Empirical Methods in Natural Language Processing (EMNLP 2018)
5th Workshop on Argument Mining (ARGMINING 2018)

Cross-Lingual Argumentative Relation Identification: from English to Portuguese

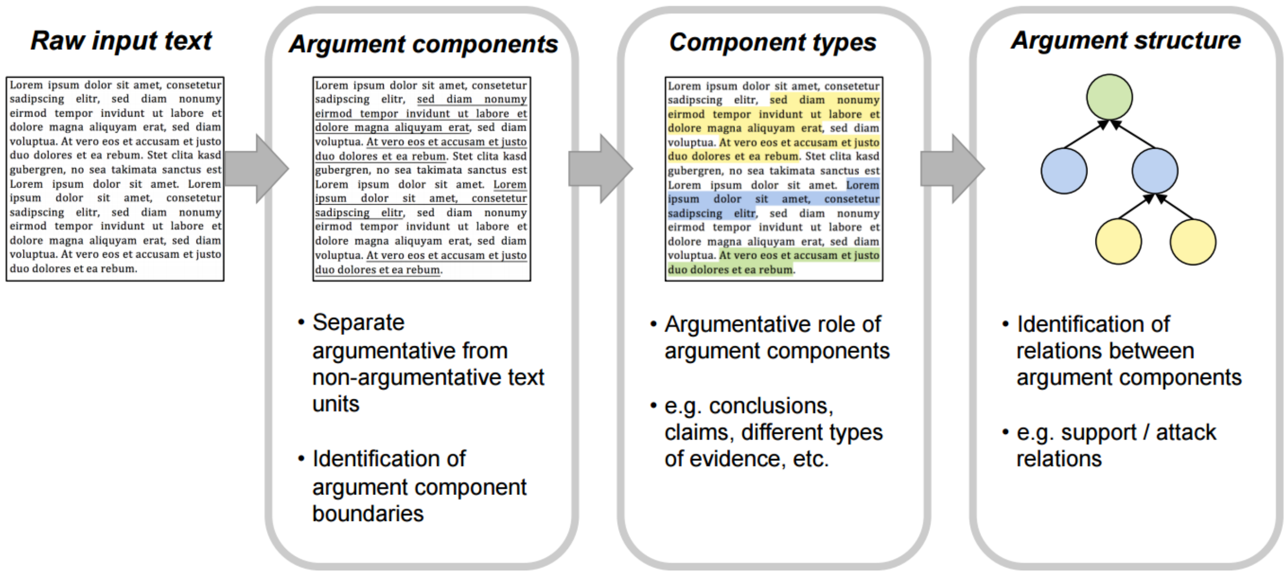
Gil Rocha, Christian Stab, Henrique Lopes Cardoso and Iryna Gurevych

LIACC/DEI, Faculty of Engineering, University of Porto

Ubiquitous Knowledge Processing Lab (UKP-TUDA), Department of Computer Science, Technische Universitat Darmstadt

01/11/2018

AM Tasks



- Focus on AM subtask of Argumentative Relation Identification [Peldszus and Stede, 2015]
- Assumption: ADUs are given as input (no ADU classification is assumed)
- Task formulation:
  - Given two ADUs determine whether they are argumentatively linked or not

AM for Less Resourced Languages

- Resources are scarce in terms of:
  - Annotations of arguments
    - Challenging and time-consuming task [Habernal et al., 2014]
    - **Proposed Approach:** Cross-Language Learning
  - Available tools and annotated resources for auxiliary NLP tasks
    - Heavily engineered NLP pipelines tend to underperform
    - **Proposed Approach:** (Multi-Lingual) Word Embeddings + Deep Neural Network Architectures

Cross-Language Learning for AM

- **Proposed approach:** explore existing corpora in different languages to improve the performance of the system on less-resourced languages
- **Hypothesis:**
  - High-level semantic representations that capture the argumentative relations between ADUs can be independent of the language
- **Contributions:**
  - First attempt to address the task of Argumentative Relation Identification in a cross-lingual setting
  - Unsupervised cross-language approaches suited for less-resourced languages

