Multi-lingual Argumentative Corpora in English, Turkish, Greek, Albanian, Croatian, Serbian, Macedonian, Bulgarian, Romanian and Arabic

Alfred Sliwa, Yuan Ma, Ruishen Liu, Niravkumar Borad
Seyedeh Fatemeh Ziyaei, Mina Ghobadi, Firas Sabbah, Ahmet Aker
Department of Computer Science and Applied Cognitive Science
University of Duisburg-Essen
Lotharstraße 65, 47057 Duisburg
{alfred.sliwa.92, yuan.ma, ruishen.liu, niravkumar.borad
seyedeh.ziyaei, mina.ghobadi, firas.sabbah}@stud.uni-due.de
a.aker@is.inf.uni-due.de

Abstract

Argumentative corpora are costly to create and are available in only few languages with English dominating the area. In this paper we release the first publicly available corpora in all Balkan languages and Arabic. The corpora are obtained by using parallel corpora where the source language is English and target language is either a Balkan language or Arabic. We use 8 different argument mining classifiers trained for English, apply them all on the source language and project the decision made by the classifiers to the target language. We assess the performance of the classifiers on a manually annotated news corpus. Our results show when at least 3 to 6 classifiers are used to judge a piece of text as argumentative an F1-score above 90% is obtained.

Keywords: Multilingual Argument Annotations, Argument Mining, Parallel Corpora

1. Introduction

Argument mining refers to the automatic extraction of arguments from natural texts. An argument consists of a claim (also referred to as the conclusion of the argument) and several pieces of evidence called premises that support or reject the claim (Lippi and Torroni, 2016b). Identifying arguments in large volumes of textual data has the potential to revolutionarise our access to information. Argument based search for information would for example facilitate individual and organisational decision-making, make learning more efficient, enable quicker reporting on present and past events, to name just a few broad applications. Even more important is argument mining in the multi-lingual context, by which argument based retrieval would be available to people in the language of their preference.

Current studies report methods for argument mining in legal documents (Reed et al., 2008), persuasive essays (Nguyen and Litman, 2015), Wikipedia articles (Levy et al., 2014; Rinott et al., 2015), discussion fora (Swanson et al., 2015), political debates (Lippi and Torroni, 2016a) and news (Sardianos et al., 2015; Al-Khatib et al., 2016). In terms of methodology, supervised machine learning is a central technique used in all these studies. This assumes the availability of data sets – argumentative texts – to train and test the argument mining models. Such data sets are readily available in English and – although in comparably smaller quantities – in very few European languages such as German or Italian. Languages other than these are currently neglected. Due to this lack of data the research and development of argumentation mining outside English and few European languages is very limited, rendering multilingual argument mining and language independent argument based retrieval impossible.

In this research we aim to fill this gap. We aim to create multi-lingual corpora annotated with argumentative structures automatically. For this purpose we make use of parallel corpora containing multiple bilingual documents aligned at sentence level, i.e. every sentence in a document written in a source language such as English is translated into a target language such as Greek. As the sentences in the documents are parallel it infers that if one of the sentence pairs is argumentative so it is also the other sentence. This also means annotating for instance English sentences for arguments leads also argumentative annotations in the target languages. One way of annotating the English sentences would be through human annotators. However, this is very intensive and costly task, especially when the task is to annotate several thousands documents which is the case in our research. Another way is to rely on existing robust argumentation tools and perform the annotation automatically through argument projection – a task recently proposed by (Aker and Zhang, 2017). This is exactly what we do in this research. We use eight different robust argument annotation tools created for English Aker et al. (2017), apply these on the English documents and annotate every sentence whether it is argumentative or not. We then project the annotations to the target languages consisting of Balkan languages such as Turkish, Greek, Albanian, Croatian, Serbian, Macedonian, Bulgarian and Romanian. The corpora we annotate is the SETimes corpora aligned at sentence level by Tyers and Alperen (2010). Since these corpora are publicly available we also release our annotations for the public. In addition, we also gathered around 3543 parallel English and Arabic documents from Huffington Post and processed them similar to the Balkan languages. By request we will make both annotated corpora available. With these pieces of information it is feasible to download the articles and also track-back the annotations.

