Argument Identification in Chinese Editorials

Marisa Chow
Princeton University
1482 Frist Campus Ctr
Princeton, NJ 08544, USA
mlchow@princeton.edu

Abstract

In this paper, we develop and evaluate several techniques for identifying argumentative paragraphs in Chinese editorials. We first use three methods of evaluation to score a paragraph's argumentative nature: a relative word frequency approach; a method which targets known argumentative words in our corpus; and a combined approach which uses elements from the previous two. Then, we determine the best score thresholds for separating argumentative and non-argumentative paragraphs. The results of our experimentation show that our relative word frequency approach provides a reliable way to identify argumentative paragraphs with a F_1 score of 0.91, though challenges in accurate scoring invite improvement through context-aware means.

1 Introduction

Argumentation – the act of reasoning in support of an opinion or idea – frequently presents itself in all types of texts, from casual chat messages to online blogs. Argumentation mining aims to identify and determine a persuasive text's argumentative components, or the atomic units of its underlying structure. For example, an argumentation mining system might seek to locate and classify sections of claims and supporting evidence within an essay. More comprehensive mining might map the relations between different units, such as the support of evidence or the opposition of counterarguments to the thesis.

Argument identification offers a wide variety of practical applications. If argumentative text can be identified accurately, then the main arguments of large sets of data may be extracted. For example, argument identification could isolate arguments surrounding subjects like U.S. immigration law, or summarize the arguments in research papers. Recent efforts in argumentation mining have included applications such as automatic essay scoring (Song et al., 2014; Ong et al., 2014), online debates (Boltuzic and Šnajder, 2014), and arguments in specific domains such as online Greek social media sites (Sardianos et al., 2015). However, to our knowledge, no work in argumentation mining to date has been performed for Chinese, a large and rich domain for NLP work.

Here we focus on the first step of argumentation mining, locating argumentative units within a text. We develop and evaluate several methods of argument identification when performed upon a corpus of Chinese editorials, making the assumption that editorials are opinionated texts, although a single editorial may contain both opinionated and non-opinionated paragraphs.

Our work met with several challenges. Although newspaper editorials can be assumed to carry an opinion of some sort, the opinion is not always explicitly expressed at a word level, and methods of argumentation can vary widely from editorial to editorial. For example, one might exhibit a thesis followed by supporting evidence, but others might only state facts until the final paragraph. Furthermore, editorials commonly build arguments by citing facts. In our work, we not only had to define 'argumentative' and 'non-argumentative', but also limit the scope of an argument. In order to capture the larger argument structure, our work focuses on identifying arguments in paragraph units of no more than 200

Proceedings of NAACL-HLT 2016, pages 16–21,
San Diego, California, June 12-17, 2016.

characters (around 3-5 sentences), although we do not concatenate shorter paragraphs to ensure a minimum size.

Our task aims to label paragraphs such as the following as argumentative: "不幸的是，如今在深圳的各十字路口，绳子在显示作用，白线却无力地趴在地上。这是法规的悲哀。" ("Unfortunately, nowadays at Shenzhen's ten road intersections, cords are used to show that the white road lines lie uselessly on the ground. This legislation is tragic.")

The contributions of this paper are collecting and annotating a dataset of Chinese editorials; manually creating a list of argumentative words; and the comparison and analysis of three methods.

2 Data

2.1 Corpora

This work makes use of two corpora, one of Chinese editorials and one of Chinese reportage, both of which are subcorpora in the Lancaster Corpus of Mandarin Chinese (McEnery and Xiao, 2004). The LCMC, a member of the Brown family corpora, is a 1M balanced word corpus with seventeen subcorpora of various topics. We used the *Press: Reportage* and *Press: Editorials* subcorpora, containing 53K and 88K words respectively. Samples in both subcorpora were drawn from mainland Mandarin Chinese newspaper issues published between 1989 and 1993, which increases the likelihood that both articles and editorials discuss the same topics and use a similar vocabulary.

Our unit of text was the paragraph, which typically contains a single argument or thought. We decided to use paragraphs that were no more than 200 characters in our experimentation, assuming that longer paragraphs might hold multiple arguments. We split our raw data into two subsets: paragraphs 200 characters and below, and paragraphs larger than 200 characters. The small paragraphs were left in their original form, but we manually split the larger paragraphs into small sections under 200 characters, with individual paragraphs no smaller than 50 characters. We omitted large paragraphs which cannot reasonably be split up into sentences (for example, a long one-sentence paragraph).

