Argument & Claim Annotation Report

COSI 140: Natural Language Annotation for Machine Learning

Hayley Ross, Meiqi Wang, Michael Zhong

Introduction

Our project studies the annotation of claim and argument, specifically the annotation of argumentation strategy. This involves categorizing sentences into classes such as anecdote, testimony, statistics, assumption, common ground, and other forms that the author used for developing arguments within news editorials. The study of argument structure can help in analyzing argument schemes and strategies, creating writing assistance systems (or automatic grading), analyzing media or public opinion on political topics or products, and structuring and querying legal documents.

Related Work

There are many different types of argument annotation in previous work, including premise and claim annotation, which is the traditional type of argument annotation (discussed in Sardianos et. al, 2015 [4]), and the related but more detailed notion of tagging propositions and support / attack / rebuttal relations. A very different type of argument annotation is argumentation strategy (classifying sentences as anecdote, testimony, statistics, assumption, common ground and other) annotation (discussed in Al Khatib et. al, 2016 [1]). This addresses the methods the author uses to persuade the reader rather than the truthfulness or coherency of their argument. A third type of argument annotation involves classifying either fact vs. opinion, degree of (belief in) opinion (discussed in Bal and Saint-Dizier, 2009 [2]), or labelling short statements as pro/contra a particular issue, involving emotion, insult and sarcasm annotation to identify main arguments pro or contra a topic (so-called "argument facets"), as discussed in Swanson, Ecker and Walker, 2015 [5].

Choice of Annotation Type

Our project decided on argumentation strategy as our type of annotation, in other words, how the author persuades, since some types of argumentation annotation listed above require substantial training of annotators, which is not practical given the time and circumstances of this current project. If annotators are minimally trained in terms of these types of argumentation annotation, the inter-annotator agreement will be low (see Musi, Ghosh and Muresan, 2016 [3]), which is not an ideal result. In addition, the project corpus consists of news editorials, which don't usually have clear support/attack relations, as found by Al Khatib et. al, 2016 [1], so argumentation strategy mining is more suitable.

The Source Corpus

The project corpus is from a news editorial corpus for mining argumentation strategies by Al Khatib et. al. in 2016 [1]. The editorials of the corpus is from three diverse news portals: Al Jazeera, Fox News, and The Guardian; the three portals have diverse cultures and styles, they are well-known and have editorial sections. Our corpus consists of 100 randomly selected editorials from each source. The criteria for selection is as follows:

1. They were published over the same time interval (December 2014 and January 2015)
2. They had at least 5 comments;
3. They contained at least 250 words.

The rationale behind the criteria was that the same time period would facilitate topic overlap between the sources, the presence of multiple comments meant that the editorial had led to a discussion, and the length would filter out short texts that did not contain much argument.

The corpus is publicly available as a CSV, and has been processed by us for loading into MAE.

Annotation

The Schema

We decided to base our annotation schema on the original schema for this corpus, as developed by Al Khatib et al. in [1]. They use the following categories:

- Common Ground
- Assumption
- Testimony
- Statistics
- Anecdote
- Other

However, they admitted that even after revising their guidelines following a pilot study, their annotators still had difficulty distinguishing Common Ground from Assumption and Anecdote. In fact, their Fleiss' kappa for Common Ground was just 0.114. We believe this is because Common Ground is a flawed category and extremely subjective, depending not only on the article but also what the annotator may happen to know or believe to be evident. This motivated us to replace Common Ground with a much more clearly defined Fact category.

Also, it seemed odd to us that the study distinguished Assumption and Testimony for all statements except statistical ones. Are we to assume that all statistics used in an article are immediately true? In fact, people are often bad at quoting statistics, and will either misrepresent them or even spread ones which they heard somewhere but are in fact unfounded. These days, quoted statistics should be treated with suspicion equal to any other assertion.

