Multilingual Argument Mining: Datasets and Analysis
-
IBM Debater: XArgMining Dataset
-
Readme (October 2020)

1 Introduction

The growing interest in argument mining and computational argumentation brings with it a plethora of Natural Language Understanding(NLU) tasks and corresponding datasets. However, as with many other NLU tasks, the dominant language is English, with resources in other languages being few and far between.

This document describes the dataset release accompanying the work of Toledo-Ronen et al. (2020), which explored three argument mining tasks in a multilingual setting: (1) *stance classification*: given a topic and an argument that supports or contests the topic, determine the argument's stance towards the topic; (2) *evidence detection*: given a topic and a sentence, determine if the sentence is an evidence relevant to the topic; (3) *argument quality*: given a topic and a relevant argument, rate the argument so that higher-quality arguments are assigned a higher score.

The dataset contains:

- 30,497 English arguments annotated for their stance and quality (Gretz et al., 2020), along with their machine-translation to 5 languages. See Section 2 for details.
- 35,211 English Wikipedia sentences annotated for whether they are valid Evidence and their stance towards the discussed topic (Ein-Dor et al., 2020), along with their machine-translation to 5 languages. See Section 2 for details.
- 6,752 human-authored arguments in 5 languages, annotated for their stance and quality. See Section 3 for details.
- 210 English arguments and 200 English Wikipedia sentences, manually and automatically translated to German and Italian, with their corresponding labels collected for English and for the translated texts. See Section 4 for details.

2 Translated Data

2.1 Description

English Datasets The sources for our translated data are two existing argument mining datasets in English, collected by our colleagues as part of our work on Project Debater.[1] One is a corpus of 30,497 arguments on 71 controversial topics (referred herein as ArgsEN), annotated for their stance towards the topic and for their quality (Gretz et al., 2020). The second dataset is a corpus of 35,211 sentences from Wikipedia on 321 controversial topics (referred herein as EviEN), annotated for their stance towards the topic and the extent to which they can serve as an evidence for the topic (Ein-Dor et al., 2020). Example 1 shows an argument and an evidence for one topic.

Example 1 **(Argument and evidence)**
Topic: *We should legalize cannabis*
Argument: *Cannabis can provide relief for a number of ailments without side effects.*
Evidence: *In 1999, a study by the Division of Neuroscience and Behavioral Health found no evidence of a link between cannabis use and the subsequent abuse of other illicit drugs.*

Translation We used the Watson Language Translator[2] to translate the English data into 5 languages: German (DE), Dutch (NL), Spanish (ES), French (FR), and Italian (IT). The labels for the MT data were projected from the English data.

[1] Project Debater is the first AI system that can debate humans on complex topics: https://www.research.ibm.com/artificial-intelligence/project-debater/

[2] www.ibm.com/watson/services/language-translator/

2.2 Data Format

The `Machine Translations` folder contains the English datasets and their machine translations:

1. The `Arguments_6L_MT.csv` file contains 30,497 English arguments annotated for their stance and quality, along with their machine-translation to 5 non-English languages. The file format is described in Table 1.

2. The `Evidence_6L_MT.csv` file contains 35,211 English Wikipedia sentences annotated for whether they are valid Evidence and their stance towards the discussed topic, along with their machine-translation to 5 non-English languages. The file format is described in Table 2.

3 Human-Authored Data

3.1 Description

Arguments written in a non-English provide a more realistic evaluation set than translated texts, specifically for tasks where labels are not well-preserved across automatic translation (Toledo-Ronen et al., 2020). Therefore, we created a new multilingual evaluation set by collecting arguments in 5 languages (ES, FR, IT, DE, and NL) for the 15 topics in the ArgsEN test set, using the *Appen*[3] crowdsourcing platform. The human-authored evaluation dataset is herein referred to as ArgsHG.

Annotation Setup Initially, crowd contributors wrote up to two pairs of arguments per topic, with one argument supporting the topic and another contesting it in each pair. Next, the arguments were assessed for their stance and quality by 10 annotators (per-language). Given an argument, they were asked to determine the stance of the argument towards the topic and to assess whether it is of high quality. The full argument annotation guidelines are included in the Supplementary Materials, and Table 3 details the number of arguments collected and labeled per language. To set a common standard, annotators were instructed to mark about half of the arguments they labeled as high quality. Annotation quality was controlled by integrating test questions (TQs) with an a-priori known answer in between the regular questions, measuring the per-annotator accuracy on these questions, and excluding underperformers.

A per-annotator average agreement score was computed by considering all peers sharing at least 50 common answers, calculating pairwise Cohen's Kappa (Cohen, 1960) with each of them, and averaging. Those not having at least 5 peers meeting this criterion were excluded and their answers were discarded. Averaging the annotator agreements yields the average inter-annotator agreement (agreement-κ) of each question.