2.2 Gold Standard Annotation

To evaluate our experiments, we employed workers through Amazon Mechanical Turk to tag our set of 719 editorial paragraphs. For each paragraph, the worker was asked, "Does the author of this paragraph express an argument?" In response, the worker categorized the paragraph by selecting "Makes argument," "Makes NO argument," or "Unsure". All text shown to the worker was written in both English and manually translated Mandarin. Instructions were screened by native speakers for clarity. Each paragraph was rated by three "Master Workers," distinguished as accurate AMT workers.

Though we provided clear instructions and examples for our categorization task, we found that the three workers for each task often did not all agree on an answer. Only 26% of paragraphs received an unambiguous consensus of "has argument" or "no argument" for the paragraph's argumentative nature. The rest of the paragraph results contain at least two different opinions about the paragraph. Since paragraphs receiving three different answers were likely unreliable for identification, we threw out those paragraphs, leaving 622 paragraphs for our methods. Around 78% of paragraphs were rated as argumentative, and 22% as non-argumentative.

Paragraph Consensus	Count	Percentage
Makes an argument	484	67.32%
Makes NO argument	138	19.19%
Unsure	43	5.98%
No consensus	54	7.51%
total	719	

Table 1: Breakdown of AMT paragraph results.

3 Models

We first score paragraphs according to the methods outlined below. Then, we determine the best score threshold for each method, and accordingly label paragraphs "argumentative" or "non-argumentative."

3.1 Method 1: Identification by Comparative Word Frequency

Our first method of evaluation is based on a process outlined by Kim and Hovy in a paper on identifying

opinion-bearing words (Kim and Hovy, 2005). We first follow Kim and Hovy's process to construct a list of word-score pairs. Then, we use these scores to evaluate our paragraphs of editorial text.

Kim and Hovy postulate that words which appear more often in editorials than in non-editorial text could be opinion-bearing words. For a given word, we use the Reportage and Editorials subcorpora to find its unigram probabilities in both corpora, then compute a score that indicates its frequency bias toward editorial or reportage text. Words that are relatively more frequent in editorial text are more likely argumentative.

$$Score(W) = \frac{EditorialProb(W)}{ReportageProb(W)} \quad (1)$$

Kim and Hovy further specify a way to eliminate words which do not have a repeated bias toward editorial or reportage text. We divide the Reportage and Editorial corpora each into three subsets, creating three pairs of reportage and editorial subsets. Then, for each word, we compute word scores as specified above, but for each pair of reportage and editorial subsets. This creates $Score_1(W), Score_2(W), Score_3(W)$, which are essentially ratios between editorial or reportage appearances of a word. We only retain words whose scores are all greater than 1.0, or all below 1.0, since this indicates repeated bias toward either editorials or reportage (opinionated or non-opinionated) text.

After scoring individual words, we rate paragraphs by assigning a score based on the scores of the individual words which comprise them. If a paragraph P contains n opinion words with corresponding frequency $f_1, f_2, \ldots f_n$ and assigned scores $s_1, s_2, \ldots s_n$, then the score for the paragraph is calculated by following:

$$Score(P) = f_1 s_1 + f_2 s_2 + \ldots f_n s_n \quad (2)$$

From these scores and our tagged data, we determine a best score threshold by tuning on our tagged data, which produced a threshold of 40.0.

3.2 Method 2: Targeting Known Argumentative Words

Our second method involves creating a list of known argumentative words that appear in the Editorials corpus and scoring paragraphs based on how many of these words appear in them. First, we constructed a list of the most frequent argumentative words that appear in the Editorials corpus. Then, we assigned each paragraph a score based on presence of these words.

We manually selected the most frequent argumentative words in the Editorials corpus by sorting a list of the words and their frequencies. Words were selected for their likelihood of indicating argumentation. Generally, the most common words which indicated opinion also possessed non-argumentative meanings. For example, the common word "要" can mean "to want" as well as "must" or "if."

Word	Translation	Count	%
我们	we	219	2.55
要	must	210	2.45
问题	problem	192	2.24
就	right away, at once	158	1.84
而	and so, yet, but	131	1.53
都	all, even (emphasis)	116	1.35
更	even more, further	87	1.01
但	but	86	1.00
还	still	84	0.98
好	good (adj)	76	0.89
人们	people	64	0.75
自己	self	61	0.71
却	however	57	0.66
人民	the people	53	0.62
必须	must	49	0.57
认为	believe	49	0.57
为了	in order to	48	0.56
我	I	47	0.55
重要	important	46	0.54
因此	consequently	46	0.54

Table 2: Constructed list of known argumentative words by frequency. Horizontal lines mark boundaries between 10-, 15-, and 20-word lists.

