20th EPIA Conference on Artificial Intelligence

Cross-Lingual Annotation Projection for Argument Mining in Portuguese

Afonso Sousa, Bernardo Leite, Gil Rocha and Henrique Lopes Cardoso

LIACC

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

7-9 September, 2021

Motivation

In Argument Mining (AM), there are not yet many works emphasizing cross-lingual approaches

- Projection
- Direct Transfer

Most works carry out a monolingual setting in English

Acquiring high-quality corpora for new languages is an expensive effort

Contributions

Build a Portuguese machine-translated version of the Persuasive Essays corpus [Stab and Gurevych, 2017]

Formalize an annotation projection algorithm and propose a tweak to enhance performance for verbose languages

Extensively evaluate the quality of the created dataset

Corpus

Persuasive Essays comprises 402 essays written by students in response to controversial topics.

The annotations distinguish between:

- Major Claim
- Claim
- Premise

*Ever since researchers at the Roslin Institute in Edinburgh cloned an adult sheep, there has been an ongoing debate about whether cloning technology is morally and ethically right or not. Some people argue for and others against and there is still no agreement whether cloning technology should be permitted. However, as far as I'm concerned, [**cloning is an important technology for humankind**]$_{MajorClaim1}$ since [it would be very useful for developing novel cures]$_{Claim1}$.*

First, [cloning will be beneficial for many people who are in need of organ transplants]$_{Claim2}$. [Cloned organs will match perfectly to the blood group and tissue of patients]$_{Premise1}$ since [they can be raised from cloned stem cells of the patient]$_{Premise2}$. In addition, [it shortens the healing process]$_{Premise3}$. Usually, [it is very rare to find an appropriate organ donor]$_{Premise4}$ and [by using cloning in order to raise required organs the waiting time can be shortened tremendously]$_{Premise5}$.

[Stab and Gurevych, 2017]

Persuasive Essays - CoNLL-Format

- 1-based indices
- tokens
- BIO Scheme
- ADU types
- indices of targeted ADU
- relation types

44	are	I	Premise	54	Support
45	watched	I	Premise	54	Support
46	closely	I	Premise	54	Support
47	and	I	Premise	54	Support
48	judged	I	Premise	54	Support
49	by	I	Premise	54	Support
50	the	I	Premise	54	Support
51	community	I	Premise	54	Support
52	.	O			
53	But	O			
54	during	B	Claim	For	
55	our	I	Claim	For	
56	childhood	I	Claim	For	

Example CoNLL-formatted snippet

Projection Pipeline

1. Translation — Perform machine translation (English to Portuguese) of the PE corpus, using a transformer-based model

⇩

2. Alignment — Align each of the corresponding words for each pair of sentences (English and Portuguese)

⇩

3. Projection — Project existing annotations from the source language (English) to the target language (Portuguese)

1. Translation

Translations conducted at the paragraph level

Use of pre-trained transformer-based translation models

Transformer-based model pretrained on OPUS dataset (https://opus.nlpl.eu/)

Example:

EN - Another technological innovations which help people around the world is related to medical equipments .

PT - Outra inovação tecnológica que ajuda pessoas em todo o mundo está relacionada com equipamentos médicos .

2. Alignment

SimAlign: Alignment tool that uses contextualized embeddings and no parallel data [Sabet *et al.*, 2020]

- EN: First$_0$ and$_1$ foremost$_2$,$_3$ email$_4$ can$_5$ be$_6$ count$_7$ as$_8$ one$_9$ of$_{10}$ the$_{11}$ most$_{12}$ beneficial$_{13}$ results$_{14}$ of$_{15}$ modern$_{16}$ technology$_{17}$.$_{18}$
- PT: Primeiro$_0$ e$_1$ acima$_2$ de$_3$ tudo$_4$,$_5$ o$_6$ email$_7$ pode$_8$ ser$_9$ contado$_{10}$ como$_{11}$ um$_{12}$ dos$_{13}$ resultados$_{14}$ mais$_{15}$ benéficos$_{16}$ da$_{17}$ tecnologia$_{18}$ moderna$_{19}$.$_{20}$
- EN $\rightarrow$ PT Alignment: [0-0, 1-1, 2-2, 2-3, 2-4, 3-5, 4-7, 5-8, 6-9, 7-10, 8-11, 9-12, 10-13, 11-16, 12-15, 13-16, 14-14, 15-17, 16-19, 17-18, 18-20]

3. Annotation Projection

Algorithm 1: Building target ADUs for a sentence

Data: s_{ADU} is the set of source ADUs as triplets $\langle start, end, label \rangle$;
$sent\_align$ is the set of aligned index pairs in the sentence
Result: set of target ADUs' triplets
for $\langle start, end, label \rangle \in s_{ADU}$ **do**
    $align_{ADU} \leftarrow sent\_align[start, end]$;
    $align^{t}_{ADU} \leftarrow align_{ADU}.targets$;
    $idx_{min} \leftarrow min(align^{t}_{ADU})$;
    $idx_{max} \leftarrow max(align^{t}_{ADU})$;
    if $idx_{min} > 0$ *and* $idx_{min} - 1 \notin sent\_align.targets$ **then**
        $idx_{min} \leftarrow idx_{min} - 1$;
    $t_{ADU} \leftarrow t_{ADU} \cup \langle idx_{min}, idx_{max}, label \rangle$;