The availability of such rich argumentative corpora is the first step to close the gap between English and under-resourced languages in terms of argument mining and kick-off efforts in creating argument mining solutions for languages other than English. Furthermore, we also think that it will start opening research direction towards multi-

Language Pair	# Documents	# Sentences
BG-EN	22,531	161,436
EL-EN	23,210	166,430
HR-EN	21,062	158,963
MK-EN	22,865	154,570
RO-EN	22,992	172,573
SQ-EN	22,947	171,885
SR-EN	22,779	164,377
TR-EN	22,800	166,510
AR-EN	3,543	85,831

Table 1: Characteristics of Parallel Corpora. All numbers refer to English data.

lingual argument mining and retrieval.

2. Data: SETimes and Huffington Post

The SETimes is an open source parallel corpus of news articles in the Balkan languages such as Turkish, Greek, Albanian, Croatian, Serbian, Macedonian, Bulgarian and Romanian and English, originating from a multi-lingual news website (Tiedemann, 2009). All documents extracted from the news website[1] are translated to XML files for each language pair and aligned at sentence level (Tyers and Alperen, 2010).

In addition to the SETimes corpora we also collected 3543 Arabic-English parallel news articles from Huffington Post. Although the articles are parallel, i.e. translations of each other, the alignment information between the sentences is not given so that we implemented simple heuristics such as sentence position, sentence length and dictionary-based translation overlap to provide this information. Table 1 shows both corpora in numbers.

3. Method

3.1. Pipeline

Our pipeline of annotating the data described in Section 2. is shown in Figure 1 – the figure shows the pipeline on SETimes as example however, processing the English-Arabic corpus happens analog. The first step is about reading English sentences from a parallel corpus such as English-Greek. For each sentence we extract rich feature sets detailed in the next section and apply 8 different argument mining models to annotate the sentence as argumentative or non-argumentative. Finally, we write the answer back to the corpus. More precisely, we record the type (argumentative or not) determined based on the majority vote among 8 annotators (at least 5 annotators are required to make clear decision) and as well as the decisions of each annotator. If we have 4:4 decisions for each type the overall result depends on the confidence ranking of each voter.

3.2. Argument Mining Tool

We used 8 different argument mining models to annotate the English sentences Aker et al. (2017). Seven of the models make use of traditional machine learning methods such as SVM, decision trees, etc. The 8^{th} model applies Convolutional Neural Networks (CNN) to predict the type labels. The CNN model uses various word embeddings. The other 7 models rely on rich set of features grouped into structural, lexical, syntactical and contextual categories.

[1] www.setimes.com, however this website is not maintained anymore.

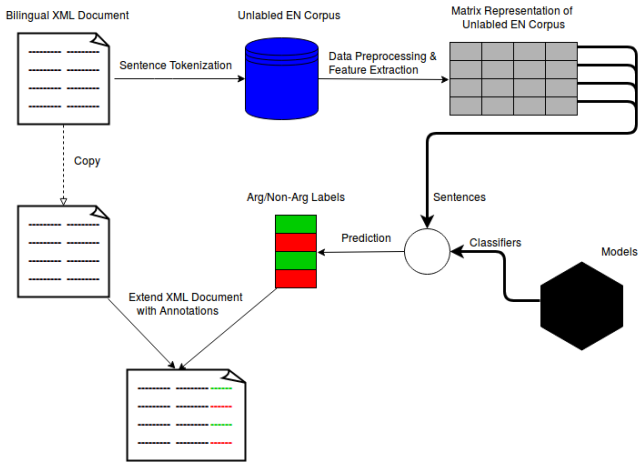

Figure 1: Data Processing Pipeline System

4. Annotation results

According to annotation results we list statistics about classification distribution for each analyzed parallel corpus. Table 2 shows the distribution of predicted argumentative and non-argumentative sentences for each bilingual data source. Note a sentence is regarded as argumentative when majority of the argument mining tools predicted that particular class. Otherwise the sentence is marked as non-argumentative. From the table we see that around 2/3 of the sentences are non-argumentative and around 1/3 are argumentative. This picture is repeated for each language pair.