Scoring paragraphs based on this list was simple: we awarded a paragraph a point for each instance of any of the words on the list. We were interested in whether the presence of a few argumentative words could indicate argumentation in the entire paragraph. We determined the best word list size and the best threshold that returned the most accurate labels, a word list size of 15 words and a threshold of

1. For this model, a threshold of 1 means if the paragraph contains at least one word from the word list, it is labeled as argumentative.

3.3 Method 3: Combined Method

Our third method of identifying arguments combines the previous two methods. Similar to the second method, we scored paragraphs based on a list of argumentative words. However, instead of manually selecting argumentative words from a word frequency distribution, we created a list of opinionated words by picking a number of the highest-scoring words from the first method.

In structuring this combination method, we theorized that the highest-scoring words are those which are the most common opinionated words, since they have the highest probability of consistently appearing throughout the Editorials corpus and not the Reportage corpus. By using these words instead of manually-picked argumentative words, we scored paragraphs using a list of words based on the composition of the corpus itself, with the intent of creating a more relevant list of words.

Scoring remained similar to the second method, where we awarded a paragraph a point for each instance of any of the words on the list. Again, the threshold which produced the best results was 1. That is, if a paragraph contained at least one word from the list, it was labeled as argumentative.

4 Results

4.1 Method 1: Identification by Comparative Word Frequency

Method	Accuracy	Precision	Recall	F_1 score
1	0.841	0.847	0.971	0.905
2	0.801	0.826	0.942	0.880
3	0.371	0.850	0.233	0.366
1 = Relative Word Frequency Method (T=40)				
2 = Targeting Argument Words (T=1,W=15)				
3 = Combined Method (T=1,W=20)				

T = threshold, W = word list size

Table 3: A comparison of the best metric scores of all three methods.

Our experiments produced the best performance under the relative word frequency method, achieving 84% accuracy and an F_1 score of 0.91. These scores were closely followed by the second method with 80% accuracy and an F_1 score of 0.88.

Despite these high scores, we were surprised to find that our relative word frequency system had scored many non-argumentative words very high. For example, the highest-scoring word was 自民党, "Liberal Democratic Party." When we eliminated words with non-argumentative POS tags (i.e. nouns and other noun forms), the highest-scoring word was 监测, "to monitor" (Table 4). These words were likely rare enough in the Reportage corpus that they were awarded high scores. Words indicative of arguments such as 必要, "necessary," scored high, but largely did not make up the highest-scoring words.

Word	Translation	Score
监测	to monitor	57.917
谈判	to negotiate	51.656
污染	to pollute	34.437
发展中国家	developing country	32.872
停火	to cease fire	29.741
整治	to rennovate, restore	28.176
腐败	to corrupt	26.610
北方	north	25.129
断	to break	25.129
匿	to hide	25.062

Table 4: Highest-scoring words from the Kim and Hovy scoring in the statistical method, words with non-argumentative POS tags removed.

As a result, our list of opinion-bearing words contained non-argumentative words along with argumentative identifiers, artificially raising paragraph scores. Paragraphs were more likely to score high, since our system labeled many non-argumentative paragraphs as argumentative. The inflated word scores are likely a result of mismatched editorial and reportage corpora, since a word that is relatively rare in the Reportage corpus and more common in the Editorials corpus will score high, regardless of its actual meaning. However, this approach still performed well, suggesting that these non-argumentative words, such as "to monitor," may be used to persuade in context (e.g. "The government monitors its people too closely").

4.2 Method 2: Targeting Known Argumentative Words

Our second method similarly performed well, with high accuracy and fewer false positives than the previous method, due to the list of words that clearly indicated argumentation. The best performance was given by a threshold of 1. That is, the system performed best when it marked a paragraph argumentative as long as it has at least one of the words from the list. Results did not significantly improve even if the list was expanded or the score threshold was raised, implying that adding words to the 10-word list, even if the new words had exclusively argumentative meanings, did not significantly improve performance. The more frequent semi-argumentative words like "而" ("and so," "yet") had a greater positive effect on accuracy than obviously argumentative words like "必须" ("must") which do not appear as often in the corpus.

4.3 Method 3: Combined Method

Since our combined method relied heavily upon the word scores generated by the relative word frequency approach, the results showed significant errors. Seeded with a word list that did not contain solely argumentative words (e.g. "to monitor" as well as "to pollute"), the combined method attempted to find argumentative paragraphs using words which did not exclusively indicate argumentation. Overall, the combined method rated many more argumentative paragraphs as non-argumentative than the reverse, and performed poorly overall with a F_1 score of 0.37.

