Predicting Argument Density from Multiple Annotations

Gil Rocha[1,3], Bernardo Leite[1,3], Luís Trigo[1,3], Henrique Lopes Cardoso[1,3], Rui Sousa-Silva[2,4], Paula Carvalho[5], Bruno Martins[5], and Miguel Won[5]

[1] Faculdade de Engenharia, Universidade do Porto, Portugal
[2] Faculdade de Letras, Universidade do Porto, Portugal
[3] Laboratório de Inteligência Artificial e Ciência de Computadores (LIACC)
[4] Centro de Linguística da Universidade do Porto (CLUP)
[5] INESC-ID, Lisboa, Portugal

NLDB 2022 June 15-17, 2022 Universitat Politècnica de Valencia, Spain

Cofinanciado por:

Outline

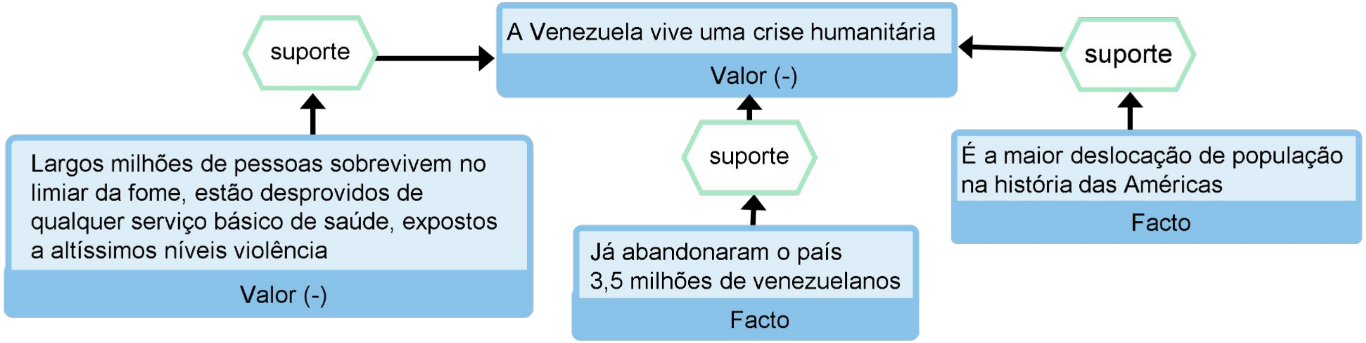
- Introduction
  - context and research goals
- Argument Density Prediction and Ranking
  - tasks, aggregation strategies
- Experiments and Results
- Conclusions

Introduction

Argument Annotation

A corpus of Portuguese **opinion articles** annotated with arguments (DARGMINTS project)

1. Primeiro, as primeiras coisas. A Venezuela vive uma crise humanitária. Não há como a esconder. Largos milhões de pessoas sobrevivem no limiar da fome, estão desprovidos de qualquer serviço básico de saúde, expostos a altíssimos níveis violência. Já abandonaram o país 3,5 milhões de venezuelanos, dos quais um terço se refugiou na Colômbia, havendo regiões onde a situação se antolha dramática. É a maior deslocação de população na história das Américas.



Argument Annotation

Annotating a corpus with **argument structures** is a complex task

- requires semantically-demanding **interpretation skills**
- argumentative **discourse markers** may be absent
- difficult to obtain **agreement** between different annotators

Different **text genres** include argumentative content of varying degrees

- **Essays** are highly structured documents: explicit argumentation, full of **discourse markers**
- **Opinion articles** tend to be more subtle: argumentative reasoning steps are harder to capture

Agreement and Perspectivism

373 opinion articles, each with 3 annotations (from a pool of 4 annotators)

- **agreement** on identifying argumentative discourse unit (**ADU**) spans [Rocha et al., LREC 2022]
  - Krippendorff's $\boldsymbol{\alpha}$ = **0.33** (**fair agreement**)

	A,B,C	A,B,D	A,C,D	B,C,D
α_U	.36	.29	.32	.35

Can we take advantage of the subjective analysis of each annotator?

- a **perspectivist** approach to disagreement in NLP [Basile, 2020; Basile et al., 2021]
- generating diverse consolidated corpora taking into account **subjective phenomena**

Research Goals

1. Study different techniques for **aggregating ADU annotations**

   We propose different strategies (**union**, **intersection** and **probabilistic**) for aggregating ADU annotations from different annotators.