First Draft

Besides the points above, we also noted that the majority of any given article was tagged as Assumption. This motivated us to break down this category into subcategories, as well as renaming it to Assertion which seemed more fitting for the type of content in an editorial. Secondly, we noted that in much modern political argumentation (especially since the Trump era), *ad hominem*-type attacks have become more common, and so we wanted to determine whether such insults could be found in this 2014-2015 corpus as well.

This yielded the following first draft of our schema. (Examples give an idea of the definition of the category; formal definitions of the categories will follow for the final version of the schema.)

Each unit is one sentence (regardless how many clauses it should contain), and should be tagged with exactly one of the following categories. If two sentences are the same type (e.g. Anecdote), they should be tagged individually.

- Assertions
    - Moral assertions and emotional appeals
      "People should always respect their elders."
      "Nobody in their right mind would think that women should be treated that way."
    - Generalizations
      "All teachers are overworked."
    - Proportional assertions
      "40% of children are obese."
      "Most children like school."
    - Individual / plain assertions
      "Brandeis is the best university in America."
- Facts
  "There are 50 states."
- Testimonies (statements with a quoted source)
  "The Washington Post reported that the Senate has passed the resolution."
  "A recent paper in Science discussed the architecture of the newly discovered protein."
- Anecdotes
  "I was in the supermarket the other day and the price of meat products was outrageous."
  "I visited New York last month and I was not impressed by the state of its public transit."
- Insults
  "He's an idiot so why should we listen to him?!"
  "What a meanie, everything he said was wrong!"
- Not part of the argument
  "Good morning, America."
- Other forms of persuasion / argument
  "Yeah right, that is totally going to work."

Outcome of Preliminary Annotation

Based on this draft schema, we each set out to annotate the first three articles of the corpus, one from each news source. This quickly showed some issues with the draft schema. Firstly, we noted that deciding which subcategory an Assertion should be took far more time and cognitive effort than choosing any of the other categories. This contradicted our desire to make the annotation easy and swift for our annotators, since time for this project was limited and it is preferable to have a larger quantity of well-annotated data than a small quantity of data or a large quantity with low agreement (because the annotators did not give the decisions enough time).

Additionally, discussing the schema in our presentation showed that the subcategories for Assertion were not mutually exclusive - for example, one might have a moral appeal that was also a generalisation or statement about a majority. The same was true of Fact and Testimony - a number of cases seemed overlapping, and in any case from the author's perspective both are considered as true. (Additionally, it is Fact that the quoted source said what they did.) Finally, we noticed that a particular type of "Other Form of Argument", rhetorical questions, were quite common in the articles we chose. This led us to revise our schema to the final version below.

Final Version

The final annotation schema uses the following list of tags. Note that there are no subcategories for Assertion anymore. Short descriptions for each of these categories are taken from our Annotation Guidelines: further clarification, detail and lists of examples can be found there.

The guidelines for what to annotate (namely one sentence per tag) remain the same as above, but we also added instructions for how to handle quoted speech involving multiple sentences to the guidelines.

- Assertion
  Main category for arguments. Includes claims made by the author which are personal opinion, which not everyone would agree with, or are simply not backed up by a verifiable source.
- Fact / Testimony
  Covers both established facts and statements with a given source, i.e. testimonies, since these are viewed as fact by the author for the purposes of the argument. A fact is defined as anything that could be looked up on Wikipedia or in a similar authoritative source; a testimony includes all statements, regardless whether true or false, which come with an accompanying source.
- Anecdote
  Stories or personal experience used to develop the author's argument.
- Rhetorical Question
  Question raised to illustrate argument, with answer left implicit or given immediately after by author.
- Insult
  Attack on a person's character, behaviour, appearance, gender, race or other personal traits, regardless whether justified or not.
- Not Part of the Argument
  Includes content such as headings, picture captions, greetings, and other segments of text in the article which do not form part of the argument.

- Other Form of Argument
  Statements that seem to be part of the author's argument, but don't fall into any of the categories above. Includes sarcasm, as well as fragments of sentences.