5. Evaluation

In terms of evaluation we measure the aggregated performance of our eight argument mining models on a manually annotated news corpus (see Section 5.1.). In order to achieve this we use three distinct corpora to train and test the models. The corpora include persuasive essays (Nguyen and Litman, 2015), Wikipedia articles (Rinott et al., 2015) and news articles. At first we divide our data set into training set, validation set and held-out test set. All instances of essay, Wikipedia corpus and 80 news articles are used for training and hyper parameter tuning our models. We perform 4-fold cross-validation with 20 news articles held-out for validation purposes. The held-out test set contains 20 news articles and is used to determine the evaluation performance of the trained models. At this point we compute an aggregated prediction vector based on the individual model votes. By introducing a threshold k for the minimum number of argumentative votes, we can decide whether a particular test instance is assigned as argumentative or not. The threshold k takes eight different values from the set $\{1, 2, 3, 4, 5, 6, 7, 8\}$. In this way we report eight different evaluation results. The procedure of the evaluation experiment can be tracked in Figure 2.

Language Pair	# Argumentative Sentences	# Non-Argumentative Sentences
BG-EN	54,171	107,265
EL-EN	55,486	110,944
HR-EN	54,417	104,546
MK-EN	51,552	103,018
RO-EN	59,079	113,494
SQ-EN	58,750	113,135
SR-EN	56,232	108,145
TR-EN	56,833	109,677
AR-EN	29,602	56,229

Table 2: Classification Distribution of Argumentative/Non-Argumentative Sentences for Parallel Corpora

5.1. Annotated news corpus

In addition to the essay and Wikipedia corpora we also manually annotated 100 news articles from The Guardian newspaper, related to the general topic of "Brexit". On average the articles have a length of $\overline{x} = 75$ sentences (range of [24,186]). Each article is annotated for claims and premises by an expert. For the purpose of training and testing for the above classifiers we do not distinguish between claims and premises but treat both annotation types as argument. Any sentence in the news article not marked as claim or premise is regarded as non argumentative.

5.2. Evaluation results

In Table 3 we report the evaluation results for each threshold k. Each evaluation report contains statistics about precision, recall and F1-score values for argumentative class, non-argumentative class and average among both classes. It is observable that in case of low threshold values the precision score for Argument class is low but the recall score for the same class is high. Because of the fact that many test instances are classified as argumentative there are more non-argumentative instances incorrectly classified as argumentative. This leads to a high recall score but to a low precision score. In case of high threshold values the recall score for Argument class is low but the precision score for the same class is high. As only high probable test instances are classified as argumentative there are more argumentative instances incorrectly classified as non-argumentative. This yields in high precision score but low recall score.
We can see that by increasing k the precision also increases but the recall simultaneously decreases. According to average F1-score the optimal threshold values are 4 and 5 with a score of 0.96.

6. Conclusion

In this paper we described the issue with argument mining in languages other than English, namely the non-availability of argumentative training data. We motivated the idea of overcoming this disadvantage using parallel data and automatic argument annotation. We processed the SETimes corpora as well as an English-Arabic corpus that we collected from HuffingtonPost. Our processing includes the annotation of English sentences as argumentative or non-argumentative. We used 8 different argument mining models and make use of majority voting to mark the class labels for the sentences. Our annotations are freely available by request.

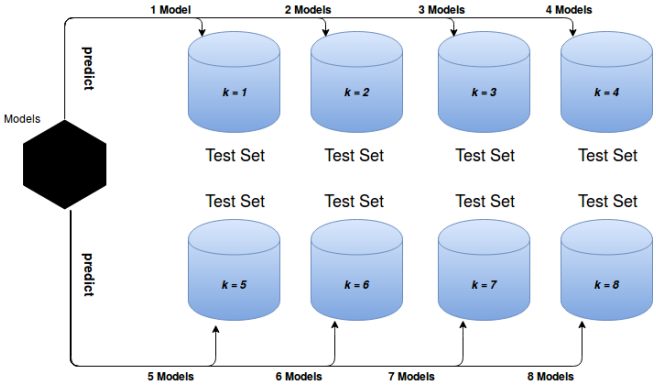

Figure 2: The Evaluation procedure containing eight replicated test sets where a various number of argument mining models k contribute to argumentative sentence prediction.

We also evaluated the performance of our classifiers on a manually annotated news corpus. Our results show that best F1-score is achieved when 3 to 6 classifiers are used to judge whether a piece of text is argumentative or not. In future we plan to use the map argumentative corpora to train argument mining systems in the respective languages.

