Projection of Argumentative Corpora from Source to Target Languages

Ahmet Aker
University of Duisburg-Essen
a.aker@is.inf.uni-due.de

Huangpan ZHANG
University of Duisburg-Essen
huangpan.zhang@stud.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 Mandarin argumentative corpus. The corpus is created by exploiting the idea of comparable corpora from Statistical Machine Translation. We use existing corpora in English and manually map the claims and premises to comparable corpora in Mandarin. We also implement a simple solution to automate this approach with the view of creating argumentative corpora in other less-resourced languages. In this way we introduce a new task of multi-lingual argument mapping that can be evaluated using our English-Mandarin argumentative corpus. The preliminary results of our automatic argument mapper mirror the simplicity of our approach, but provide a baseline for further improvements.

1 Introduction

Identifying argument, i.e. claims and their associated pieces of evidence (premises) 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 search would be available to people in the language of their preference.

Argument mining is a new, but rapidly growing area of research within Computational Linguistic that has gained a great popularity in the last five years. For instance, since 2014 the meeting of the Association for Computational Linguistics (ACL) is hosting a workshop specifically dedicated to Argument Mining.[1] 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, 2016) 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. We are only aware of the study conducted by Chow (2016) who manually annotated Chinese news editorial paragraphs about whether they contain an argument or not. However, the boundaries of the arguments and their claims and premises were not annotated. Due to this lack of data the research and development of argumentation mining outside English and few European languages is very limited, rendering multi-lingual argument mining and language independent argument based search impossible.

In this research we aim to fill this gap. We aim to map existing argument annotations from a source language to a target language. For this purpose an ideal situation would be if there existed parallel documents where the source docu-

[1] http://argmining2016.arg.tech/

ments are annotated for arguments and where every sentence in the source document had a translation in the target document. In this case one could easily map any argumentative annotation from the source language to the target one. However, parallel data are sparse. In particular there exist no annotated argumentative corpora with parallel documents in any other language, expect the one described by Peldszus and Stede (2015) who report argumentative microtexts corpora in German that is also translated into English. Instead, inspired by the statistical machine translation (SMT) methods, we explore the idea of comparable corpora to obtain argumentative data sets. A comparable corpus contains pairs of documents written in two different languages. The document pairs usually share the same topic but the documents in a pair are not necessarily entirely translations of each other. However, they may share few sentences that are translation of each other. Related work has shown the usefulness of such corpora for training SMT system for under-resourced languages, cross-lingual information retrieval and assisted machine translation (Marton et al., 2009; Aker et al., 2013; Hashemi and Shakery, 2014; Kumano et al., 2007; Sharoff et al., 2006; Aker et al., 2012; Skadiņa et al., 2012; Munteanu and Marcu, 2005, 2002; Rapp, 1999). Given the difficulty and the cost of creating an argumentative corpus, extracting arguments from comparable corpora by automatically mapping arguments from the source language corpus to their translations in the target language seems an attractive avenue. In this work, we take a preliminary step to evaluating the viability of such an approach.

This paper reports on the first Mandarin argumentative corpus that is obtained using comparable corpora. We make use of the existing corpora, in which English documents are annotated for arguments, i.e. where sentences within the documents are marked as claims and premises. We manually map these English sentences to the target documents, by determining sentences in Mandarin that are translations of the English argumentative sentences. In addition, we report the results of our attempt to automatise this manual process of cross-lingual argument mapping. This data set will be publicly available for the research community.

Overall the paper contributes the following:

- We make available a first freely available argumentative corpus of Mandarin, also containing projected argumentative sentences from English to Mandarin comparable articles.
- We introduce a new task of creating multilingual argumentative corpora based on the idea of mapping argumentative sentences between articles that are comparable. Our manually generated data can be used to evaluate performance of automatic approaches.
- We establish and evaluate the possibility of obtaining argumentative corpora in any language with lower cost. To this end we propose a first baseline system for mapping English argumentative sentences into Mandarin.

