns das Recht vor, Inhalte zu löschen oder zu deaktivieren, die vorgeblich gegen Urheberrechtsgesetze oder diese Bedingungen verstoßen.* $\langle /cr2 \rangle$ $\langle ter2 \rangle$ *Zusätzlich behalten wir uns das Recht vor, den/die Account(s) des Beschuldigten zu sperren.* $\langle ter2 \rangle$

Example 4.2 instead shows one sentence in the English version, pertaining to the use of third-party subcontractors as data processors, corresponding to two sentences in German. Since the message contained in both still fulfils the criteria for the same tag, the annotation used needs to be manually copied to the neighboring sentence.

Example 4.2 (PP Tumblr, line 112)**.**
$\langle tpr2 \rangle$ *Information Shared with Our Agents in Order to Operate and Improve the Services: In some cases, we share information that we store (such as IP Addresses) with third parties, such as service providers, consultants, and other agents ("Agents"), for the purposes of operating, enhancing, and improving the Services, and developing new products and services.* $\langle tpr2 \rangle$
$\langle tpr2 \rangle$ *Daten, die mit unseren Beauftragten ausgetauscht werden, um die Dienste zu betreiben und zu verbessern: In manchen Fällen geben wir von uns gespeicherte Daten (beispielsweise IP-Adressen) an Dritte wie Dienstanbieter, Berater und andere Beauftragte, ("Beauftragte"), weiter.* $\langle /tpr2 \rangle$ $\langle tpr2 \rangle$ *Dies geschieht zum Zwecke des Betreibens, der Erweiterung und Verbesserung der Dienste sowie zur Entwicklung neuer Produkte und Dienste.* $\langle tpr2 \rangle$

The presence of such discrepancies makes projection a particularly challenging task, and affects performance evaluation, as explained in Section 5.

5 Experimental Results

We performed projection as a fully unsupervised process. To translate the documents from German to English we used `Apache Joshua`, a standalone tool that poses no usage restrictions. To compute the dissimilarity measures described in Section 3.1 between the D_E and D_G^t we used the Python `textdistance` library. To create the embeddings of the source sentences and of the translated target sentences, we use the pre-trained ELMo model available on TensorFlowHub, which follows Peters et al. (2018), with an embedding size of 1024, and create sentence embeddings by average of word embeddings. The corpus used for pre-training is 1B Word benchmark (Chelba et al., 2014), which includes a variety of sources in English such as news commentaries and parliamentary debates. To compute the distance between embedding vectors we used the `SciPy` library (Virtanen et al., 2020).

After obtaining the projected labels for D_G, we evaluated the performance by comparing such projected labels with the original golden annotations. In order to measure the performance, we considered the task as a multi-label classification problem on the sentences of the target documents. Tables 2 and 3 report F1 scores, precision and recall associated with each approach. These scores are computed with no distinction between different documents. The F1-micro score is computed considering the total number of true positives, false positives, and false negatives, without making any distinction between classes. The weighted score is computed as a weighted average between the F1 scores obtained for each class, where the weight is given by the number of times the label appears in the ground truth. The macro score is obtained with a similar procedure, without using different weights, but considering all the classes alike.

We shall remark that, in principle, the task does not guarantee that an F1 equal to 1 is always achievable, due to possible discrepancies between the two sets of labels in the original English and German corpora. With reference to Example 4.1, even if the English sentence had been correctly associated with both German sentences, each of the latter would inherit both tags of the source language, thus resulting in a false positive for what concerns the evaluation metrics.

Table 2 compares the results obtained by the baselines and the use of ELMo embeddings without DTW. The strongest baseline is Needleman-Wunsch, which is largely outperformed by the use of the embeddings. Indeed, embeddings improve the F1 scores by about 0.20, especially due to increased precision. Moreover, the Bray-Curtis dissimilarity performs slightly better than the cosine one.