We believed that this list of categories should be comparatively swift for annotators to choose between, compared to other types of argument annotation, while also adding benefit to the previous study. In particular, these categories better distinguish verifiable fact or testimony from unverified statements, as discussed above by clarifying the Fact (formerly Common Ground) category and removing the ambiguous Statistics category, as well as adding two new categories: Insult and Rhetorical Question.

Some difficulty in distinguishing categories still exists, most notably the distinction between Assertion and Fact/Testimony is quite fine at times. We wrote an extensive section in our Annotation Guidelines on this topic; please refer to that for more information.

The Annotation Process

Division of Annotation

Although there are 300 editorials available, based on the experience during our own test annotations, we decided that it was impractical, given the time constraint, to fully annotate all of the editorials. In fact, only 18 editorials could be annotated by an annotator during the given period. Therefore, we needed to decide which editorials to select and how to assign them to the annotators.

In order to obtain reliable gold standards, each annotated article must be processed by at least two annotators. Thus, the decision came down to whether we would prefer three annotators for each editorial, providing us with a majority vote that could guide the gold standard selection, or two annotators for each editorial, relying on a selection algorithm to construct the gold standard but receiving a maximal amount of tags.

We opted to have three annotators annotating each editorial because for our task, we chose to prioritize the quality of the tags over the quantity of tags when we could meet the minimum requirement for number of tags. A majority vote will provide us with a good overview of how our schema is used in practice and a large amount of gold standard already in place without resorting to further adjustments. If the inter-annotator agreement (IAA) is very high, then we could potentially proceed with fewer annotators in the future, as that would suggest that our schema is executable and accurate.

Difficulties

During the annotation process, some difficulties came up, the first one is the tagging of Assertion vs. Fact/Testimony, a second one is the tagging of Anecdote vs. Fact/Testimony. Even though we wrote about this in the annotation guidelines, for complicated sentences, the boundaries between Assertion vs. Fact/Testimony and Anecdote vs. Fact/Testimony are sometimes unclear and hard to decide for the annotators. For example, part of the anecdote paragraphs might belong to other categories, so it leaves the annotator to decide whether to tag the anecdote as a whole or to tag each sentence separately. Thus even with the guidelines it's still a subjective decision.

The third difficulty is tagging sentences with multiple parts, since annotators don't always have the intuition about which final tag the sentence with parts of different categories might belong to. For example, in the sentence *"As 2014 ends, the stock market is at record highs but our traditional institutions and self-confidence are in decline."*, it at first seems to be a Fact/Testimony, but right at the end of the sentence there is definitely an Assertion: *"our [...] self-confidence is in decline"*. Sometimes annotators didn't spot parts like these and simply tagged the whole sentence as Fact/Testimony.

The fourth difficulty is that only one sentence should be annotated at one time, while annotators sometimes annotate adjacent sentences with same tag together. Thus our algorithm for creating the final gold standard data needs to deal with this issue.

Inter-Annotator Agreement

Al Khatib et al. achieved a Fleiss' Kappa of 0.56. As we developed our schema based on theirs, we aimed for a similar Fleiss' Kappa. After compiling the annotations, we used a Fleiss' Kappa module developed by Shamya (see https://github.com/Shamya/FleissKappa) to calculate the Fleiss' Kappa.

Our IAA is 0.54, with a PA of 0.75 and PE of 0.45. This is, for our purpose, very similar to what Al Khatib et al. achieved and much higher than the general threshold of 0.2. We are satisfied with our IAA, although it does leave room for further improvements.

The New Corpus

Creating the Gold Standard

Automatic Processing

Since the 21 annotated articles contain over 800 sentences, it seemed more efficient to use Python to create the gold standard and determine the majority vote of the annotators where possible. We chose a CSV format for our gold standard since it is easy to read with e.g. Pandas for the machine learning part of the problem.