Related Work

Mono-Lingual Setting

- Argumentative Relation Identification
  - Subtask addressed in isolation
    - Feature-based approach [Nguyen and Litman, 2016]
    - NN architecture (LSTMs for sentence encoding) [Bosc et al., 2016; Cocarascu and Toni, 2017]
  - Jointly modeled with previous subtasks
    - Feature-based approach and ILP [Stab and Gurevych, 2017]
    - End-to-End AM System [Eger et al., 2017]
    - Encoder-decoder formulation employing a pointer network [Potash et al., 2017]
- Discourse Parsing
  - NN architecture: Sentence Encoding using word embeddings + lexical + syntactic info) [Braud et al., 2017; Li et al., 2014]
- Recognizing Textual Entailment
  - Different sentence encoding techniques
    - Recurrent [Bowman et al., 2015a] and Recursive neural networks [Bowman et al., 2015a]
  - Complex aggregation functions [Rocktaschel et al., 2015; Chen et al., 2017; Peters et al., 2018]

Related Work

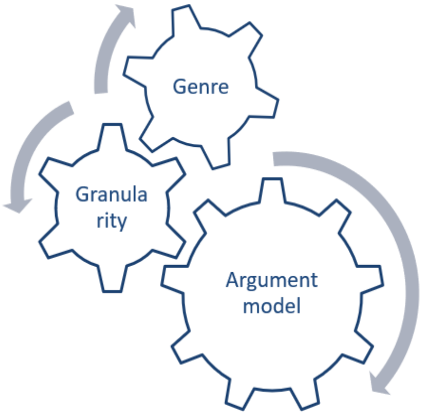
Cross-Lingual Setting

- **Cross-Language Learning:** obtain an intermediate and shared representation of the data that can be employed to address a specific task across different languages
- Current approaches can be divided in:
  - Projection
  - Direct Transfer
    - Training only on the source language
    - Re-Training on the target language
- Related tasks:
  - Textual Entailment and Semantic Similarity
  - Sequence Tagging approaches
    - NER, PoS Tagging, Sentiment classification, Discourse parsing
  - Argumentation Mining
    - Argument Component Identification and Classification [Eger et al., 2018a]
    - Argumentative Sentence Detection (PD3) [Eger et al., 2018b]

AM Corpora with relations

Lang	Corpus	#Docs	#Rel	#None	#Support	#Attack	Arg. Schema	Type
EN	Argumentative Essays	402	22,172	17,923	3,918	331	Premise, Claim, Major Claim	Essays
PT	ArgMine	75	778	621	153	4	Premise, Claim	Opinion Articles

Table 2. Corpora Statistics: Argumentative Essays (EN) [Stab and Gurevych, 2017] and ArgMine corpus (PT) [Rocha and Lopes Cardoso, 2017]



Lang.	Source ADU	Target ADU	Label
EN	Teachers are not just teachers, they are also friends and conseileurs	In conclusion, there can be no school without a teacher	support
	computers need to be operated by people	no one can argue that technological tools are must-haves for the classroom	none
PT	Durante a última década, a saúde, o meio ambiente, a biodiversidade, assim como a evolução humana tem sido temas recorrentes em todos os meios de comunicação. *(During the last decade, health, environment, biodiversity, as well as human evolution have been recurring topics in all sorts of media)*	O século XXI é sem sombra de dúvida a era da Biologia *(The 21st century is undoubtedly the era of biology)*	support
	Seria da mais elementar prudência não voltar a precisar de lhe pedir dinheiro *(It would be most prudent not to need asking it money again)*	O fluxo de migrantes agravou o peso do euroceptismo nos governos *(The flow of migrants has increased the weight of euroscepticism in governments)*	none