7. Acknowledgements

This work was supported by the Deutsche Forschungsgemeinschaft (DFG) under grant No. GRK 2167, Research Training Group "User-Centred Social Media".

8. Bibliographical References

Aker, A. and Zhang, H. (2017). Projection of argumentative corpora from source to target languages. In *Proceedings of the 4th Workshop on Argument Mining*, pages 67–72.

Aker, A., Sliwa, A., Ma, Y., Lui, R., Borad, N., Ziyaei, S., and Ghobadi, M. (2017). What works and what does not: Classifier and feature analysis for argument mining. In *Proceedings of the 4th Workshop on Argument Mining*, pages 91–96.

Al-Khatib, K., Wachsmuth, H., Kiesel, J., Hagen, M., and Stein, B. (2016). A news editorial corpus for mining argumentation strategies. In *Proceedings of the 26th International Conference on Computational Linguistics*.

Levy, R., Bilu, Y., Hershcovich, D., Aharoni, E., and Slonim, N. (2014). Context dependent claim detection.

Threshold k	Class	Precision	Recall	F1-Score
1	Argument	**0.61**	**1.0**	0.76
	Non-Argument	1.0	0.08	0.14
	Average/Total	0.77	.0.62	0.51
2	Argument	**0.74**	**0.99**	0.85
	Non-Argument	0.98	0.51	0.67
	Average/Total	0.84	.0.79	0.78
3	Argument	**0.89**	**0.98**	0.94
	Non-Argument	0.97	0.83	0.9
	Average/Total	0.93	.0.92	0.92
4	Argument	**0.96**	**0.96**	0.96
	Non-Argument	0.95	0.95	0.95
	Average/Total	0.96	.0.96	0.96
5	Argument	**0.98**	**0.95**	0.96
	Non-Argument	0.93	0.97	0.95
	Average/Total	0.96	.0.96	0.96
6	Argument	**0.99**	**0.91**	0.95
	Non-Argument	0.89	0.99	0.94
	Average/Total	0.95	.0.95	0.95
7	Argument	**1.0**	**0.82**	0.9
	Non-Argument	0.79	1.0	0.88
	Average/Total	0.91	.0.89	0.89
8	Argument	**1.0**	**0.44**	0.61
	Non-Argument	0.55	1.0	0.71
	Average/Total	0.82	.0.67	0.65

Table 3: Evaluation Results of Argumentative Sentence Detection on the News Articles held-out test set for varying threshold values k.

Lippi, M. and Torroni, P. (2016a). Argument mining from speech: Detecting claims in political debates. In *AAAI*, pages 2979–2985.

Lippi, M. and Torroni, P. (2016b). Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology (TOIT)*, 16(2):10.

Nguyen, H. V. and Litman, D. J. (2015). Extracting argument and domain words for identifying argument components in texts. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 22–28.

Reed, C., Mochales Palau, R., Rowe, G., and Moens, M.-F. (2008). Language resources for studying argument. In *Proceedings of the 6th conference on language resources and evaluation-LREC 2008*, pages 91–100. ELRA.

Rinott, R., Dankin, L., Perez, C. A., Khapra, M. M., Aharoni, E., and Slonim, N. (2015). Show me your evidence-an automatic method for context dependent evidence detection. In *EMNLP*, pages 440–450.

Sardianos, C., Katakis, I. M., Petasis, G., and Karkaletsis, V. (2015). Argument extraction from news. *NAACL HLT 2015*, page 56.

Swanson, R., Ecker, B., and Walker, M. (2015). Argument mining: Extracting arguments from online dialogue. In *Proceedings of 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 2015)*, pages 217–227.

Tiedemann, J. (2009). News from OPUS - A collection of multilingual parallel corpora with tools and interfaces. In N. Nicolov, et al., editors, *Recent Advances in Natural Language Processing*, volume V, pages 237–248. John Benjamins, Amsterdam/Philadelphia, Borovets, Bulgaria.

Tyers, F. M. and Alperen, M. S. (2010). South-east european times: A parallel corpus of balkan languages. In *Proceedings of the LREC Workshop on Exploitation of Multilingual Resources and Tools for Central and (South-) Eastern European Languages*, pages 49–53.