The script parses the XML created by MAE and creates a list of GoldStandardLine objects, which capture the file ID, start and end index of the sentence, the tags by each annotator and the majority vote for that sentence. After the sentences by the first annotator are read, fuzzy matching is used to try to find the existing GoldStandardLine object which the sentence in the MAE XML corresponds to. This is necessary not only because annotators may highlight different amounts of whitespace or accidentally miss off punctuation (despite the instructions in the guidelines), but also because due to operating system differences the start and end indices may not be exactly the same for the same sentence between annotators. Further, we noticed that it was a common mistake for annotators to tag multiple sentences at once if they had the same tag (again, despite the instructions). The code attempts to address this, and will split a “sentence” into multiple sentences if it can find each of those smaller sentences in its existing list. It will then apply the same tag to all of them.

Finally, the code iterates over all the GoldStandardLine objects and calculates the majority vote where one is available, leaving it blank if no two annotators chose the same tag. This may happen because the sentence was truly difficult to tag, but more likely due to segmentation differences or annotators not tagging certain pieces of text (e.g. image captions) at all instead of tagging them as Not Part of Argument. (The presence of these and other pieces of non-argument text such as headings presumably arose from how the original authors scraped these articles off the internet. Unfortunately we didn’t have time during this project to do additional sanitisation on the data.)

Final Adjustments By Hand

The final adjustment to the gold_standard.csv deals with several problems of the data, the major problem is due to the sentence segmentation differences between the annotators: parts of the sentences might be tagged differently by some annotators when they should be tagged as a complete sentence. For this problem, the sentence as a whole needs to be retagged by the annotators. Similarly, sentences that need to be tagged separately also need to be retagged by some annotators. These data were collected and sent back to the annotators for the retagging work, then they were included into the final corpus data.

After they were included, there still existed some sentences in the gold standard file with no gold standard tag, since for some sentences the three annotators might all have different tags, thus, our group did the final manual adjudication work by hand. The final corpus contains 21 articles and 829 sentences.

Creating the Training, Development and Test Sets

Through discussion, our group decided not to use a sequential model (as discussed below), since for the annotators, the annotation of the sentence rely little on the previous sentence, thus the data was split by 8-1-1 (training data - development test data - test data) within each annotated article, and each part was included in the final training data file, dev test data file and test data file. This part of the work was completed by code, which basically counts the number of sentences by article id, split them proportionally and write them into three separate csv files.

The Model

Choice of the Underlying Model

There are two routes for applying machine learning to this problem. On one hand, we have a sequence of sentences, so it might seem obvious to use a sequential model such as an HMM or a CRF. On the other hand, watching human annotators choose the tag for a sentence, they rely remarkably little on the tag that came before, instead using the content of the sentence plus a comparatively small amount of context. After all, for cases such as Fact/Testimony and Assertion, it does not matter what the previous sentence was, only if the statement is verifiable or not. Even for Anecdote, while it may be useful to know that the previous sentence was an Anecdote as they normally go on for multiple sentences, in fact it is the presence of keys such as personal pronouns and other storytelling words that determine whether the current sentence is still part of the Anecdote. Thus we decided that a regular classifier would be a good first choice as it mimics how humans classify the sentences. Nevertheless it’s still possible that a sequential model would be good for a machine and so it would be a very interesting extension of this project to compare the two.

Secondly, we needed to decide whether to use a generative or discriminative model. Given the wide variety of sentences going into our model (recall that sentences are the units, not words), most of which will only be seen once and thus have a tiny probability, a generative model seemed like a poor choice. Between these two restrictions, then, choosing a Logistic Regression (as used in Homework 1) seemed the obvious choice for a first pass.

Features

The most basic feature was bag of words, which was a good place to start, given that we knew little about what could distinguish the different argumentation strategies.
Somewhat surprisingly, the plain bag of words with counts yielded a weighted average F1 score of 0.71, as shown below in table 1. Perhaps more surprisingly, removing rare words (with count less than/equal to 2) didn’t change the numbers at all.