Table 3. Annotated examples extracted from the corpora

Data Preparation

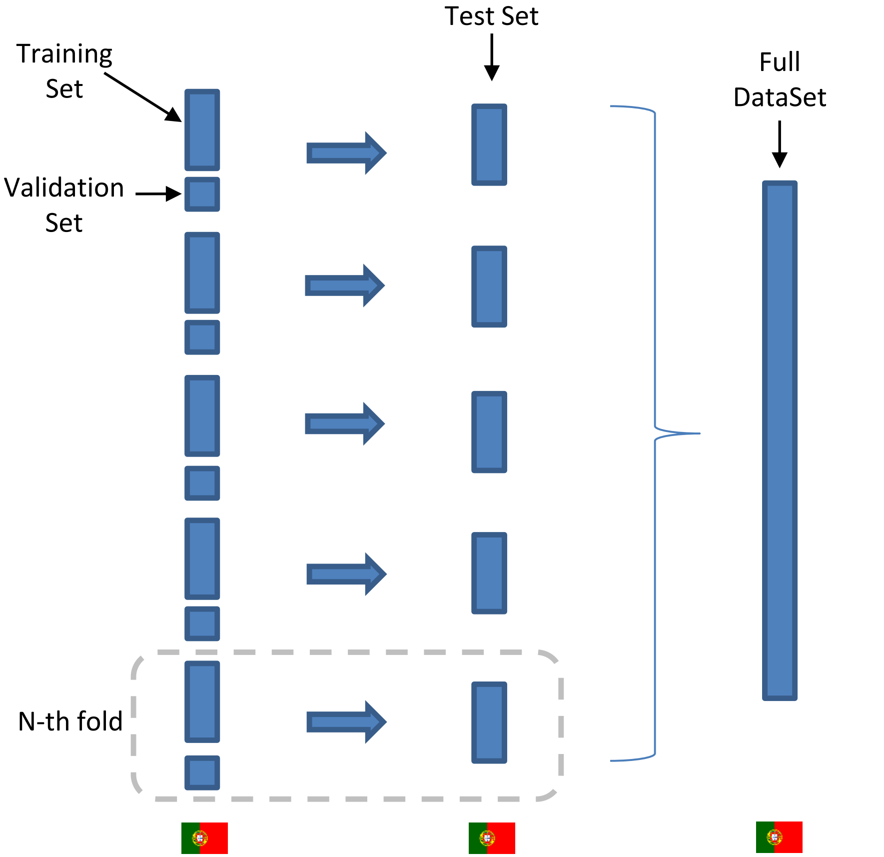
- Input: text annotated with argumentative content at the token level
- Output: ADU pairs annotated with labels: None, Support and Attack

- Procedure:
  - For each pair of ADUs $\langle A_1, A_2 \rangle$ in the same paragraph:
    - If A_1 is connected to A_2 with label L, with $L \in \{Support, Attack\}$
      - use label L
    - Otherwise,
      - use label $None$