In Table 3 we report results with DTW, which improves the performance for both the text-based and the embedding-based approaches. The Needleman-Wunsch dissimilarity receives a minor boost of about

Table 2: Evaluation of projection on the whole dataset, using different dissimilarity functions, without the use of DTW.

	P_t						P_e	
	Hamming	Levenshtein	Damerau-Levenshtein	Needleman-Wunsch	Jaccard	NCD	Cosine	Bray-Curtis
F1-macro	0.21	0.54	0.50	0.59	0.41	0.18	0.76	**0.77**
F1-micro	0.25	0.58	0.59	0.62	0.47	0.22	0.80	**0.81**
F1-weighted	0.25	0.58	0.58	0.62	0.48	0.22	**0.80**	**0.80**
Precision	0.26	0.62	0.63	0.63	0.48	0.20	**0.86**	**0.86**
Recall	0.24	0.55	0.55	0.61	0.45	0.23	**0.75**	**0.75**

Table 3: Evaluation of projection on the whole dataset, using different dissimilarity functions, with the use of DTW.

	P_{t+DTW}						P_{e+DTW}	
	Hamming	Levenshtein	Damerau-Levenshtein	Needleman-Wunsch	Jaccard	NCD	Cosine	Bray-Curtis
F1-macro	0.78	0.79	0.79	0.70	0.78	0.74	0.81	**0.82**
F1-micro	0.82	0.83	0.83	0.72	0.83	0.79	**0.86**	**0.86**
F1-weighted	0.82	0.83	0.83	0.72	0.83	0.79	**0.86**	**0.86**
Precision	0.90	0.91	0.91	0.68	0.89	0.84	0.92	**0.93**
Recall	0.76	0.77	0.77	0.76	0.77	0.75	**0.80**	**0.80**

0.10 points. On the contrary, Hamming and NCD dissimilarities improve dramatically, outperforming Needleman-Wunsch and reaching results similar to Levenshtein and Damerau-Levenshtein. The embedding-based approach (P_e) still yields the better results with respect to P_t, although with a smaller margin. In spite of the small margin, embedding-based approaches offer a dramatic gain in terms of computational footprint. Processing our whole corpus by the embedding-based method required less than a minute, while the P_t methods required several hours.

We have also observed that all the methods perform better on PPs than on ToSs, especially with respect to precision. We speculate that the reason for that may the different distribution of the labels between the English and German documents, which is more pronounced in ToS than in PP. Table 4 reports a comparison between the best methods for each approach on ToS and PP.

Finally, Figure 2 summarizes the F1-weighted scores obtained for each different category in the setting without DTW, in ToS and PP, respectively. The plot shows that the embedding-based approaches consistently outperform the other dissimilarity scores. It is noteworthy that the categories of ToS in which we perform best and worst are the ones with less data points. A meaningful qualitative analysis should not focus on those two classes but rather sample on the whole dataset. We will further investigate these differences in our future work.

One issue worth consideration is the impact of the translation service on the overall performance. We have conducted preliminary experimentation using professional tools such as Google Translate. We did notice a performance improvement overall. However, the relations between the performance of P_t, P_e, P_{t+DTW} and P_{e+DTW} are unchanged.

Table 4: Comparison of the various approaches on the two corpora.

	P_t		P_e		P_{t+DTW}		P_{e+DTW}	
	ToS	PP	ToS	PP	ToS	PP	ToS	PP
F1-macro	0.52	0.66	0.70	0.84	0.73	0.85	0.75	0.88
F1-micro	0.54	0.67	0.76	0.83	0.81	0.84	0.82	0.88
F1-weighted	0.54	0.66	0.75	0.83	0.80	0.84	0.82	0.88
Precision	0.50	0.73	0.81	0.89	0.87	0.93	0.87	0.96
Recall	0.58	0.62	0.71	0.77	0.75	0.77	0.78	0.81