Table 1. Bag of Words with Counts, on the dev-test set

	Precision	Recall	F1-score	Support
Anecdote	0.50	1.00	0.67	1
Assertion	0.82	0.87	0.84	31
Fact-Testimony	0.33	0.14	0.20	7
Non-Argument	0.00	0.00	0.00	1
Other-Form	0.00	0.00	0.00	0
Rh-Question	0.00	0.00	0.00	0
Micro avg	0.72	0.72	0.73	40
Macro avg	0.28	0.34	0.29	40

Weighted avg	0.70	0.72	**0.71**	40

Based on a decently informative bag of words feature, we added counts of names and numbers to recognize elements that are very common in the news editorial genre such as person or organization names and reported figures, which improved weighted average F1 score to **0.82** on the dev-test set. There was a lone Other-Form tag that was recognized by the new features too. We also tried adding counts of pronouns and connectives such as “because” and “although”, but this didn’t improve our accuracy.

Both of these features failed to distinguish the lone rhetoric question in the set, however, which was covered by adding counts of POS tags, although that lowered the weighted average F1 score to **0.74** on the dev-test set, which would still be better than the plain bag of words. The rationale behind adding POS tags was that we speculated that some argumentation strategies might have a higher amount of a part of speech than others. For example, assertions might employ more adjectives while facts/testimonies might use more nouns. We did not have any background data about this, however, so we started with a plain count of POS tags.

All of the currently selected features could also be analyzed further by examining most informative features and observing the distribution of features such as count of POS tags, which along with other possible features will be discussed further in the Discussion section. This might allow us to isolate the good parts of each of these attempts and bring us up to the highest seen score of 0.82 while also capturing the rhetorical question and non-argument tags.

For completing this stage of the project, we settled on using the POS tags as well as all of the count features discussed. On the test set, this yielded the following results. Interestingly, they are substantially higher than on the dev-test set for the tags, which must be down to the precise data in each set, but unlike the dev-test set the test set doesn’t contain any anecdotes or other forms of argument, so we can’t see how well the algorithm performed on those. As discussed below, this is problematic. (The precise make-up of the test set is 57 assertions, 15 facts/testimonies, 3 rhetorical questions and 2 non-argument sentences.)

Table 2. Final run with all features on the test set

	Precision	Recall	F1-score	Support
Assertion	0.97	0.92	0.94	71
Fact-Testimony	0.73	0.85	0.79	13
Non-Argument	1.00	1.00	1.00	2
Rh-Question	0.33	1.00	0.50	1
Micro avg	0.91	0.91	0.91	87

Macro avg	0.76	0.94	0.81	87
Weighted avg	0.93	0.91	**0.91**	87

Discussion

Learning

First of all, it is encouraging that our schema managed to obtain a Fleiss' Kappa of 0.54, which is very close to what Al Khatib et al. achieved, despite having a lot less annotator training time. We now know that the schema has good potential to be employed more widely to annotate more data and could also allow us to reduce the number of annotators needed in order to obtain more tags. However, there is one significant problem: in the 18 articles, insult was not found. We might need to reconsider whether we should still include it even though it is very interesting. A possible alternative could be scathing language, which could describe some of the editorial excerpts in the corpus, which, however, presents its own issue that is scathing language itself does not really pertain argumentation strategy, compared to ad hominem personal attacks, which, although fallacious, is a fairly common argumentation strategy.

Another possible improvement should be handling untagged sentences. Currently, we have a placeholder tag to account for this issue which is included in the Fleiss' Kappa. If we could ensure that we have full coverage in the future, the Fleiss' Kappa could be higher than what it is now.

In addition, we could still use a sentence segmentation algorithm that could break down composite sentences such as conjunctions and clauses. Although the IAA is satisfactory without such segmentation tools, it must be noted that the schema also went to great length to explain how we would treat certain complex sentences. For example, if a conjunction contains both facts and assertions, we would consider it as an assertion as a whole with the caution that the author might want to lead the readers to believe in a less accepted assertion by presenting it alongside a more credible fact. Another example is when an author writes in length about an anecdote, it becomes problematic to the annotators as they are confused whether to tag the whole sentence as an anecdote even if it might be very similar to an assertion. With a good sentence segmentation algorithm, we could provide more defined segments for the annotators, which could further improve the IAA, and obtain higher resolution data where each category may share fewer but more unique features while currently we are simply relying on bag of words and POS tags.