Experimental Setup

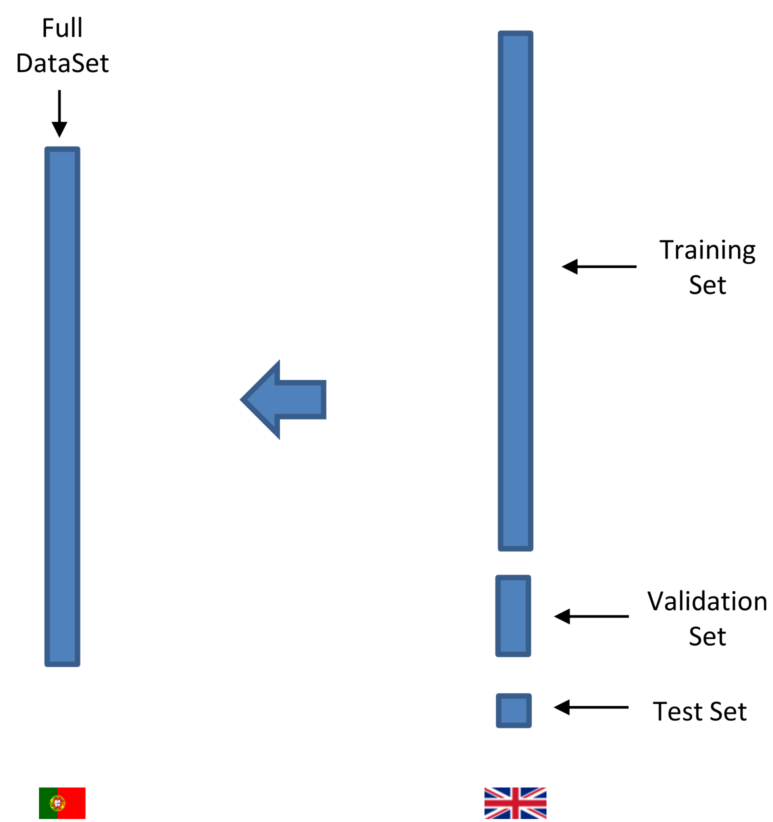
In-Language experiments:

(e.g. PT)



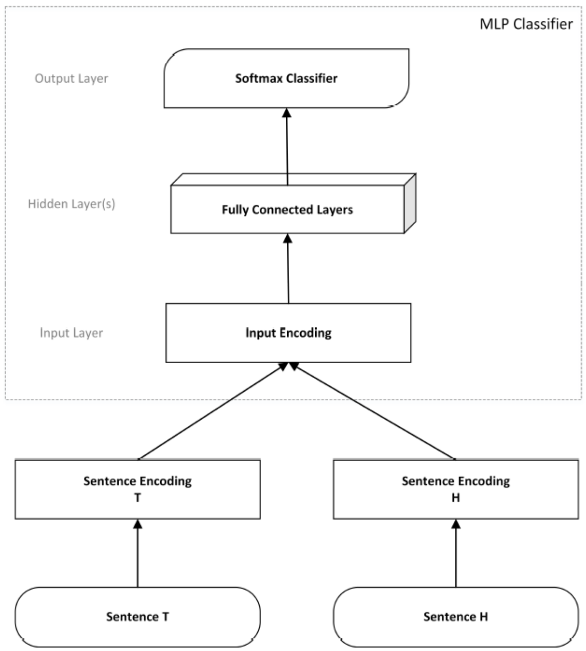
Cross-Language experiments:

(e.g. Direct Transfer from EN to PT)



Methods

- Baselines
  - BoW encoding + Logistic Regression
  - Enhanced Sequential Inference Model (ESIM) [Chen et al., 2017]
  - AllenNLP TE model [Peters et al., 2018]
- Explored architectures
  - Different ways of encoding the sentence
    - Sum of Word Embeddings
    - LSTMs and BiLSTMs
    - Convolutional
    - Conditional Encoding
- Dealing with unbalanced datasets
  - Random Undersampling
  - Cost-Sensitive Learning



Results: In-Language EN

- NN architectures outperform baselines
- State-of-the-Art RTE models perform poorly
  - Tasks are conceptually different
  - Models are too complex for the relatively small amount of data
- Skewed nature of the dataset plays an important role

Baselines

Model	*Macro-F1*	*F1-None*	*F1-Supp*
Random	.447	.625	.269
Peters et al. (2018)	.512	.903	.121
Chen et al. (2017)	.577	.879	.275
BoW+LR	.604	.898	.311
LSTM	.606	.877	.336
BiLSTM	.624	.867	.381
Conv1D	.634	.879	.390
Inner-Att	.621	.882	.360

Results: In-Language EN

- CSL and RU do not improve overall performance
- Simple BoW + LR obtains better macro f1-score
- Results are worst than existing SOTA work:
  - [Potash et al., 2017] reports 0,767 macro f1-score
  - Notice that existing SOTA work:
    - Do not scaled for cross-lingual settings targeting less-resourced languages
    - Modeled the problem differently