PADDING HEURISTIC

3. Annotation Projection Example

- EN: In_0 fact_1 $,_2$ [stringent_3 gun_4 control_5 does_6 not_7 decrease_8 violence_9 and_{10} crime_{11}] because_{12} [most_{13} gun_{14} violence_{15} is_{16} committed_{17} with_{18} guns_{19} obtained_{20} illegally_{21}] $.22$
- PT: De_0 fato_1 $,_2$ [o_3 rigoroso_4 controle_5 de_6 armas_7 não_8 diminui_9 a_{10} violência_{11} e_{12} o_{13} crime_{14}] porque_{15} [a_{16} maioria_{17} da_{18} violência_{19} de_{20} armas_{21} é_{22} cometida_{23} com_{24} armas_{25} obtidas_{26} ilegalmente_{27}] $.28$
- EN $\rightarrow$ PT Alignment: [0-0, 1-1, 2-2, 3-4, 4-7, 5-5, 6-9, 7-8, 8-9, 9-11, 10-12, 11-14, 12-15, 13-17, 14-21, 15-19, 16-22, 17-23, 18-24, 19-25, 20-26, 21-27, 22-28]

Intrinsic Evaluation

Construct English dataset with projection from annotations of the Portuguese corpus

- Compute back-alignment
- Compute back-projection

	# Tokens	Precision	Recall	F1
B-Claim	1506	0.83	0.83	0.82
B-MajorClaim	751	0.81	0.81	0.81
B-Premise	3832	0.90	0.90	0.89
I-Claim	20937	0.94	0.99	0.96
I-MajorClaim	10215	0.92	0.99	0.95
I-Premise	63326	0.98	0.99	0.98
O	47615	1.00	0.94	0.96
Overall	148182	0.91	0.92	0.91

Token level evaluation of the English back-projected corpus (on test set)

Error Analysis

- EN_{gold}: First_0 of_1 all_2 $,_3$ it_4 seems_5 to_6 be_7 true_8 that_9 [if_{10} there_{11} is_{12} gun_{13} control_{14} somehow_{15} crime_{16} has_{17} to_{18} decrease_{19}] because_{20} [fewer_{21} guns_{22} available_{23} mean_{24} less_{25} crime_{26}] $.27$
- $\text{PT}_{projected}$: Em_0 primeiro_1 lugar_2 $,_3$ parece_4 ser_5 verdade_6 que_7 [se_8 há_9 controle_{10} de_{11} armas_{12} de_{13} alguma_{14} forma_{15} o_{16} crime_{17} tem_{18} de_{19} diminuir_{20}] porque_{21} [menos_{22} armas_{23} disponíveis_{24} significa_{25} menos_{26} crime_{27}] $.28$
- PT $\rightarrow$ EN Alignment: [0-1, 1-0, 2-2, 3-3, 4-5, 5-7, 6-8, 7-9, 8-10, 9-11, 10-14, 11-12, 12-13, 13-15, 15-15, 16-4, 17-16, 18-17, 19-18, 20-6, 20-19, 21-20, 22-21, 23-22, 24-23, 25-24, 26-25, 27-26, 28-27]
- $\text{EN}_{back-projected}$: First_0 of_1 all_2 $,_3$ [$\underline{\text{it}}_4$ $\underline{\text{seems}}_5$ $\underline{\text{to}}_6$ $\underline{\text{be}}_7$ $\underline{\text{true}}_8$ $\underline{\text{that}}_9$ if_{10} there_{11} is_{12} gun_{13} control_{14} somehow_{15} crime_{16} has_{17} to_{18} decrease_{19}] because_{20} [fewer_{21} guns_{22} available_{23} mean_{24} less_{25} crime_{26}] $.27$

Extrinsic Evaluation

Sequence tagging (classes: B-Claim, B-MajorClaim, B-Premise, I-Claim, I-MajorClaim, I-Premise, O)

Perform **Projection** and **Direct Transfer**
from English to Portuguese (PT) (**ours**),
Spanish (ES) and German (DE) [Eger *et al.*, 2018]

Experiment with BLCRF, BLCRF+Char, and mBERT+CRF

Extrinsic Evaluation - Results

			Machine Translated		
		EN	PT	ES	DE
Projection	BLCRF + Char	73.02 ± 1.32	68.59 ± 0.77	69.35 ± 0.97	67.76 ± 0.36
	BLCRF	72.49 ± 1.69	65.46 ± 0.85	66.23 ± 1.03	65.79 ± 1.43
	mBERT + CRF	75.74 ± 0.37	70.12 ± 0.97	71.29 ± 1.06	67.89 ± 0.56
Direct Transfer	mBERT + CRF		63.76 ± 2.44	66.07 ± 2.49	57.12 ± 2.73

Projection and direct transfer (EN→PT, EN→ES and EN→DE) sequence tagging results on the test set (token-level macro-F1)

Conclusions

Proposed an annotation projection procedure

Created a Portuguese version of the Persuasive Essays corpus
(https://github.com/AfonsoSalgadoSousa/argumentation_mining_pt)

Experiments show promising results in the quality of the projected annotations and sequence tagging results on par with other languages

Our approach is reliant on the quality of the translations and alignments

Future Work

Extend the analysis to other translation and alignment methods

Employ this approach for other languages and compare results

Employ this approach in other NLP annotated corpora

20th EPIA Conference on Artificial Intelligence

Thank you!

Cross-Lingual Annotation Projection for Argument Mining in Portuguese

Afonso Sousa, Bernardo Leite, Gil Rocha and Henrique Lopes Cardoso

LIACC

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

7-9 September, 2021