Last but not least, it is important to note here that the data we obtained were extremely skewed as a significant majority of the tags in the gold standard was assertion (above 66%). This means that even if an algorithm were to guess assertion all the time, it would still reach 66% accuracy. Therefore, a high accuracy is still encouraging but not as significant as if the tags were more evenly distributed. Furthermore, some of the tags were so rare that it could only be present in one of the three subsets of the data, which only adds to the problem of data scarcity, and made the metrics produced by the machine learning algorithm more difficult to interpret. For example, our micro average varied wildly depending on whether our features captured the one rhetorical question or not. In retrospect, we should perhaps have split our data along a 60-20-20 split simply in order to have more of the rare categories in our dev-test and test set, at the expense of the training set.

Thus, combined with the points mentioned above, we could adjust the schema and develop a segmentation tool to fetch more and higher quality data to both solve the issue of data scarcity and provide our machine learning algorithms with better category definitions.

Future Direction

Besides what we could do about the corpus and the annotation workflow, there are also some more features that we would like to explore. There are a few levels of possible additions and we will address them from lower level to higher level.

At sentence level, we think it may be helpful to include syntactic and semantic information. Argumentation strategy, regardless of how well bag of words could distinguish some tags given limited data, contains far more information than mere words would convey. Strategies such as fact/testimony may have hallmark expressions such as "<News Source> reported that…", "<City Name>, where this incident took place, is the capital of <Region/Country Name>", and "<Person Name> said that they would investigate this matter further". Within sentences like those, vocabulary such as "report" and "investigate" could also be informative, but even better, the syntactic structure of a statement or quotation sentence and the semantic mapping of the verbs "...report…" and "...said…" could reveal a lot about the strategy, as could identifying verbs such as "believe" or "should". Furthermore, better person/organization name recognition would likely enable us to better recognize testimonies where reference and citation are common. These new features could also help distinguishing unique strategies such as rhetoric questions, which have a distinct syntactic structure. (Surprisingly, the presence of a question mark doesn't seem to be enough for the model to identify them - this also warrants further investigation.)

At document level, we also want to consider CRF as a potential ML model. We speculate that beyond sentence level features, document level context clues could also help with argumentation strategy recognition. In this case, a certain strategy may be more likely to occur given a context or sequence. Plus, news editorials form an interesting genre that usually hinges on a central topic like news articles but still reflects authorship through its less rigid formatting. This way, document level analysis could reveal a lot about common editorial writing practice, should there be a pattern to exploit.

References

[1] Al Khatib, K., Wachsmuth, H., Kiesel, J., Hagen, M., & Stein, B. (2016, December). A news editorial corpus for mining argumentation strategies. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (pp. 3433-3443).

[2] Bal, B. K., & Saint-Dizier, P. (2009). Towards an analysis of argumentation structure and the strength of arguments in news editorials. In *Persuasive Technology and Digital Behaviour Intervention Symposium* (pp. 64-67). The Society for the Study of Artificial Intelligence and the Simulation of Behaviour.

[3] Musi, E., Ghosh, D., & Muresan, S. (2016). Towards feasible guidelines for the annotation of argument schemes. In *Proceedings of the third workshop on argument mining (ArgMining2016)* (pp. 82-93).

[4] Sardianos, C., Katakis, I. M., Petasis, G., & Karkaletsis, V. (2015). Argument extraction from news. In *Proceedings of the 2nd Workshop on Argumentation Mining* (pp. 56-66).

[5] Swanson, R., Ecker, B., & Walker, M. (2015). Argument mining: Extracting arguments from online dialogue. In *Proceedings of the 16th annual meeting of the special interest group on discourse and dialogue* (pp. 217-226).