Understanding and Detecting Supporting Arguments of Diverse Types

Xinyu Hua and **Lu Wang**
College of Computer and Information Science
Northeastern University
Boston, MA 02115
hua.x@husky.neu.edu luwang@ccs.neu.edu

Abstract

We investigate the problem of sentence-level *supporting argument detection* from relevant documents for user-specified claims. A dataset containing claims and associated citation articles is collected from online debate website idebate.org. We then manually label sentence-level supporting arguments from the documents along with their types as STUDY, FACTUAL, OPINION, or REASONING. We further characterize arguments of different types, and explore whether leveraging type information can facilitate the supporting arguments detection task. Experimental results show that LambdaMART (Burges, 2010) ranker that uses features informed by argument types yields better performance than the same ranker trained without type information.

1 Introduction

Argumentation plays a crucial role in persuasion and decision-making processes. An argument usually consists of a central claim (or conclusion) and several supporting premises. Constructing arguments of high quality would require the inclusion of diverse information, such as factual evidence and solid reasoning (Rieke et al., 1997; Park and Cardie, 2014). For instance, as shown in Figure 1, the editor on idebate.org – a Wikipedia-style website for gathering pro and con arguments on controversial issues, utilizes arguments based on study, factual evidence, and expert opinion to support the anti-gun claim "legally owned guns are frequently stolen and used by criminals". However, it would require substantial human effort to collect information from diverse resources to support argument construction. In order to facilitate this process, there is a pressing need for tools that can automatically detect supporting arguments.

To date, most of the argument mining research focuses on recognizing argumentative components and their structures from constructed arguments based on curated corpus (Mochales and Moens, 2011; Stab and Gurevych, 2014; Feng and Hirst, 2011; Habernal and Gurevych, 2015; Nguyen and Litman, 2016). Limited work has been done for retrieving supporting arguments from external resources. Initial effort by Rinott et al. (2015) investigates the detection of relevant factual evidence from Wikipedia articles. However, it is unclear whether their method can perform well on documents of different genres (e.g. news articles vs. blogs) for detecting distinct types of supporting information.

- A June 2013 IOM report states that "almost all guns used in criminal acts enter circulation via initial legal transaction". [study]
- Between 2005 and 2010, 1.4 million guns were stolen from US homes during property crimes (including bulglary and car theft), a yearly average of 232,400. [factual]
- Ian Ayres, JD, PhD, ... states, "with guns being a product that can be easily carried away and quickly sold at a relatively high fraction of the initial cost, the presence of more guns can actually serve as a stimulus to burglary and theft." [expert opinion]

Figure 1: Three different types of arguments used to support the claim "Legally owned guns are frequently stolen and used by criminals".

In this work, *we present a novel study on the task of sentence-level supporting argument detection from relevant documents for a user-specified claim*. Take Figure 2 as an example: assume we are given a claim on the topic of "banning cosmetic surgery" and a relevant article (cited for argument construction), we aim to automatically pinpoint the sentence(s) (in *italics*) among all sentences in the cited article that can be used to back up the claim. We define such tasks as *supporting argument detection*. Furthermore, another goal of

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 203–208
Vancouver, Canada, July 30 - August 4, 2017.
https://doi.org/10.18653/v1/P17-2032

- **Topic**: This house would ban cosmetic surgery
- **Claim**: An outright ban would be easier than the partial bans that have been enacted in some places.
- **Human Constructed Argument**: . . .This potentially leaves difficulty drawing the line for what is allowed.[1] . . .

Citation Article

[1]: *"Australian State Ban Cosmetic Surgery for Teens"*

- . . . It is unfortunate that a parent would consider letting a 16-year-old daughter have a breast augmentation."

- But others worry that similar legislation, if it ever comes to pass in the United States, would draw a largely arbitrary line – and could needlessly restrict some teens from procedures that would help their self-esteem.

- Dr. Malcolm Z. Roth, director of plastic surgery at Maimondes Medical Center in Brooklyn, N.Y. , said he believes that some teens are intelligent and mature enough to comprehend the risks and benefits of cosmetic surgery.. . .

Figure 2: A typical debate motion consists of a Topic, Claims, and Human Constructed Arguments. Citation article is marked at the end of sentence. Our goal is to find out supporting argument (in *italics*) from citation article that can back up the given claim.

this work is to understand and characterize different types of supporting arguments. Indeed, human editors do use different types of information to promote persuasiveness as we will show in Section 3. Prediction performance also varies among different types of supporting arguments.