To derive a label (or score) for each question we use the WA-score of Gretz et al. (2020). Roughly, answers are aggregated with a weight proportional to the agreement score for the annotators who chose them. At least 5 answers were required for a question to be considered as labeled.

Scaling the annotation from English to new languages required some adjustments, such as restricting participation to countries in which the TL is commonly spoken, and the use of TQs for the argument quality question.

Results Table 3 presents the agreement-κ for all TLs and each task for the human-generated dataset. For stance, the agreement is comparable to previously reported values for English (0.69 by Toledo et al. (2019) and 0.83 for ArgsEN). For quality, the agreement is significantly better than previously reported on ArgsEN (0.12 by Gretz et al. (2020)), presumably due to the use of TQs in this task, which were not included before. The annotation in each of the non-English languages involved a distinct group of annotators, producing varying annotation quality among languages which is reflected in their agreement-κ values.

The results also include the percentage of arguments labeled as supporting arguments, computed separately for each annotator and averaged over all annotators. All values are close to 0.5, confirming that the collected arguments are balanced for stance, as instructed. Similarly, the results show the percentage of arguments labeled as high quality, averaged over all annotators, confirming that annotators mostly followed the instruction to label about half of the arguments as high quality.

3.2 Data Format

The `Human Authored Arguments` folder contains the human-generated arguments and their

[3] www.appen.com

Field	Description	Example
set	The partition of the data into `train`, `dev` and `test` sets in the release created by Gretz et al. (2020).	`train`
argument_EN	An English argument discussing the topic.	`A ban on naturopathy creates a cohesive front between scientists and the government that can combat the anti-science rhetoric of naturopathic industries.`
topic_EN	The topic (in English) discussed by the argument.	`We should ban naturopathy`
quality_score_EN	The argument's quality score (within the range $[0..1]$). A high value indicates an argument of high quality.	0.753
stance_label_EN	The argument's stance label (1 for supports the topic, -1 for contests the topic).	1
stance_conf_EN	The confidence in the stance label (within the range $[0..1]$). A higher value indicates a more confident label.	1
argument_ES	An automatic translation of the argument.	`Una prohibición de la naturopathy crea un frente cohesivo entre los científicos y el gobierno que puede combatir la retórica anticientífica de las industrias naturopaticas.`
topic_ES	An automatic translation of the topic.	`Deberíamos prohibir la naturopatía.`

Table 1: The format of the `Arguments_6L_MT.csv` file containing English arguments and their stance and quality labels, collected for the English texts by Gretz et al. (2020), along with machine-translations to 5 languages. The table shows the translation fields for Spanish. The fields for the other 4 languages are similar. See Section 2.2.

Field	Description	Example
set	The partition of the data into `train`, `dev` and `test` sets in the release created by Ein-Dor et al. (2020).	`train`
sentence_EN	A sentence from English Wikipedia.	`The Court declared forcefully that content-based restrictions on games are unconstitutional`
topic_EN	The topic (in English) discussed by the sentence.	`We should ban the sale of violent video games to minors`
evidence_label_EN	A binary label specifying whether the sentence is valid Evidence (1) or not (0).	1
evidence_conf_EN	The evidence score within the range $[0..1]$.	0.8
stance_label_EN	The sentence's stance label (1 for supports the topic, -1 for contests the topic, 0 for neither).	-1
sentence_ES	An automatic translation of the sentence.	`El Tribunal declaró a la fuerza que las restricciones basadas en el contenido de los juegos son inconstitucionales`
topic_ES	An automatic translation of the topic.	`Deberíamos prohibir la venta de videojuegos violentos a menores de edad`

Table 2: The format of the `Evidence_6L_MT.csv` file containing Wikipedia sentences annotated for whether they are valid Evidence and their stance towards the discussed topic (collected for the English texts by Ein-Dor et al. (2020)), along with machine-translations to 5 languages. The table shows the translation fields for Spanish. The fields for the other 4 languages are similar. See Section 2.2.

			Stance			***Quality***		
Language		**#C**	**#L**	κ	**Sup.**	**#L**	κ	**HQ**
Spanish	**ES**	2995	2995	0.73	0.51	828	0.29	0.62
French	**FR**	2201	1109	0.66	0.49	903	0.41	0.53
Italian	**IT**	3018	987	0.82	0.50	969	0.24	0.67
German	**DE**	1962	801	0.60	0.50	801	0.39	0.56
Dutch	**NL**	925	599	0.72	0.47	382	0.40	0.49

Table 3: Statistics of the ArgsHG multilingual arguments dataset, collected in five languages (See Section 3): the number of unique arguments collected (**#C**); the number of arguments labeled (**#L**) for their ***stance*** and ***quality***; the agreement-κ obtained for each task; the average percentage of arguments labeled by each annotator as supporting the topic (**Sup.**) and as high-quality (**HQ**).