2 Data

We work with the argumentative data published by Aharoni et al. (2014). The data contains the annotation of English Wikipedia articles for topic specific claims called Context Dependent Claims (CDCs) and premises referred as Context Dependent Evidence (CDE). A topic is a short phrase and frames the discussion within the article (Levy et al., 2014). A CDC is a general, concise statement that directly supports or contests the given topic (Levy et al., 2014). A CDE is a text segment that directly supports a claim in the context of the topic (Rinott et al., 2015). The data released in 2014 contains 1392 labeled claims for 33 different topics, and 1291 labeled premises for 350 distinct claims in 12 different topics (Aharoni et al., 2014). The average number of premises for each claim is 3.69.

To create the comparable corpora we used the inter-language links provided by Wikipedia to link the English articles to the articles in the target language Mandarin. The original data has 315 English articles of which have 160 corresponding Mandarin articles. These 160 pairs of English-Mandarin articles build the basis for mapping arguments from the English to Mandarin.

3 Manual mapping

In our manual process we first mapped Context Dependent Claims (CDCs) and then for each successfully mapped CDC its Context Dependent Evidence (CDEs). To do this we first automatically determined the sentences within the English Wikipedia articles that contained those CDCs and

Language	CDC	CDE
Only English	1392	1291
English-Mandarin	79	27

Table 1: Statistics about the CDCs and CDEs.

CDEs.[2] Next, we manually marked sentences that convey the same meaning as the English argumentative sentences. This process was performed by an annotator who is a native speaker of Mandarin and fluent in English.[3]

Table 1 summarizes the results of this process. In total 79 CDCs out of 1392 (5.7%) were mapped. These mappings were found in 34 English-Mandarin article pairs. The remaining 126 article pairs did not share any argumentative sentences. For the 79 CDCs we also analysed their premises (CDEs) and repeated the mapping process to determine corresponding Mandarin CDEs. In total we found 27 CDEs belonging to 18 CDCs. Table 2 shows an example CDC along with its CDEs in both languages.

Compared to the English the number of CDCs and CDEs mapped into Mandarin is substantially smaller. We have noted three major reasons for this data reduction:

- **No article to match an English one:** In this case there is no Mandarin article to match an English one. In most cases this is due to the topic of the article being very specific, so there are only limited language versions available. This is the reason why only 160 Mandarin articles could be identified for 315 English articles.

- **Dissimilar contents:** In this case there is a matched Mandarin article for the English one, but the contents of the article are not similar. This happens in articles which talk about topics whose content is country specific. Like Google China (`https://en.wikipedia.org/wiki/Google_China`) that talks about country specific events or government control wheres the corresponding English version does not contain any of Mandarin topics.

- **Missing sections:** The matched Mandarin article has missing sections. When the English claims are in those missing parts then there is no corresponding Mandarin mapping.

In a final step it was important to verify that the matched argument pairs are indeed comparable. This assessment was performed by three native Mandarin speakers fluent in English educated to post-graduate level. Their task was to indicate whether the sentences containing claims and premises in English translate into Mandarin claims and premises identified by our annotator. These assessors worked independently of each other. Only one argument was judged as not being an identified translation. The assessors agreed on all identified translation with our annotator and with each other, leading to the inter-annotator agreement of $kappa = 1$ based on Cohen's kappa.

4 Automatic mapping

As our manual effort indicates, there is a substantial reduction in data set, when comparable corpora are used to identify arguments that match in source and target languages. For this reason, an automatic approach to argument matching is mandatory in order to achieve larger data set sizes for multi-lingual argument mining approaches. In addition, successfull automatation of matching would open up the possibiity of creation argumentative corpora from any less-resourced language for which comparable corpora are available.

To evaluate the viability of an automatic approach and create a first benchmark we also performed a simple automatic mapping of English CDCs and CDEs into Mandarin. Our approach relies on automatic machine translation using MOSES (Koehn et al., 2007) and Google translate[4].