Given that none of the existing datasets is suitable for our study, we collect and annotate a corpus from Idebate, which contains hundreds of debate topics and corresponding claims.[1] As is shown in Figure 2, each claim is supported with some human constructed argument, with cited articles marked on sentence level. After careful inspection on the supporting arguments, we propose to label them as STUDY, FACTUAL, OPINION, or REASONING. Substantial inter-annotator agreement rate is achieved for both supporting argument labeling (with Cohen's κ of 0.8) and argument type annotation, on 200 topics with 621 reference articles.

Based on the new corpus, we first carry out a study on characterizing arguments of different types via type prediction. We find that arguments of STUDY and FACTUAL tend to use more concrete words, while arguments of OPINION contain more named entities of person names. We then investigate whether argument type can be leveraged to assist supporting argument detection. Experimental results based on LambdaMART (Burges, 2010) show that utilizing features composite with argument types achieves a Mean Reciprocal Rank (MRR) score of 57.65, which outperforms an unsupervised baseline and the same ranker trained without type information. Feature analysis also demonstrates that salient features have significantly different distribution over different argument types.

For the rest of the paper, we summarize related work in Section 2. The data collection and annotation process is described in Section 3, which is followed by argument type study (Section 4). Experiment on supporting argument detection is presented in Section 5. We finally conclude in Section 6.

2 Related Work

Our work is in line with argumentation mining, which has recently attracted significant research interest. Existing work focuses on argument extraction from news articles, legal documents, or online comments without given user-specified claim (Moens et al., 2007; Palau and Moens, 2009; Mochales and Moens, 2011; Park and Cardie, 2014). Argument scheme classification is also widely studied (Biran and Rambow, 2011; Feng and Hirst, 2011; Rooney et al., 2012; Stab and Gurevych, 2014; Al Khatib et al., 2016), which emphasizes on distinguishing different types of arguments. To the best of our knowledge, none of them studies the interaction between types of arguments and their usage to support a user-specified claim. This is the gap we aim to fill.

3 Data and Annotation

We rely on data from `idebate.org`, where human editors construct paragraphs of arguments, either supporting or opposing claims under controversial topics. We also extract textual citation articles as source of information used by editors during argument construction. In total we collected 383 unique debates, out of which 200 debates are randomly selected for study. After removing invalid ones, our final dataset includes 450 claims

[1]The labeled dataset along with the annotation guideline will be released at `xyhua.me`.

STUDY: Results and discoveries, usually quantitative, as a result of some research investment.
FACTUAL: Description of some occurred events or facts, or chapters in law or declaration.
OPINION: Quotes from some person or group, either direct or indirect. It usually contains subjective, judgemental and evaluative languages, and might reflect the position or stance of some entity.
REASONING: Logical structures. It usually can be further broken down into causal or conditional substructures.

Table 1: Annotation scheme for our dataset. Due to space limit, we do not show detailed explanations and examples.

and 621 citation articles with about 53,000 sentences.

Annotation Process. As shown in Figure 2, we first annotate which sentence(s) from a citation articles is used by the editor as supporting arguments. Then we annotate the type for each of them as STUDY, FACTUAL, OPINION, or REASONING, based on the scheme in Table 1.[2] For instance, the highlighted supporting argument in Figure 2 is labeled as REASONING.

Two experienced annotators were hired to identify supporting arguments by reading through the whole cited article and locating the sentences that best match the reference human constructed argument. This task is rather complicated since human do not just repeat or directly quote the original sentences from citation articles, they also paraphrase, summarize, and generalize. For instance, the original sentence is "The global counterfeit drug trade, a billion-dollar industry, is thriving in Africa", which is paraphrased to "This is exploited by the billion dollar global counterfeit drug trade" in human constructed argument.

The annotators were asked to annotate independently, then discuss and resolve disagreements and give feedback about current scheme. We compute inter-annotator agreement based on Cohen's κ for both supporting arguments labeling and argument type annotation. For supporting arguments we have a high degree of consensus, with Cohen's κ ranges from 0.76 to 0.83 in all rounds and 0.80 overall. For argument type annotation, we achieve Cohen's κ of 0.61 for STUDY, 0.75 for FACTUAL, 0.71 for OPINION, and 0.29 for REASONING[3]

[2]We end up with the four-type scheme as a trade-off between complexity and its coverage of the arguments.

[3]Many times annotators have different interpretation on REASONING, and frequently label it as OPINION. This results in a low agreement for REASONING.