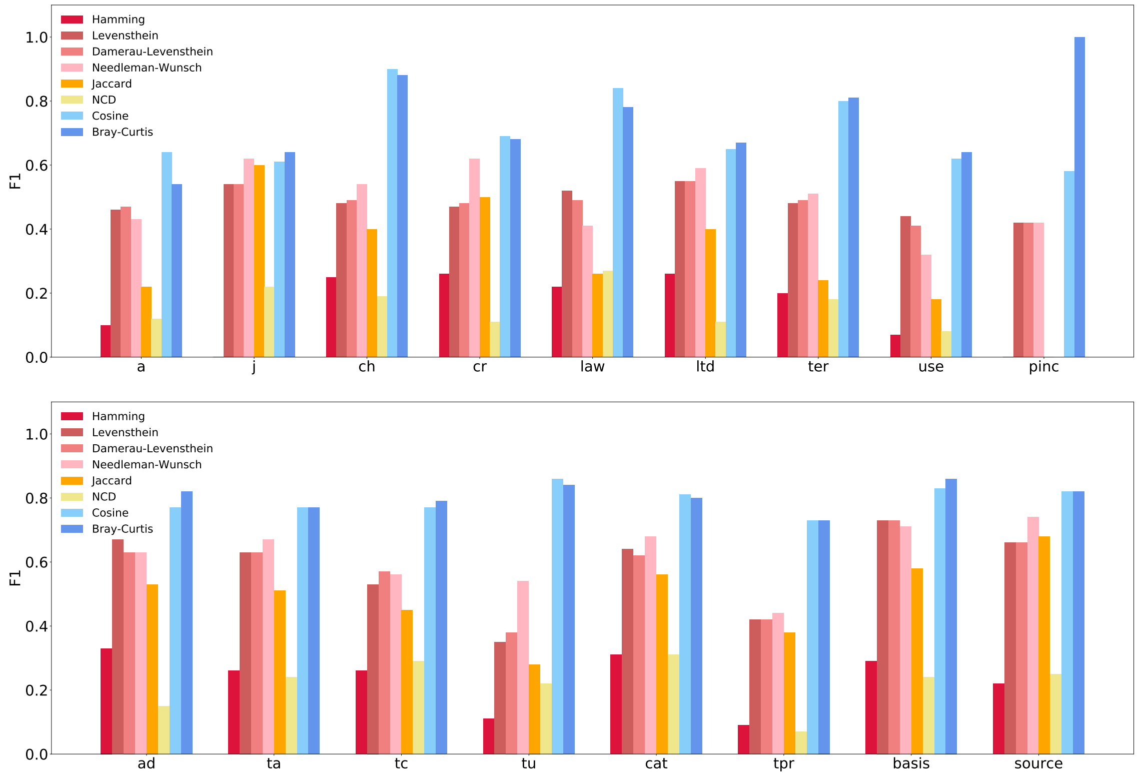

Figure 2: F1-weighted scores per each different category, achieved for ToS (top) and PP (bottom) without the use of DTW.

6 Related Works

Our work is closely related to two research strands, namely sentence alignment and annotation projection.

The sentence alignment task is addressed in the work of Simard and Plamondon (1998) by defining a “corridor of alignment” based on global information. Basically, a candidate matching between sentences takes into account the position of the sentences inside each document. Then, corresponding words are matched, and this local information is further exploited to align full sentences following an ordered one-to-one relationship. This is somehow similar to our approach based on DTW, although we directly operate at the sentence level. To address the task of aligning documents in which sentences do not appear in the same order, Zamani et al. (2016) present an approach based on Integer Linear Programming, while Quan et al. (2018) propose instead an approach based on the length of the sentences to create one-to-one matches. Other examples of alignement between parallel corpora can be found in (Santos, 2011). With respect to these works, our task is different, since we do not aim for the perfect alignment of words, or paragraphs, but we are interested only in a faithful projection of labels.