We trained MOSES using the publicly available parallel corpora from the HIT IR-lab[5]. For each English article we first translate all CDCs and CDEs into Mandarin. Next, we compare each of those translated argumentative pieces of text with every sentence from the corresponding Mandarin article. Our comparison is based on cosine similarity without stop-word removal. To perform tokenisation we used THULAC[6] an efficient Chi-

[2] In the data of (Aharoni et al., 2014) few CDCs and CDEs go over sentence boundaries. We ignore such cases and focus only on those that are bordered by a single sentence.

[3] Note that the annotation or mapping does not contain exact boundary information of the actual argument but only that the Mandarin sentence conveys similar meaning as its English counterpart.

[4] https://translate.google.com/

[5] `http://ir.hit.edu.cn/demo/ltp/Sharing_Plan.htm`

[6] `https://github.com/thunlp/THULAC`

	English	Mandarin
CDC	there was a connection between video games and violence	暴力事件与电子游戏之间有着必然的联系
CDE 1	Academic studies have attempted to find a connection between violent video games and the rate of violence and crimes from those that play them; some have stated a connection exists	不少学术研究试图找出一些人的犯罪行为同他们玩电子游戏行为之间的联系，有些研究表明这种联系是存在的
CDE 2	Incidents such as the Columbine High School massacre in 1999 have heightened concerns of a potential connection between video games and violent actions	如在1999年的科伦拜校园事件中，有人认为凶手的暴力行为与电子游戏之间就存在潜在的联系

Table 2: Example CDCs and CDEs.

nese lexical analyser.

We evaluate the performance using accuracy of our automatic mapping solution in retrieving correct pairs. For each CDC and CDE we check whether the most similar Mandarin sentence (according to cosine similarity) is also the correct pair. If yes, this is regarded as correct mapping, otherwise it is marked as wrong. Our evaluation results give us an accuracy of 24% for MOSES based translation and 49% for Google based translation. The Google based results are substantially better than those obtained through MOSES translation. This is because the MOSES decoder fails to translate many cases correctly.

5 Discussion

Our simple approach to tackling the automatic mapping of CDCs and CDEs achieves very low accuracy scores. Although the accuracy of the argument mapper based on the Google translation is substantially higher than the one achieved through MOSES translation, 49% of correct matches are still not satisfactory. This indicates that the task of argument matching in comparable corpora requires more sophisticated methods. One venue for improvement could be to extract richer features capturing sequential translations. Another direction for improvement could be towards a two phases approach. In the first phase one could reduce the Mandarin sentences by using an argumentative cue filter. In the second phase rich features could be extracted from the remaining candidates to perform the final pairing.

In terms of size the closest corpus to the one presented in this work is the one reported by Boltužić and Šnajder (2014) with 300 sentences. However, despite its small size at present, our corpus has important potential applications. Apart from training initial Mandarin argument mining solutions it can serve as a benchmark data for the task of mapping argumentative sentences from English to Mandarin. Systems performing with high precision on this data can be used to extend the given corpus by (1) determining annotated documents in the source language, (2) finding comparable documents in Mandarin and (3) using the mapping tool to map the source annotations to Mandarin.

6 Conclusion

In this paper we release the first Mandarin argumentative corpus containing Context Dependent Claims (CDCs) and Context Dependent Evidence (CDE). We obtained the corpus by manually mapping existing CDCs and CDEs from English Wikipedia articles to corresponding Mandarin articles. With this corpus we provide the basis for developing first argumentation mining solutions for Mandarin. The data can be downloaded from git-hub.[7]

By tackling the need for multi-lingual arguments in this paper we also introduced a new task: mapping argumentative sentences from one language to another. With this task we open up possibilities for obtaining argumentative resources in less-resourced languages with substantially lower cost than the manual effort.

Finally, we introduced a simple automatic tool for performing the argument mapping between English and Mandarin. The modest accuracy results achieved by this simple approach indicate that more sophisticated methods are necessary for argument mapping. We plan to improve the performance of our tool by investigating richer features and also the idea of filtering out sentences in the Mandarin languages that bare no argumentation. Our method reported in this work is a baseline system for argument mapping, and its scores can serve as a benchmark for further more sophisticated methods.