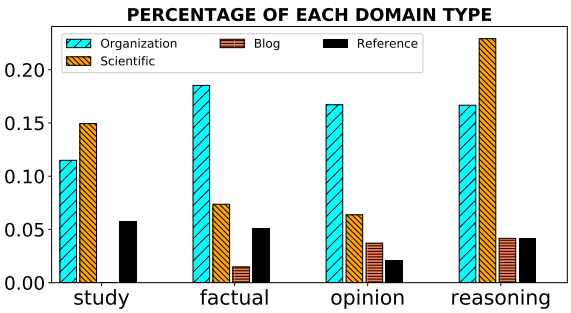

Figure 3: For each supporting argument type, from left to right shows the percentage of domain names of organizations, scientific, blog, and reference. We do not display statistics for news, because news articles take the same portion in all types (about 50%).

Statistics. In total 995 sentences are identified as supporting arguments. Among those, 95 (9.55%) are labeled as STUDY, 497 (49.95%) as FACTUAL, 363 (36.48%) as OPINION, and 40 (4.02%) as REASONING.

We further analyze the source of the supporting arguments. Domain names of the citation articles are collected based on their URL, and then categorized into "news", "organization", "scientific", "blog", "reference", and others, according to a taxonomy provided by Alexa[4] with a few edits to fit our dataset. News articles are the major source for all types, which account for roughly 50% for each. We show the distribution of other four types in Figure 3. Arguments of STUDY and REASONING are mostly from "scientific" websites (14.9% and 22.9%), whereas "organization" websites contribute a large portion of arguments of FACTUAL (18.5%) and OPINION (16.7%).

4 A Study On Argument Type Prediction

Here we characterize arguments of different types based on diverse features under the task of predicting argument types. Supporting arguments identified from previous section are utilized for experiments. We also leverage the learned classifier in this section to label the sentences that are not supporting arguments, which will be used for supporting argument detection in the next section. Four major types of features are considered.

Basic Features. We calculate frequencies of unigram and bigram words, number of four major types of part-of-speech tags (verb, noun, adjective, and adverb), number of dependency relations, and

[4]http://www.alexa.com/topsites/category

	Acc	F1
Majority class	0.520	0.171
Random	0.240	0.199
Log-linear (ngrams)	0.535	0.277
Log-linear (all features)	**0.622**	**0.436**

Table 2: Results for argument type prediction. One-vs-rest classifiers are learned for Log-linear models.

number of seven types of named entities (Chinchor and Robinson, 1997).

Sentiment Features. We also compute number of positive, negative and neutral words in MPQA lexicon (Wilson et al., 2005), and number of words from a subset of semantic categories from General Inquirer (Stone et al., 1966).[5]

Discourse Features. We use the number of discourse connectives from the top two levels of Penn Discourse Tree Bank (Prasad et al., 2007).

Style Features. We measure word attributes for their concreteness (perceptible vs. conceptual), valence (or pleasantness), arousal (or intensity of emotion), and dominance (or degree of control) based on the lexicons collected by Brysbaert et al. (2014) and Warriner et al. (2013).

We utilize Log-linear model for argument type prediction with one-vs-rest setup. Three baselines are considered: (1) random guess, (2) majority class, and (3) unigrams and bigrams as features for Log-linear model. Identified supporting arguments are used for experiments, and divided into training set (50%), validation set (25%) and test set (25%). From Table 2, we can see that Log-linear model trained with all features outperforms the ones trained with ngram features. To further characterize arguments of different types, we display sample features with significant different values in Figure 4. As can be seen, arguments of STUDY and FACTUAL tend to contain more concrete words and named entities. Arguments of OPINION mention more person names, which implies that expert opinions are commonly quoted.

5 Supporting Argument Detection

We cast the sentence-level supporting argument detection problem as a ranking task.[6] Features

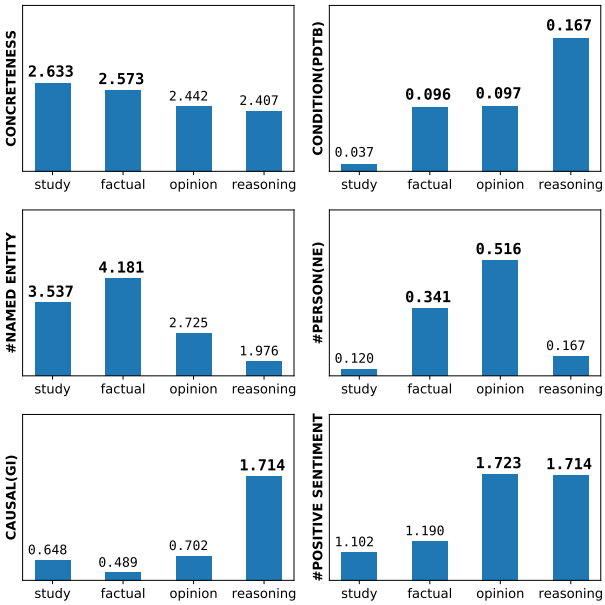

Figure 4: Average features values for different argument types. Numbers in **boldface** are significantly higher than the others based on paired t-test ($p < 0.05$).

in Section 4 are also utilized here as "Sentence features" with additional features considering the sentence position in the article. We further employ features that measure similarity between claims and sentences, and the composite features that leverage argument type information.