5 Related Work

Prior work on argument identification has been largely domain-specific. Among them, Sardianos et al. (2015) produced work on argument extraction from news in Greek, and Boltûzic and Jan Ŝnajder (2015) worked on recognizing arguments in online discussions. Kiesel et al. have worked on a shared task for argument mining in newspaper editorials (Kiesel et al., 2015). They contributed a data set of tagged newspaper editorials and a method for modeling an editorial's argumentative structure.

Because argument mining is a relatively new field within the NLP community, there has been no argument identification study performed on Chinese editorials, although there has been a significant amount of work on opinion identification. In particular, Bin Lu's work on opinion detection in Chinese news text (Lu, 2010) has produced a highest F-measure of 78.4 for opinion holders and 59.0 for opinion targets.

6 Conclusion and Future Work

In this study, we sought to computationally identify argumentative paragraphs in Chinese editorials through three methods: using relative word frequencies to score paragraphs; targeting known argumentative words in paragraphs; and combining the two methods. Our experiments produced the best performance under the relative word frequency method, achieving 84% accuracy and an F_1 score of 0.91.

Despite these high scores, we found our relative word frequency system scored many non-argumentative words very high. These words were likely rare enough in the Reportage corpus (but common enough in the Editorials corpus) that they were awarded high scores. As a result, our list of opinion-bearing words contained non-argumentative words along with argumentative identifiers, raising paragraph scores and producing false positives.

Future work could be done to improve upon Kim and Hovy's method in order to more accurately score words. In particular, it is necessary to avoid scoring non-argumentative words high, simply due to their presence in the Editorials corpus and absence in the Reportage corpus. In our experiment, we eliminated high-scoring non-argumentative words like "自民党" ("Liberal Democratic Party") by removing nouns from scoring. However, this also eliminates argumentative nouns like "问题," meaning "problem" or "issue." One solution to removing topic-specific nouns while keeping argumentative nouns is identifying the editorial topic and its domain-specific words, which would prevent the method from scoring rare but non-argumentative words high. Another benefit to determining word context is distinguishing between argumentative and non-argumentative senses of a word. For example, if the word "garbage" appears in an article discussing landfills, it is likely not argumentative. However, if it appears in an editorial discussing recent movies, it is more likely to be an argumentative word (e.g.

"That movie is garbage."). In our current system, we cannot distinguish between these two uses. If the word "garbage" appeared equally in news text as a neutral word (say, in an article discussing landfills) and in the editorials corpus as an argumentative word ("that's garbage!"), then the score of "garbage" would be low, and we would be unable to identify the argumentative nature of "garbage" in editorials. Another solution is to observe the context in which words appear. If the word "garbage" appears in proximity to words like "landfill" and "recycling," then we could guess that the usage of this word is non-argumentative.

By improving the list of opinionated words in the relative word frequency method, we could not only improve its scoring system, but perhaps even improve upon our combined method to produce even better, more accurate results than the first two methods used. We hope our research provides a benchmark or foundation for future research in the growing field of argument mining.

Acknowledgments

I would like to thank the Princeton University Department of Computer Science for their generous support in the funding of this project, without which this work would not be possible. I would also like to thank Professor Christiane Fellbaum for her invaluable guidance and advice throughout this project.

References

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

Johannes Kiesel, Khalid Al-Khatib, Matthias Hagen, and Benno Stein. 2015. A shared task on argumentation mining in newspaper editorials. *NAACL HLT 2015*, page 35.

Soo-Min Kim and Eduard Hovy. 2005. Automatic detection of opinion bearing words and sentences. In *Companion Volume to the Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP)*.

Bin Lu. 2010. Identifying opinion holders and targets with dependency parser in chinese news texts. In *Proceedings of the NAACL HLT 2010 Student Research Workshop*, pages 46–51.

Anthony McEnery and Zhonghua Xiao. 2004. The lancaster corpus of mandarin chinese: A corpus for monolingual and contrastive language study. *Religion*, 17:3–4.

Nathan Ong, Diane Litman, and Alexandra Brusilovsky. 2014. Ontology-based argument mining and automatic essay scoring. In *Proceedings of the First Workshop on Argumentation Mining*, pages 24–28.

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

Yi Song, Michael Heilman, Beata Beigman, and Klebanov Paul Deane. 2014. Applying argumentation schemes for essay scoring.