[7] https://github.com/ahmetaker/MandarinArguments

Acknowledgements

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

References

Ehud Aharoni, Anatoly Polnarov, Tamar Lavee, Daniel Hershcovich, Ran Levy, Ruty Rinott, Dan Gutfreund, and Noam Slonim. 2014. A benchmark dataset for automatic detection of claims and evidence in the context of controversial topics. In *Proceedings of the First Workshop on Argumentation Mining*. pages 64–68.

Ahmet Aker, Evangelos Kanoulas, and Robert J Gaizauskas. 2012. A light way to collect comparable corpora from the web. In *LREC*. Citeseer, pages 15–20.

Ahmet Aker, Monica Lestari Paramita, and Robert J Gaizauskas. 2013. Extracting bilingual terminologies from comparable corpora. In *ACL (1)*. pages 402–411.

Khalid Al-Khatib, Henning Wachsmuth, Johannes Kiesel, Matthias Hagen, and Benno Stein. 2016. A news editorial corpus for mining argumentation strategies. In *Proceedings of the 26th International Conference on Computational Linguistics*.

Filip Boltužić and Jan Šnajder. 2014. Back up your stance: Recognizing arguments in online discussions. In *Proceedings of the First Workshop on Argumentation Mining*. Citeseer, pages 49–58.

Marisa Chow. 2016. Argument identification in chinese editorials. In *Proceedings of NAACL-HLT*. pages 16–21.

Homa B Hashemi and Azadeh Shakery. 2014. Mining a persian–english comparable corpus for cross-language information retrieval. *Information Processing & Management* 50(2):384–398.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*. Association for Computational Linguistics, pages 177–180.

Tadashi Kumano, Hideki Tanaka, and Takenobu Tokunaga. 2007. Extracting phrasal alignments from comparable corpora by using joint probability smt model. *Proceedings of TMI* .

Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. 2014. Context dependent claim detection .

Marco Lippi and Paolo Torroni. 2016. Argument mining from speech: Detecting claims in political debates. In *AAAI*. pages 2979–2985.

Yuval Marton, Chris Callison-Burch, and Philip Resnik. 2009. Improved statistical machine translation using monolingually-derived paraphrases. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*. Association for Computational Linguistics, pages 381–390.

Dragos Stefan Munteanu and Daniel Marcu. 2002. Processing comparable corpora with bilingual suffix trees. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*. Association for Computational Linguistics, pages 289–295.

Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics* 31(4):477–504.

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

Andreas Peldszus and Manfred Stede. 2015. An annotated corpus of argumentative microtexts. In *Proceedings of the First Conference on Argumentation, Lisbon, Portugal, June. to appear*.

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated english and german corpora. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*. Association for Computational Linguistics, pages 519–526.

Chris Reed, Raquel Mochales Palau, Glenn Rowe, and Marie-Francine Moens. 2008. Language resources for studying argument. In *Proceedings of the 6th conference on language resources and evaluation-LREC 2008*. ELRA, pages 91–100.

Ruty Rinott, Lena Dankin, Carlos Alzate Perez, Mitesh M Khapra, Ehud Aharoni, and Noam Slonim. 2015. Show me your evidence-an automatic method for context dependent evidence detection. In *EMNLP*. pages 440–450.

Christos Sardianos, Ioannis Manousos Katakis, Georgios Petasis, and Vangelis Karkaletsis. 2015. Argument extraction from news. *NAACL HLT 2015* page 56.

Serge Sharoff, Bogdan Babych, and Anthony Hartley. 2006. Using comparable corpora to solve problems difficult for human translators. In *Proceedings of the COLING/ACL on Main conference poster sessions*. Association for Computational Linguistics, pages 739–746.

Inguna Skadiņa, Ahmet Aker, Nikos Mastropavlos, Fangzhong Su, Dan Tufis, Mateja Verlic, Andrejs Vasiļjevs, Bogdan Babych, Paul Clough, Robert Gaizauskas, et al. 2012. Collecting and using comparable corpora for statistical machine translation. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC), Istanbul, Turkey*.

Reid Swanson, Brian Ecker, and Marilyn Walker. 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.