The idea of projecting annotations between two parallel corpora in different languages, to enable the construction of a machine learning model in both languages, is firstly framed in (Yarowsky et al., 2001). The authors there address various word-level NLP tasks. Interestingly, they address the problem of noisy projection by using the uncertainty level given by the alignment model. The same problem is addressed by Fossum and Abney (2005) by exploiting multiple sources for the same target document. Projection has been used also for argumentation mining (Eger et al., 2018; Rocha et al., 2018), to create training data for machine learning models for low-resource languages. In particular, Eger et al. (2018) argue against the necessity of human-translated parallel corpora as a resource, since they obtain comparable results using machine translated parallel documents. In contrast with these approaches, which use projected labels for supervised training, in the work by Das and Petrov (2011) the projected labels are used as features for unsupervised training. Projection has also been used by Bentivogli and Pianta (2005) to create a parallel version of an existing corpus. The projection of structural information between parallel documents is tackled by Bamman et al. (2010), where alignment is performed firstly sentence-wise (1-1) and then word-wise.

None of the reviewed approaches is directly applicable to our setting. The works of Das and Petrov (2011), Moore (2002), Quan et al. (2018), and Zamani et al. (2016) are all based on symmetric corpora, while imposing a 1-1 alignment is not possible in our domain. Bamman et al. (2010), Bentivogli and Pianta (2005), Eger et al. (2018), Fossum and Abney (2005), and Yarowsky et al. (2001) all address word-level alignment instead of sentence level-alignment. These methods cannot be extended to address sentence-level alignment without introducing an element of subjectivity, for instance, in resolving conflicts of words matched across different sentences. Finally, Simard and Plamondon (1998) rely on the language-dependent concept of "cognate" words, which is applicable only to selected language pairs, while our approach is language-agnostic.

Besides projection, other approaches can be used to apply machine-learning methods for low-resources languages. Direct transfer (Zhang et al., 2016) consists of exploiting features which are shared across the languages, such as multilingual word embeddings, and then training a model on the labelled corpus, so as to use it directly on the test data, without any need of parallel corpora or projection. While this approach may seem less noisy than projection, in Eger et al. (2018) direct transfer approaches have achieved worse performance. Other works (Cotterell and Heigold, 2017; Kim et al., 2017) have made use of weak supervision, a mixture of supervised and unsupervised learning, which requires a setting where a few labelled documents of the target language are available. Finally, another approach consists in the cross-lingual transfer/alignment of word embedding spaces (Xu et al., 2018; Lample et al., 2018), where a mapping between the word embeddings of the source and the target language is first learned; then, word embeddings in the target language are mapped into the ones in the source language word, so that they can be finally fed into any model trained on the source language documents.

For what concerns other methods to asses the distance between two sentences, Gomaa and Fahmy (2013) present an extensive survey from which we have selected the most common and significant algorithms. Mueller and Thyagarajan (2016) and He et al. (2015) train neural model to assess the distance between two sentence, but this approach is applicable only to languages where a dataset of sentence distances is available. Finally, Lopez-Gazpio et al. (2019) and Chen et al. (2018) involve the training of advanced neural models and require a large database for training, thus do not suit our setting, since ours is not a learning-based approach.

7 Conclusion

We devised and compared several methods for annotation projection. The motivation behind this study stemmed from our long-term goal of extending the scope of existing classification systems of legal texts, currently available only in English, so as to have them work with a variety of languages, without having to resort to legal experts in such languages. We studied the performance of a number of alternative methods for text encoding, measuring sentence dissimilarity, and integrating sequence information in the alignment problem.

We tested several possibilities with an English-German corpus created by legal experts in the two languages, and interpreted the results as an indication that a combination of word embeddings and dynamic time warping seems most effective. This is fortunate since the methods involved in such a combination also yield a much smaller computational footprint compared to other methods we have investigated.

These results pave the way to several directions for future research. One of them is architectural. As we remarked when we defined the problem, our approach could be implemented by relying on automated translations of source documents to the target language or vice versa. In our study, we arbitrarily opted for the second alternative. We thus plan to implement the option we have not yet considered, and measure its performance in comparison to the results we obtained so far. Moreover, we plan to make an experimental comparison between the results obtained with a tool for legal text analysis trained to work on a target language using the approach presented here, and the results obtained by translating the query document and performing the analysis with the original tool. We speculate that answers to the above questions may depend partly on the task, partly on the availability and quality of language resources and tools in different languages. Finally, another direction worth of investigation is the use of state-of-the-art sentence embeddings, such as Sentence-BERT (Reimers and Gurevych, 2019) and Universal Sentence Encodings (Cer et al., 2018).