Field	Description	Example
topic_EN	The topic (in English) discussed by the argument.	`Social media brings more harm than good`
topic_DE	An automatic translation of the topic.	`Social Media bringt mehr Schaden als gut`
argument_DE	The human-authored argument discussing the topic, in a non-English language.	`Es ist eine Möglichkeit, mit Menschen auf der ganzen Welt in Kontakt zu treten, sei es aus persönlichen oder geschäftlichen Gründen.`
stance_label	The argument's stance label (1 for supports the topic, -1 for contests the topic, 0 for neither).	-1
stance_conf	The confidence in the stance label (within the range $[0..1]$). A higher value indicates a more confident label.	0.923
stance_num_labelers	The number of annotators that annotated the stance of the argument towards the topic.	10
stance_pro	The number of annotators that annotated the argument as supporting the topic.	1
stance_neutral	The number of annotators that annotated the argument as not having a stance towards the topic.	0
stance_con	The number of annotators that annotated the argument as contesting the topic.	9
quality_score	The argument's quality score (within the range $[0..1]$). A high value indicates an argument of high quality.	1
quality_num_labelers	The number of annotators that annotated the argument for its quality.	8
quality_positive	The number of annotators that annotated the argument as high-quality.	8
quality_negative	The number of annotators that annotated the argument as low-quality.	0

Table 4: The format of the `human_authored_arguments_de.csv` file containing human-authored arguments in German and their stance and quality labels. The annotation was performed directly on the German text. The format of the files for other languages is similar. See Section 3.2

	German	Italian
Stance: Arguments	0.74	0.82
Stance: Evidence	0.82	0.76
Argument Quality	0.26	0.12
Evidence Detection	0.38	0.33

Table 5: The agreement-κ obtained in the annotations of the translated texts, for stance classification on arguments or evidence, argument quality and evidence detection.

corresponding labels in `csv` files. Each file includes arguments collected for one language. The file format is described in Table 4.

4 Labeled Translations

4.1 Description

An important prerequisite for training and evaluating models on automatically translated texts is that the labels of the original texts are preserved under translation, which depends on the specific task at hand. Example 2 shows one argument and its translation to Spanish and back to English. The translation preserves the original stance, but the argument quality is degraded. Hence, we annotated a sample of the translated texts to assess how often this happens in each task. The annotation focuses on one Romance and one West-Germanic language – Italian and German.

Example 2 (Translation quality)
***Topic**: We should ban algorithmic trading*
***English argument**: Algorithmic trading results in unfair advantages for those able to access it to the detriment of ordinary investors.*
***Back-translation**: The algorithmic trading of results in unjust advantages for those able to access it to the detriment of common investors.*

Annotation Setup 14 arguments were randomly sampled from each topic of the ArgsEN test set, yielding 210 arguments per language. Similarly, two sentences were sampled from each topic in the EviEN test set, producing 200 sentences per language. All texts were machine translated and human translated by native speakers of each TL. Both translations of each argument were labeled for their stance and quality, as in Section 3. Similarly, the potential evidence sentences were annotated for whether they are valid evidence, and those which are so were also annotated with their stance towards the topic, as in Ein-Dor et al. (2020). In this annotation, TQs were formed from translated texts, with the correct answer taken from the English labels. The full evidence annotation guidelines are included in the Supplementary Materials.

Results Table 5 shows the agreement-κ obtained for all tasks and the two languages. These values are on par with previously reported values for these tasks (as detailed in Section 3), though somewhat lower for evidence detection. For a detailed analysis of the correlation between the English and the translated labels, see Toledo-Ronen et al. (2020).

4.2 Data Format

The `Labeled Translations` folder contains the annotated translations and their corresponding labels in `csv` files.

The file format for the English arguments, their human translation, and the corresponding labels is exemplified in Table 6, for German (file `labeled_translated_arguments_de.csv`). Similarly, the automatic translation and its labels are in the same file in fields suffixed with `_mt`. A similar file is provided for Italian: `labeled_translated_arguments_it.csv`. The labels for the English texts are taken from (Gretz et al., 2020), and the labels for the translated texts are from Toledo-Ronen et al. (2020).

The file format for the evidence data is similarly described in Table 7. Here, the labels for the English sentences are taken from Ein-Dor et al. (2020), and the labels for the translated texts are from Toledo-Ronen et al. (2020).

References

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46.