2. Address the task of **argumentative density prediction** and **ranking**

   We study the impact of using different aggregation strategies on these tasks.

   We provide source code and BERT-based models: https://github.com/DARGMINTS/argument-density

Argument Density Prediction and Ranking

Argument Density

Text input: $\mathcal{T} = \langle t_1, ..., t_m \rangle$

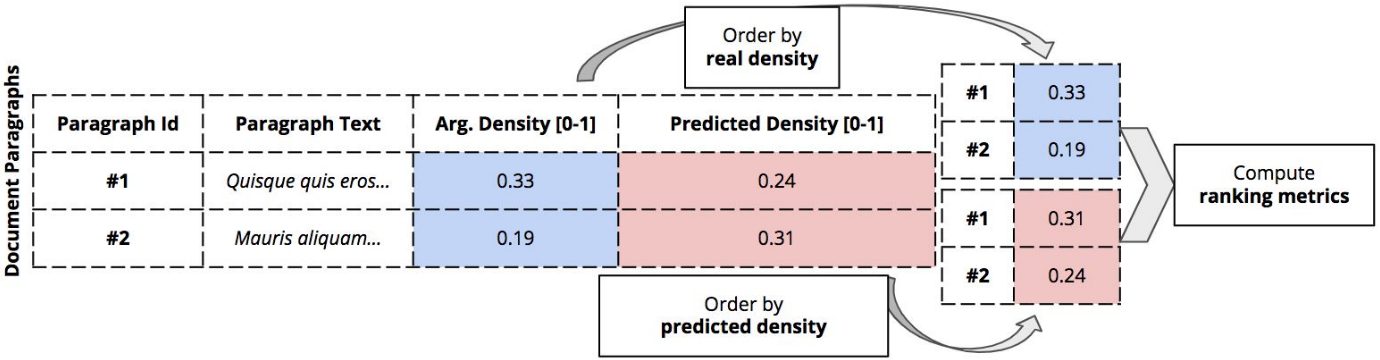
Argumentative content: $\mathcal{T}^\star = \langle t_1^\star, ..., t_n^\star \rangle$, such that $n \leq m, \forall i : t_i^\star \in \mathcal{T}$

Argument density (AD) is the proportion of argumentative tokens: $\rho = |\mathcal{T}^\star|/m$

AD Prediction: a **regression** task

- ***ρ* = 1**: all tokens in the input sequence are included in ADUs
- ***ρ* = 0**: none of the tokens in the input sequence are included in ADUs

Argument Density Prediction and Ranking



Workflow for both Argument Density Prediction and Ranking tasks.

Computing Argument Density from Multiple Annotations

Annotation aggregation strategies: **union** (***U***), **intersection** (***I***), **probabilistic** (***P***)

Let $\mathcal{T}^k$ be the set of tokens annotated by annotator $k \in K$

- **Union (*U*)**: set of tokens that were annotated by at least one annotator

$$\mathcal{U} = \langle t_i : \bigvee t_i \in \mathcal{T}^k, \forall k \in K, \forall i \in [1, m] \rangle \qquad \rho(\mathcal{U}) = |\mathcal{U}|/m$$

- **Intersection (*I*)**: set of tokens that were annotated by all annotators

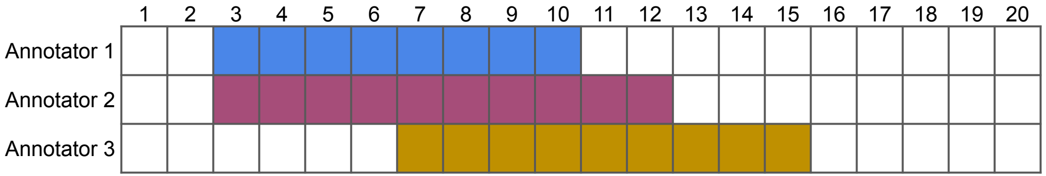
$$\mathcal{I} = \langle t_i : \bigwedge t_i \in \mathcal{T}^k, \forall k \in K, \forall i \in [1, m] \rangle \qquad \rho(\mathcal{I}) = |\mathcal{I}|/m$$

- **Probabilistic (*P*)**: set of tokens weighed by the ratio of annotators that have annotated them

$$\mathcal{P} = \langle w_i : w_i \in [0, 1], \forall i \in [1, m] \rangle \qquad \rho(\mathcal{P}) = (\textstyle\sum_i^m w_i)/m$$