In conclusion, we shall remark that although our analysis has focussed on a specific set of legal documents, cross-lingual annotation projection in legal texts has, in fact, a much wider import. In the longer run we would like to apply this method to other areas of cross-lingual legal text analysis, with the ultimate purpose of contributing positively to the relations between citizens and institutions across nations characterized by linguistic and cultural diversity and united under a common institutional framework.

References

David Bamman, Alison Babeu, and Gregory Crane. 2010. Transferring structural markup across translations using multilingual alignment and projection. In *JCDL 2010*, pages 11–20, New York, NY, USA. ACM.

L. Bentivogli and E. Pianta. 2005. Exploiting parallel texts in the creation of multilingual semantically annotated resources: the multisemcor corpus. *Natural Language Engineering*, 11(3):247–261.

J. Roger Bray and J. T. Curtis. 1957. An ordination of the upland forest communities of southern wisconsin. *Ecological Monographs*, 27(4):325–349.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174, Brussels, Belgium, November. Association for Computational Linguistics.

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. 2014. One billion word benchmark for measuring progress in statistical language modeling. In *INTERSPEECH*, pages 2635–2639. ISCA.

Qin Chen, Qinmin Hu, Jimmy Xiangji Huang, and Liang He. 2018. CA-RNN: using context-aligned recurrent neural networks for modeling sentence similarity. In *AAAI*, pages 265–273. AAAI Press.

Giuseppe Contissa, Koen Docter, Francesca Lagioia, Marco Lippi, Hans-W Micklitz, Przemysław Pałka, Giovanni Sartor, and Paolo Torroni. 2018. CLAUDETTE meets GDPR. Automating the evaluation of privacy policies using artificial intelligence. Study Report, Funded by The European Consumer Organisation.

Ryan Cotterell and Georg Heigold. 2017. Cross-lingual character-level neural morphological tagging. In *EMNLP*, pages 748–759, Copenhagen, Denmark, September.

Fred J. Damerau. 1964. A technique for computer detection and correction of spelling errors. *Commun. ACM*, 7(3):171–176, March.

Dipanjan Das and Slav Petrov. 2011. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *ACL*, pages 600–609, Portland, Oregon, USA, June.

Steffen Eger, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2018. Cross-lingual argumentation mining: Machine translation (and a bit of projection) is all you need! In *COLING*, pages 831–844, Santa Fe, New Mexico, USA, August.

Victoria Fossum and Steven Abney. 2005. Automatically inducing a part-of-speech tagger by projecting from multiple source languages across aligned corpora. In *IJCNLP*.

Wael H. Gomaa and Aly A. Fahmy. 2013. Article: A survey of text similarity approaches. *International Journal of Computer Applications*, 68(13):13–18, April. Full text available.

Richard W Hamming. 1980. *Coding and Theory*. Prentice-Hall Englewood Cliffs.

Hua He, Kevin Gimpel, and Jimmy J. Lin. 2015. Multi-perspective sentence similarity modeling with convolutional neural networks. In *EMNLP*, pages 1576–1586.

Paul Jaccard. 1912. The distribution of the flora in the alpine zone. *New Phytologist*, 11(2):37–50.

Tom Kenter and Maarten de Rijke. 2015. Short text similarity with word embeddings. In *CIKM*, pages 1411–1420. ACM.

Joo-Kyung Kim, Young-Bum Kim, Ruhi Sarikaya, and Eric Fosler-Lussier. 2017. Cross-lingual transfer learning for POS tagging without cross-lingual resources. In *EMNLP*, pages 2832–2838, Copenhagen, Denmark, September.