Similarity Features. We compute similarity between claim and candidate sentence based on TF-IDF and average word embeddings. We also consider ROUGE (Lin, 2004), a recall oriented metric for summarization evaluation. In particular, ROUGE-L, a variation based on longest common subsequence, is computed by treating claim as reference and each candidate sentence as sample summary. In similar manner we use BLEU (Papineni et al., 2002), a precision oriented metric.

Composite Features. We adopt composite features to study the interaction of other features with type of the sentence. Given claim c and sentence s with any feature mentioned above, a composite feature function $\phi_{M\,\texttt{(type, feature)}}(s, c)$ is set to the actual feature value if and only if the argument type matches. For instance, if the ROUGE-L score is 0.2, and s is of type STUDY, then $\phi_{M\,\texttt{(study, ROUGE)}}(s, c) = 0.2$ $\phi_{M\,\texttt{(factual, ROUGE)}}(s, c)$, $\phi_{M\,\texttt{(opinion, ROUGE)}}(s, c)$, $\phi_{M\,\texttt{(reasoning, ROUGE)}}(s, c)$ are all set to 0.

[5] Categories used: Strong, Weak, Virtue, Vice, Ovrst (Overstated), Undrst (Understated), Academ (Academic), Doctrin (Doctrine), Econ (Economic), Relig (Religious), Causal, Ought, and Perceiv (Perception).

[6] Many sentences in the citation article is relevant to the topic to various degrees. We focus on detecting the most relevant ones, and thus treat it as a ranking problem instead of a binary classification task.

Feature set	MRR	NDCG
Baselines		
TFIDF similarity	45.48	56.48
W2V similarity	47.65	59.00
Ngrams	27.26	43.83
Separate feature sets		
Sentence (Sen)	55.38*	65.09*
Similarity (Simi)	43.13	55.16
Comp(type, Sen) + Comp(type, Simi)	55.75*	64.91*
Additive Feature Test		
Sen + Ngrams + Simi	56.43*	65.79*
+ Comp(type, Sen) + Comp(type, Simi)	**57.65***	**66.51***
+ Comp(type, Claim)	56.58*	65.68*

Table 3: Supporting argument detection results. Comp(type, Sen) stands for composite features of argument type and sentence features, similarly for Comp(type,Simi). Comp(type,Claim) represents composite features of type and claim features. Results that are statistically significantly better than all three baselines are marked with $*$ (paired t-test, $p < 0.05$).

We choose LambdaMART (Burges, 2010) for experiments, which is shown to be successful for many text ranking problems (Chapelle and Chang, 2011). Our model is evaluated by Mean Reciprocal Rank (MRR) and Normalized Discounted Cumulative Gain (NDCG) using 5-fold cross validation. We compare to TFIDF and Word embedding similarity baselines, and LambdaMART trained with ngrams (unigrams and bigrams).

Results in Table 3 show that using composite features with argument type information (Comp(type, Sen) + Comp(type, Simi)) can improve the ranking performance. Specifically, the best performance is achieved by adding composite features to sentence features, similarity features, and ngram features. As can be seen, supervised methods outperform unsupervised baseline methods. And similarity features have similar performance as those baselines. The best performance is achieved by combination of sentence features, N-grams, similarity, and two composite types, which is boldfaced. Feature sets that significantly outperform all three baselines are marked with $*$.

For feature analysis, we conduct t-test for individual feature values between supporting arguments and the others. We breakdown features according to their argument types and show top salient composite features in Table 4. For all sentences of type STUDY, relevant ones tend to contain more "percentage" and more concrete words. We also notice those sentences with more hedging words are more likely to be considered. For sentences of FACTUAL, position of sentence in article

Feature	STUDY	FACTUAL	OPINION	REASONING
# PERC, NE	** ↑↑↑↑	–	–	–
# LOC, NE	–	** ↑↑	–	** ↑
position of sentence	** ↓↓	* * ** ↓↓	–	**** ↓↓↓↓
concreteness of sentence	* * * ↑↑	–	** ↑↑	* * * ↓
arousal of sentence	* * * ↑↑	–	** ↑↑	** ↓
# hedging word	** ↑↑↑	–	–	–
ROUGE	* * *↑↑	* * * ↑	** ↑↑	–
concreteness of claim	* * * ↑↑	–	** ↑↑	* * * ↓
arousal of claim	* * * ↑↑	–	** ↑↑	* * * ↓