Computing Argument Density from Multiple Annotations

Example annotations:

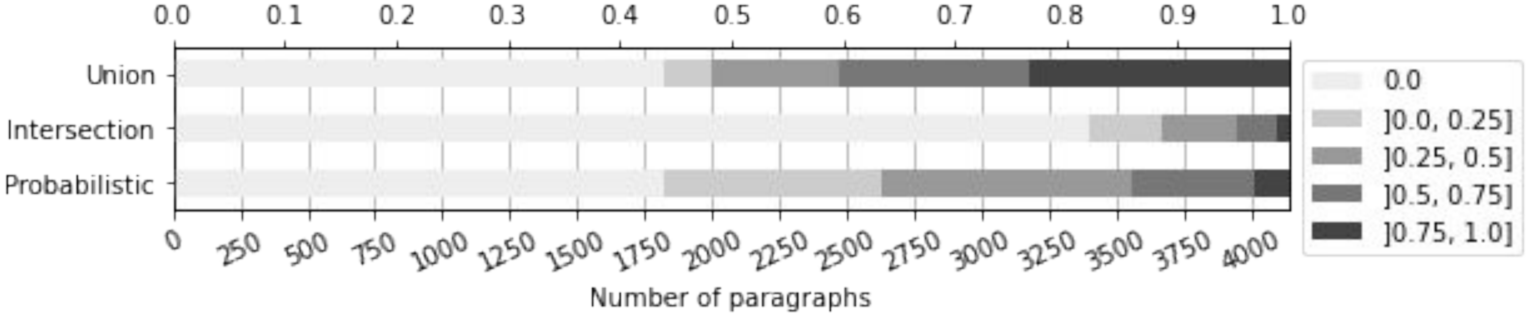


- **Union (*U*)**: [3-15]
  - ***ρ*(*U*)** = 13/20 = **0.65**

- **Intersection (*I*)**: [7-10]
  - ***ρ*(*I*)** = 4/20 = **0.2**

- **Probabilistic (*P*)**: [7-10]x3 + [3-6]x2 + [11-12]x2 + [13-15]x1
  - ***ρ*(*P*)** = (4x3 + 6x2 + 3x1)/3/20 = **0.45**

Paragraph-level Argument Density Distributions



Experiments and Results

Experimental Setup

Goal: compare **AD prediction** for different **aggregation strategies**

Data Preparation

- **15 generated datasets**:
  5 annotator combinations × 3 aggregation techniques
- **Density prediction** is made at the **paragraph-level**

	All	⟨A,B,C⟩	⟨A,B,D⟩	⟨A,C,D⟩	⟨B,C,D⟩
Union	dataset 1	...	...	...	...
Int.	...	...	...	...	...
Prob.	...	...	...	...	dataset 15

Setup

- 10-fold cross-validation with 8-1-1 train-validation-test splits (with similar mean density)
- Fine-tune **mBERT** for the regression task of **AD prediction** (loss = mean squared error)
- Baseline: dummy regressor (predict mean of the training set)

Density Prediction Results

	All			⟨A, B, C⟩			⟨A, B, D⟩			⟨A, C, D⟩			⟨B, C, D⟩		
	bl	*BERT*	*mean AD*	*bl*	*BERT*	*mean AD*	*bl*	*BERT*	*mean AD*	*bl*	*BERT*	*mean AD*	*bl*	*BERT*	*mean AD*
$\mathcal{U}$	.14	.09	.37	.12	.09	.31	.13	.09	.33	.12	.08	.29	.13	.09	.34
$\mathcal{I}$	.03	.02	.07	.04	.04	.10	.04	.03	.09	.03	.03	.08	.05	.04	.11
$\mathcal{P}$	.06	.03	.20	.06	.04	.20	.06	.04	.21	.06	.04	.18	.07	.04	.22

bl / ***BERT*** = MSE for baseline / mBERT; ***mean AD*** = mean argument density in the dataset

- Comparing **aggregation techniques**
  - **Higher MSE** for **Union** (despite improvements over baseline)
  - **Lowest MSE** for **Intersection** (but very sparse dataset, hence least improvements)
- Comparing **sets of annotators**
  - **All** yields results with **reduced MSE** for both **Intersection** and **Probabilistic**
  - MSE results do not seem to be aligned with IAA scores

	(A,B,C)	(A,B,D)	(A,C,D)	(B,C,D)	Mean
α_U	.43	.36	.38	.41	.39

Paragraph Ranking

Are the paragraphs with **highest predicted AD** the ones with **higher AD**?