Joseph B Kruskall. 1983. The symmetric time warping algorithm: From continuous to discrete. *Time warps, string edits and macromolecules*.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Vladimir I Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710.

Marco Lippi, Giuseppe Contissa, Francesca Lagioia, Hans-Wolfgang Micklitz, Przemysław Pałka, Giovanni Sartor, and Paolo Torroni. 2019a. Consumer protection requires artificial intelligence. *Nature Machine Intelligence*, 1(4):168.

Marco Lippi, Przemysław Pałka, Giuseppe Contissa, Francesca Lagioia, Hans-Wolfgang Micklitz, Giovanni Sartor, and Paolo Torroni. 2019b. CLAUDETTE: An automated detector of potentially unfair clauses in online terms of service. *Artificial Intelligence and Law*, 27(2):117–139.

Iñigo Lopez-Gazpio, Montse Maritxalar, M. Lapata, and Eneko Agirre. 2019. Word n-gram attention models for sentence similarity and inference. *Expert Syst. Appl.*, 132:1–11.

Robert C. Moore. 2002. Fast and accurate sentence alignment of bilingual corpora. In Stephen D. Richardson, editor, *Machine Translation: From Research to Real Users*, pages 135–144, Berlin, Heidelberg. Springer Berlin Heidelberg.

Jonas Mueller and Aditya Thyagarajan. 2016. Siamese recurrent architectures for learning sentence similarity. In *AAAI*, pages 2786–2792. AAAI Press.

Saul B. Needleman and Christian D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443 – 453.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *HLT-NAACL*, pages 2227–2237, New Orleans, Louisiana, June.

Xiaojun Quan, Chunyu Kit, and Wuya Chen. 2018. Collaborative matching for sentence alignment. In Maosong Sun, Ting Liu, Xiaojie Wang, Zhiyuan Liu, and Yang Liu, editors, *CCL*, pages 39–52, Cham. Springer International Publishing.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, November. Association for Computational Linguistics.

Gil Rocha, Christian Stab, Henrique Lopes Cardoso, and Iryna Gurevych. 2018. Cross-lingual argumentative relation identification: from English to Portuguese. In *ArgMining*, pages 144–154, Brussels, Belgium, November.

Stan Salvador and Philip Chan. 2007. Toward accurate dynamic time warping in linear time and space. *Intell. Data Anal.*, 11(5):561–580, October.

André Santos. 2011. A survey on parallel corpora alignment. In *MI-STAR*, pages 117–128, Braga, Portugal, January.

Michel Simard and Pierre Plamondon. 1998. Bilingual sentence alignment: Balancing robustness and accuracy. *Machine Translation*, 13(1):59–80, Mar.

Ralf Steinberger, Mohamed Ebrahim, Alexandros Poulis, Manuel Carrasco-Benitez, Patrick Schlüter, Marek Przybyszewski, and Signe Gilbro. 2014. An overview of the european union's highly multilingual parallel corpora. *Language Resources and Evaluation*, 48(4):679–707, Dec.

Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J van der Walt, Matthew Brett, Joshua Wilson, K Jarrod Millman, Nikolay Mayorov, Andrew R J Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E A Quintero, Charles R Harris, Anne M Archibald, Antônio H Ribeiro, Fabian Pedregosa, and Paul van Mulbregt. 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3):261–272, mar.

Ruochen Xu, Yiming Yang, Naoki Otani, and Yuexin Wu. 2018. Unsupervised cross-lingual transfer of word embedding spaces. In *EMNLP*, pages 2465–2474, Brussels, Belgium, november.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *HLT*, pages 1–8.

Hamed Zamani, Heshaam Faili, and Azadeh Shakery. 2016. Sentence alignment using local and global information. *Computer Speech & Language*, 39:88 – 107.

Yuan Zhang, David Gaddy, Regina Barzilay, and Tommi Jaakkola. 2016. Ten pairs to tag – multilingual POS tagging via coarse mapping between embeddings. In *HLT-NAACL*, pages 1307–1317, San Diego, California, June.