Cross-lingual Annotation Projection in Legal Texts
Andrea Galassi¹, Kasper Drazewski², Marco Lippi³,⁴, Paolo Torroni¹,⁴

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

Projection Architecture

Input
- Dₑ: source, annotated, English
- D₉: target, non-annotated, German
- Dₑ*: target translation, non-annotated, English

Goal
Match each sentence of Dₑ with (at least) one sentence of D₉.
Dissimilarity computed using
- Textual representation (Pₑ)
- ELMo embedded representation (Pₑ)

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.

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

Conclusion
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