Liat Ein-Dor, Eyal Shnarch, Lena Dankin, Alon Halfon, Benjamin Sznajder, Ariel Gera, Carlos Alzate, Martin Gleize, Leshem Choshen, Yufang Hou, Yonatan Bilu, Ranit Aharonov, and Noam Slonim. 2020. Corpus Wide Argument Mining – a Working Solution. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. 2020. A Large-scale Dataset for Argument

Field	Description	Example
topic_EN	The topic (in English) discussed by the argument.	`We should ban algorithmic trading`
topic_DE	An automatic translation of the topic.	`Wir sollten den algorithmischen Handel verbieten`
argument_EN	An English argument discussing the topic.	`a simple error in the algorithm ...`
stance_label_EN	The argument's stance label (1 for supports the topic, -1 for contests the topic).	1
stance_conf_EN	The confidence in the stance label (within the range $[0..1]$). A higher value indicates a more confident label.	0.893
quality_score_EN	The argument's quality score (within the range $[0..1]$). A high value indicates an argument of high quality.	0.854
argument_DE_ht	A human translation of the English argument to German.	`ein einfacher Fehler im Algorithmus ...`
num_labelers_ht	The number of annotators that annotated the human translation.	6
stance_label_ht	The human translation's stance label (1 for supports the topic, -1 for contests the topic, 0 for neither).	1
stance_conf_ht	The confidence in the stance label (within the range $[0..1]$). A higher value indicates a more confident label.	0.811
stance_pro_ht	The number of annotators that annotated the human translation as supporting the topic.	5
stance_neutral_ht	The number of annotators that annotated the human translation as not having a stance towards the topic.	0
stance_con_ht	The number of annotators that annotated the human translation as contesting the topic.	1
quality_score	The human translation's quality score (within the range $[0..1]$). A high value indicates an argument of high quality.	1
quality_positive	The number of annotators that annotated the human translation as high-quality.	8
quality_negative	The number of annotators that annotated the human translation as low-quality.	0

Table 6: The format of the `labeled_translated_arguments_de.csv` file containing English arguments and their human and automatic translations to German, along with the corresponding labels. The automatic translation and its fields are similarly suffixed with `_mt`. See Section 4.2

Field	Description	Example
topic_EN	The topic (in English) discussed by the sentence.	`We should increase the workweek`
topic_DE	An automatic translation of the topic.	`Wir sollten die Wochenarbeitszeit erhöhen`
sentence_EN	A sentence from English Wikipedia.	`Some governments create ...`
evidence_label_EN	A binary label specifying whether the sentence is valid Evidence (1) or not (0).	0
evidence_score_EN	The evidence score within the range $[0..1]$.	0
stance_label_EN	The sentence's stance label (1 for supports the topic, -1 for contests the topic, 0 for neither).	0
sentence_DE_ht	A human translation of the English sentence to German.	`Manche Regrierungen ...`
evidence_label_ht	A binary label specifying whether the human translation is valid Evidence (1) or not (0).	0
evidence_conf_ht	The confidence in the evidence label (within the range $[0..1]$). A higher value indicates a more confident label.	0.892
evidence_num_labelers_ht	The number of annotators that annotated the human translation.	10
evidence_accept_ht	The number of annotators that annotated the human translation as valid evidence.	2
evidence_reject_ht	The number of annotators that annotated the human translation as non-evidence.	8
stance_label_ht	The human translation's stance label (1 for supports the topic, -1 for contests the topic, 0 for neither).	
stance_conf_ht	The confidence in the stance label (within the range $[0..1]$). A higher value indicates a more confident label.	
stance_num_labelers_ht	The number of annotators that annotated the stance of the human translation towards the topic.	2
stance_pro_ht	The number of annotators that annotated the human translation as supporting the topic.	2
stance_con_ht	The number of annotators that annotated the human translation as contesting the topic.	0

Table 7: The format of the `labeled_translated_evidence_de.csv` file containing sentences from English Wikipedia, their human and automatic translations to German, along with the corresponding labels. The automatic translation and its fields are similarly suffixed with `_mt`. See Section 4.2

Quality Ranking: Construction and Analysis. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)*.

Assaf Toledo, Shai Gretz, Edo Cohen-Karlik, Roni Friedman, Elad Venezian, Dan Lahav, Michal Jacovi, Ranit Aharonov, and Noam Slonim. 2019. Automatic Argument Quality Assessment - New Datasets and Methods. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5625–5635, Hong Kong, China. Association for Computational Linguistics.

Orith Toledo-Ronen, Matan Orbach, Yonatan Bilu, Artem Spector, and Noam Slonim. 2020. Multilingual argument mining: Datasets and analysis. In *Findings of EMNLP*.