Model	*Macro-F1*	*F1-None*	*F1-Supp*
Random	.447	.625	.269
Peters et al. (2018)	.512	.903	.121
Chen et al. (2017)	.577	.879	.275
BoW+LR	.604	.898	.311
LSTM	.606	.877	.336
BiLSTM	.624	.867	.381
Conv1D	.634	.879	.390
Inner-Att	.621	.882	.360
Cost Sensitive Learning			
BoW+LR	**.641**	.875	.407
LSTM	.616	.822	.410
BiLSTM	.634	.835	**.434**
Conv1D	.631	.832	.430
Inner-Att	.606	.822	.410
Random Undersampling			
BoW+LR	.574	.748	.401
LSTM	.566	.734	.399
BiLSTM	.609	.796	.422
Conv1D	.598	.786	.410
Inner-Att	.586	.775	.397

Results: In-Language PT

- Similar trend compared to In-Language EN results
  - CSL and RU are more effective to increase the scores on the Support label

Baselines

	In-Language		
Model	*Macro*	*None*	*Supp*
Random	.448	.613	.283
BoW+LR	.457	.888	.025
Peters et al. (2018)	.485	.887	.082
Chen et al. (2017)	.522	.856	.188
LSTM	.489	.868	.110
BiLSTM	.510	.840	.180
Conv1D	.459	.882	.035
Inner-Att	**.534**	.764	.305
Cost Sensitive Learning			
BoW+LR	.520	.846	.193
LSTM	.496	.680	.312
BiLSTM	.523	.786	.259
Conv1D	.503	.827	.178
Inner-Att	.479	.637	.321
Random Undersampling			
BoW+LR	.264	.191	.337
LSTM	.494	.668	.321
BiLSTM	.464	.581	.348
Conv1D	.423	.554	.292
Inner-Att	.487	.621	**.352**

Results: Cross-Language EN to PT

- Cross-Language scores are close to in-language scores (better in some settings)

	In-Language			Direct Transfer			Projection		
Model	*Macro*	*None*	*Supp*	*Macro*	*None*	*Supp*	*Macro*	*None*	*Supp*
LSTM	.489	.868	.110	.461	.887	.036	.462	.884	.041
BiLSTM	.510	.840	.180	.463	.870	.057	.466	.877	.055
Conv1D	.459	.882	.035	.459*	.880	.038	.462*	.884	.039
Inner-Att	**.534**	.764	.305	.454	.883	.025	.456	.882	.030
Cost Sensitive Learning									
LSTM	.496	.680	.312	.489	.870	.109	.493	.849	.137
BiLSTM	.523	.786	.259	.485	.861	.109	.503	.845	**.162**
Conv1D	.503	.827	.178	.497	.854	.141	.494	.841	.147
Inner-Att	.479	.637	.321	.477	.867	.088	.484*	.844	.123
Random Undersampling									
LSTM	.494	.668	.321	.494*	.870	.118	.495*	.859	.131
BiLSTM	.464	.581	.348	**.500***	.856	**.145**	**.512***	.865	.158
Conv1D	.423	.554	.292	.499*	.855	.144	.492*	.849	.134
Inner-Att	.487	.621	**.352**	.482	.878	.087	.495*	.861	.128

