Cross-lingual Annotation Projection in Legal Texts

Andrea Galassi[1] , Kasper Drazewski[2], Marco Lippi[3,4], Paolo Torroni[1,4]

[1]University of Bologna, Italy; [2]Bureau Européen des Unions de Consommateurs, Belgium;
[3]University of Modena and Reggio Emilia, Italy; [4]European University Institute, Italy

Problem Definition

Context

- Two parallel asymmetric corpora in two different languages.
- One of the two is annotated (source), the other is not (target).
- Goal: **unsupervised transfer of annotations from one corpus to the other**.
  - Desired property: **language agnosticism**.
- Case study: Terms of Service and Privacy Policies. Sources in English, targets in German.

Method

1) Automatically translate the target document in English.
2) Find correspondence between sentences of the two documents in English.
3) Transfer labels between similar English sentences.
4) Transfer them to the original German document.

Corpus

Content

10 documents (~4500 sentences) available in English and German

- 5 Privacy Policies.
- 5 Terms of Service contracts.

Annotations

Sentences labeled for unfair clause detection.

- 50 labels: reason and degree of unfairness
- Sentences may have more multiple labels.

Characteristics

Parallelism:
English and German version of each document have the same content.

Asymmetry:
Discrepancies between the two versions, usually in the structure of the sentences.
e.g. One sentence in English may correspond to multiple sentences in German, possibly with different labels.

Projection Architecture

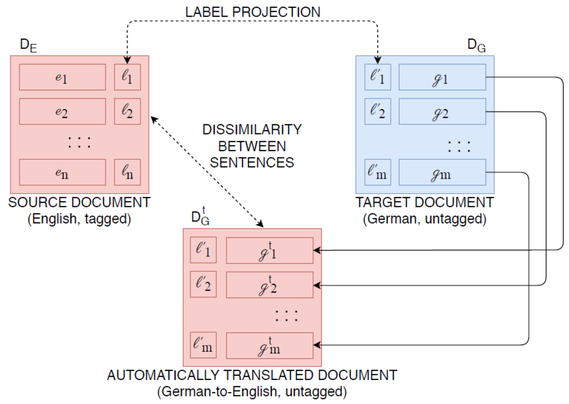
Input

D_E : source, annotated, English
D_G : target, non-annotated, German
D_G^t : target translation, non-annotated, English

Goal

Match each sentence of D_G^t with (at least) one sentence of D_e.
Dissimilarity computed using

- Textual representation (P_t)
- ELMo embedded representation (P_e).



Results

Evaluation

- Multi-label classification of target documents sentences.
- Reaching scores of 1.00 may be impossible due to asymmetry.

Results

- Use of embeddings outperforms classical approaches.
- DTW improves all approaches.

	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**

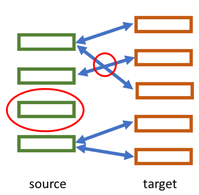
	P_{t+DTW}						P_{e+DTW}	
	Hamming	Levenshtein	Damerau-Levenshtein	Needleman-Wunsch	Jaccard	NCD	Cosine	Bray-Curtis
F1-macro	0.78	0.79	0.79	0.70	0.78	0.74	0.81	**0.82**
F1-micro	0.82	0.83	0.83	0.72	0.83	0.79	**0.86**	**0.86**
F1-weighted	0.82	0.83	0.83	0.72	0.83	0.79	**0.86**	**0.86**
Precision	0.90	0.91	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**

Matching Criterion

Baseline approach

Match each unlabelled sentence with the least dissimilar one.

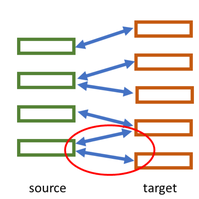
- “Crossings” are allowed in the matches.
- Some source sentences may not be matched.
- Only one match for target sentence.



Dynamic Time Warping

Algorithm for time series alignment. Helps dealing with asymmetries.

- No “crossings” in the matches.
- All sentences are matched.
- Multiple matches are possible.



Conclusion

Discussion

- Best result: embeddings + DTW small computational footprint!
- Given an automatic translation method, the approach is language agnostic

This method can be used to **create new annotated corpora** in any language.

Future Works

- Use of advanced sentence embeddings
- Train classification systems on corpora generated through projection.

Corpus and code available at https://bitbucket.org/a-galaxy/cross-lingual-annotation-projection-in-legal-texts