Table 4: Comparison of feature significance under composition with different types. The number of $*$ stands for the p-value based on t-test between supporting argument sentences and the others after Bonferroni correction. From one $*$ to four, the p-value scales as: 0.05, 1e-3, 1e-5, and 1e-10. When mean value of supporting argument sentences is larger, ↑ is used; otherwise, ↓ is displayed. Number of arrows represents the ratio of the larger value over smaller one. "-" indicates no significant difference.

plays an important role, as well as their similarity to the claim based on ROUGE scores. For type OPINION, unlike all other types, position of sentence seems to be insignificant. As we could imagine, opinionated information might scatter around the whole documents. For sentences of REASONING, the ones that can be used as supporting arguments tend to be less concrete and less emotional, as opposed to opinion.

6 Conclusion

We presented a novel study on the task of sentence-level supporting argument detection from relevant documents for a user-specified claim. Based on our newly-collected dataset, we characterized arguments of different types with a rich feature set. We also showed that leveraging argument type information can further improve the performance of supporting argument detection.

Acknowledgments

This work was supported in part by National Science Foundation Grant IIS-1566382 and a GPU gift from Nvidia. We thank Kechen Qin for his help on data collection. We also appreciate the valuable suggestions on various aspects of this work from three anonymous reviewers.

References

Khalid Al Khatib, Henning Wachsmuth, Johannes Kiesel, Matthias Hagen, and Benno Stein. 2016. A news editorial corpus for mining argumentation strategies. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. Osaka, Japan, pages 3433–3443.

Or Biran and Owen Rambow. 2011. Identifying justifications in written dialogs by classifying text as argumentative. *International Journal of Semantic Computing* 5(04):363–381.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods* 46(3):904–911.

Christopher JC Burges. 2010. From ranknet to lambdarank to lambdamart: An overview. *Learning* 11(23-581):81.

Olivier Chapelle and Yi Chang. 2011. Yahoo! learning to rank challenge overview. In *Yahoo! Learning to Rank Challenge*. pages 1–24.

Nancy Chinchor and Patricia Robinson. 1997. Muc-7 named entity task definition. In *Proceedings of the 7th Conference on Message Understanding*. volume 29.

Vanessa Wei Feng and Graeme Hirst. 2011. Classifying arguments by scheme. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, pages 987–996.

Ivan Habernal and Iryna Gurevych. 2015. Exploiting debate portals for semi-supervised argumentation mining in user-generated web discourse. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, pages 2127–2137.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*. Barcelona, Spain, volume 8.

Raquel Mochales and Marie-Francine Moens. 2011. Argumentation mining. *Artificial Intelligence and Law* 19(1):1–22.

Marie-Francine Moens, Erik Boiy, Raquel Mochales Palau, and Chris Reed. 2007. Automatic detection of arguments in legal texts. In *Proceedings of the 11th international conference on Artificial intelligence and law*. ACM, pages 225–230.

Huy Nguyen and Diane Litman. 2016. Context-aware argumentative relation mining. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 1127–1137.

Raquel Mochales Palau and Marie-Francine Moens. 2009. Argumentation mining: the detection, classification and structure of arguments in text. In *Proceedings of the 12th international conference on artificial intelligence and law*. ACM, pages 98–107.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, pages 311–318.

Joonsuk Park and Claire Cardie. 2014. Identifying appropriate support for propositions in online user comments. In *Proceedings of the First Workshop on Argumentation Mining*. pages 29–38.

Rashmi Prasad, Eleni Miltsakaki, Nikhil Dinesh, Alan Lee, Aravind Joshi, Livio Robaldo, and Bonnie L Webber. 2007. The penn discourse treebank 2.0 annotation manual .

Richard D Rieke, Malcolm Osgood Sillars, and Tarla Rai Peterson. 1997. *Argumentation and critical decision making*. New York: Longman.

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.

Niall Rooney, Hui Wang, and Fiona Browne. 2012. Applying kernel methods to argumentation mining. In *Twenty-Fifth International FLAIRS Conference*.

Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *EMNLP*. pages 46–56.

Philip J Stone, Dexter C Dunphy, and Marshall S Smith. 1966. The general inquirer: A computer approach to content analysis. .

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods* 45(4):1191–1207.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*. Association for Computational Linguistics, pages 347–354.