Results: Cross-Language EN to PT

- CSL and RU consistently improves the overall macro f1-score

	In-Language			Direct Transfer			Projection		
Model	*Macro*	*None*	*Supp*	*Macro*	*None*	*Supp*	*Macro*	*None*	*Supp*
LSTM	.489	.868	.110	.461	.887	.036	.462	.884	.041
BiLSTM	.510	.840	.180	.463	.870	.057	.466	.877	.055
Conv1D	.459	.882	.035	.459*	.880	.038	.462*	.884	.039
Inner-Att	**.534**	.764	.305	.454	.883	.025	.456	.882	.030
Cost Sensitive Learning									
LSTM	.496	.680	.312	.489	.870	.109	.493	.849	.137
BiLSTM	.523	.786	.259	.485	.861	.109	.503	.845	**.162**
Conv1D	.503	.827	.178	.497	.854	.141	.494	.841	.147
Inner-Att	.479	.637	.321	.477	.867	.088	.484*	.844	.123
Random Undersampling									
LSTM	.494	.668	.321	.494*	.870	.118	.495*	.859	.131
BiLSTM	.464	.581	.348	**.500***	.856	**.145**	**.512***	.865	.158
Conv1D	.423	.554	.292	.499*	.855	.144	.492*	.849	.134
Inner-Att	.487	.621	**.352**	.482	.878	.087	.495*	.861	.128

Results: Cross-Language EN to PT

- Projection approach >> Direct Transfer (in most of the settings)

	In-Language			Direct Transfer			Projection		
Model	*Macro*	*None*	*Supp*	*Macro*	*None*	*Supp*	*Macro*	*None*	*Supp*
LSTM	.489	.868	.110	.461	.887	.036	.462	.884	.041
BiLSTM	.510	.840	.180	.463	.870	.057	.466	.877	.055
Conv1D	.459	.882	.035	.459*	.880	.038	.462*	.884	.039
Inner-Att	**.534**	.764	.305	.454	.883	.025	.456	.882	.030
Cost Sensitive Learning									
LSTM	.496	.680	.312	.489	.870	.109	.493	.849	.137
BiLSTM	.523	.786	.259	.485	.861	.109	.503	.845	**.162**
Conv1D	.503	.827	.178	.497	.854	.141	.494	.841	.147
Inner-Att	.479	.637	.321	.477	.867	.088	.484*	.844	.123
Random Undersampling									
LSTM	.494	.668	.321	.494*	.870	.118	.495*	.859	.131
BiLSTM	.464	.581	.348	**.500***	.856	**.145**	**.512***	.865	.158
Conv1D	.423	.554	.292	.499*	.855	.144	.492*	.849	.134
Inner-Att	.487	.621	**.352**	.482	.878	.087	.495*	.861	.128

Error Analysis

- Text genre shift:
  - Linguistic indicators
    - Prevail in Argumentative Essays (EN) [Stab and Gurevych, 2017]
    - Ambiguous and rare in ArgMine Corpus (PT) [Rocha and Lopes Cardoso, 2017]
  - ArgMine Corpus (PT) is more demanding in terms of common-sense knowledge and temporal reasoning

ADU_S: "Greece, last year, tested the tolerance limits of other European taxpayers"
ADU_T: "The European Union of 2016 is no longer the one of 2011."

- Distinction between linked and convergent arguments
  - During data preparation both cases were considered as convergent

Conclusions

- Competitive results can be obtained using unsupervised language adaptation when compared to in-language supervised approach
  - Cross-lingual transfer loss is relatively small (always below 10% macro f1)
    - In some settings cross-language approaches outperform in-language approaches
- Higher-level representations of argumentative relations can be obtained that can be transferred across languages
    - **Future work:** Evaluate approach in other languages
- Existing corpora poses many challenges
  - Annotations using different argument models
    - Cross-lingual approaches are hard to explore (requires extra pre-processing steps)
    - **Solution:** Frame the problem as MTL; PD3 approach [Eger et al., 2018b]
  - Domain shift needs to be investigated in more detail
    - **Future work:** employ MTL and/or adversarial training approaches

Questions?

Code available:

https://github.com/GilRocha/emnlp2018-argmin-workshop-xLingArgRelId

Contact:

Gil Rocha

Artificial Intelligence and Computer Science Lab (LIACC)

Faculty of Engineering, University of Porto (FEUP)

Email: gil.rocha@fe.up.pt