- Normalized Discounted Cumulative Gain (**NDCG**)
  - sums the scores **ranked in the order induced by the predicted AD** (with logarithmic discount)
  - normalizes by the best possible score (**induced by the true AD**)

$$NDCG = \frac{DCG_p}{IDCG_p} = \frac{\sum_{i=1}^{p} \frac{rel_i}{\log_2(i+1)}}{\sum_{i=1}^{|REL_p|} \frac{rel_i}{\log_2(i+1)}}$$

- **Top-*k*** accuracy
  - computes how many of the *k* paragraphs with **higher AD** are among the *k* paragraphs with **higher predicted AD**

Paragraph Ranking (NDCG)

		All			$\langle A,B,C\rangle$			$\langle A,B,D\rangle$			$\langle A,C,D\rangle$			$\langle B,C,D\rangle$		
	k	*1*	*5*	*all*	*1*	*5*	*all*	*1*	*5*	*all*	*1*	*5*	*all*	*1*	*5*	*all*
$\mathcal{U}$	**baseline**	.45	.56	.75	.41	.54	.73	.42	.54	.73	.38	.52	.71	.43	.55	.74
	BERT	.74	.81	.89	.66	.76	.86	.72	.80	.88	.71	.78	.87	.70	.79	.87
$\mathcal{I}$	**baseline**	.16	.34	.48	.20	.39	.54	.19	.37	.52	.18	.36	.50	.21	.39	.55
	BERT	.44	.61	.66	.47	.65	.71	.49	.65	.71	.47	.64	.69	.50	.66	.73
$\mathcal{P}$	**baseline**	.37	.52	.71	.36	.51	.70	.36	.52	.70	.34	.49	.68	.37	.52	.70
	BERT	.71	.81	.88	.65	.77	.86	.70	.79	.87	.70	.79	.87	.69	.79	.87

- **Worse baseline results** for **Intersection** (in contrast with AD prediction)
  - The only strategy where **All** gets worse results than any annotator trio (for baseline and BERT)
- **Best values** are obtained for the **Union** and **Probabilistic** strategies
  - Avg improvement for BERT over baseline: 48% (Union), 90% (Intersection), 61% (Probabilistic)
- Results improve as *k* increases (as expected), with **best results** for **Union** and **Probabilistic**
- Again, no clear alignment with IAA scores

Paragraph Ranking (Top-*k*)

	k	All		(A, B, C)		(A, B, D)		(A, C, D)		(B, C, D)	
		1	5	1	5	1	5	1	5	1	5
$\mathcal{U}$	**baseline**	.09	.54	.10	.56	.10	.54	.09	.57	.10	.54
	BERT	.21	.69	.22	.67	.19	.70	.25	.70	.21	.69
$\mathcal{I}$	**baseline**	.21	.80	.16	.75	.18	.75	.19	.77	.15	.73
	BERT	.29	.59	.28	.63	.31	.62	.30	.61	.29	.63
$\mathcal{P}$	**baseline**	.10	.54	.13	.56	.09	.55	.10	.57	.10	.54
	BERT	.33	.72	.28	.70	.33	.71	.35	.72	.32	.70

- For **Intersection**, BERT is **unable to improve** accuracy for *k*=5
- Results improve as *k* increases (as expected)
  - *k*=1 setup is very challenging, most results below 33%
  - BERT obtains the **best scores** for *k*=5 with the **Probabilistic** strategy
- Again, no observed alignment with IAA scores

Conclusions

Conclusions

Argument annotation in opinion articles is a **demanding task**

Aggregation strategies are needed to **leverage different annotator biases**

- **Intersection** strategy is too demanding and negatively impacts downstream tasks
- **Probabilistic** strategy seems to be the most sensible approach (looking at both AD and ranking)
- Annotator selection did not bring any significant improvement and does not correlate with IAA

Argument Density Prediction and Ranking

- Simpler argument mining tasks useful for measuring the merits of annotation aggregation strategies
- BERT-based models shown to be **able to learn from multiple annotations**

Future Work

- Explore other **aggregation techniques** (e.g., Bayesian and vector-based)
- For argument density, train the models directly in the **ranking task**

Predicting Argument Density from Multiple Annotations

Thank you!

Questions?

NLDB 2022 June 15-17, 2022 Universitat Politècnica de Valencia, Spain

Cofinanciado por: