Datasets for the NLP task of Detecting Contradictions

Beatriz Baldaia

June 11, 2020

This document provides a description of the built datasets used for my Master's research project which explores transfer learning for the task of detecting contradictions. For the target task domain we considered contradictions in a political domain, therefore, we built two datasets containing pairs of contradictory statements from two different sources: an online article exposing Donald Trump contradictory claims, and government-related instances of the MultiNLI corpus. Then for the source task domains, we collected data from five publicly available corpora: MultiNLI, US2016, Argumentative Microtext, Argument Annotated Essays, and W2E.

Disclaimer

The datasets presented here were constructed by a single and untrained annotator which makes them not strongly reliable. Particularly the DonaldTrump dataset, that was built manually by us, missing robustness due to the lack of annotation guidelines, and the lack of professional training, preparation and knowledge of how to create a reliable and representative corpus for a specific task. Furthermore, since this dataset was constructed by a single annotator, we do not employ any technique for calculating annotators agreement level, such as the Inter Annotator Agreement (IAA) metric. Besides all the above, the number of examples incorporated in that dataset is small. Thus, polishing, enriching and increasing this dataset is essential.

Datasets Overview

An instance of a dataset is a pair of two texts with a lable that can be either 1 (the positive class, meaning the two texts are contradictory), or 0 (the negative class, meaning the two texts are not contradictory).

A dataset can be composed by three Tab Separated Values files, for training ("train.tsv"), for validation ("dev.tsv"), and for testing ("test.tsv"). Each of these files has five columns: "Quality" (the example label, 0 or 1), "#1 ID" (id of the first document of the example pair),

Table 1: Datasets not initialy designed for the task of detecting contradictions.

Dataset	Language	Date of publish	Short-detail	Relation being tested	Availability address
Multi-Genre Natural Language Inference (MultiNLI)	English	2018	A collection of 433k sentence pairs annotated with textual entailment information, containing examples of different genres: "fiction", "government", "slate", "telephone", "travel", "9/11", "face-to-face", "letters", "oup", and "verbatim".	Contradiction in texts of different genres.	NYU
US2016	English	2019	Transcriptions of television debates leading up to the 2016 US presidential elections, and reactions to the debates on Reddit. The annotation of the corpus is based on Inference Anchoring Theory (IAT), containing three types of relations: inference, conflict and rephrase.	Arguments of disagreement between different speakers.	AIFdb
Argumentative Microtext Corpus	English and German	2015	Short texts that respond to a trigger question. The argumentation structure identifies the central claim of the text, supporting premises, possible objections and counters to these objections. The annotation guidelines are available online.	Author's counter--arguments attacking his\her own claims.	University of Potsdam
Argument Annotated Essays	English	2017	Argument annotated persuasive essays including annotations of argument components ("Major Claim", "Claim", and "Premise") and argumentative relations ("Support" and "Attack").	Author's counter--arguments attacking his\her own claims.	TU Darmstadt
Worldwide Event (W2E)	English	2018	Dataset for topic detection and tracking. 207,722 news articles covering a large set of 4,501 popular events, each belonging to one out of 10 categories.	Topic similarity.	W2E: A dataset for TDT

"#2 ID" (id of the second document of the example pair), "#1 String"(the first document of the example pair), and "#2 String" (the second document of the example pair).

As said before, we studied transfer learning, thus we investigate the effect of resorting to different document relationships, but still related with our task of detecting contradictions. Table 1 shows the selected corpora for our research, and the correspondent new relations between documents that we exploited.

Table 2 presents the dimension of the developed datasets.

Table 2: Datasets dimension. The two first rows correspond to the target task domain in the context of transfer learning.

Dataset (our given name)	Short description	Total examples	Total positives	Total negatives
DonaldTrump	Manually created dataset, based on the article from POLITICO about moments where Donald Trump contradicts himself.	250	144	106
MultiNLIGovernment	All instances of genre "government" from the MultiNLI corpus.	79350	26418	52932
MultiNLI-	All instances, except the ones of "government" genre, from the MultiNLI corpus.	333352	110938	222414
US2016	US2016, the largest publicly available set of corpora of annotated dialogical argumentation.	1882	941	941
ArgumentativeMicrotext	The argumentative microtext corpus consists of short texts that respond to a trigger question.	1133	403	730
ArgumentEssays	Argument Annotated Essays corpus.	6673	715	5958
W2E	A Worldwide-Event Benchmark Dataset for TopicDetection and Tracking.	4800	2400	2400

Datasets Description

In our research, we focused on contradictions, and use the political domain as a case study. For that purpose, we use two datasets as target task domain, one built by us from scratch, based on an online article, and the other containing a specific section of the publicly available corpus MultiNLI. Next we describe each dataset developed to construct the source and target task domains.

Target Task Domains

DonaldTrump

For this dataset we focus on a specific entity, Donald Trump, the president of the United States, as a case study. The reason why we chose this well-known person is that there is a lot of controversy around his allegations, and the online magazine POLITICO Magazine [1] has an article exposing some of Trump's self-contradictions[2]. Hence, our domain is expected to be, mainly, political statements, although it can contain other topics escaping from our scope.

The article has a list of Trump's quotes in interviews, debates, posts in the social media network Twitter, and propositions from his published books. However, the list does not have a pattern, meaning that, you do not always have one quote followed by another that contradicts it. Thus, it requires to read and to analyse each quote. Moreover, not all the instances provide the source link (e.g. when it is quotes from Trump's books), or the provided source link is sometimes unavailable or requires website subscription (like some blocked articles from The New York Times). So, in order to approve and verify quotes, we occasionally had to manually search on the internet for the quote, resorting to different sources, until we could find means to prove the reliability of the sentences.

Politico is an American political opinion company that produces contents covering politics and policy in the United States and internationally. It has professional journalists working for them to provide interesting, true and authentic content, so we assume we can trust this source. Nevertheless, some quotes are not easily identified as contradictions, as we can see in the following four examples:

Example 1:

> *"I love the poorly educated."*
>
> *"I see no value whatsoever in believing ignorance to be an attribute."*

We can argue that it is contradictory to be against ignorance, considering it an unacceptable "attribute", and, at the same time, claim to adore ignorant and poorly educated people.

[1] https://www.politico.com/section/magazine
[2] https://www.politico.com/magazine/story/2016/05/donald-trump-2016-contradictions-213869

However, it is possible to judge the ignorance of people, but still like them.

Example 2:

> *"I'm very pro-choice."*
>
> *"And I am very, very proud to say that I am pro-life."*

Here it is impossible to capture the contradiction if you don't know the meaning of the concepts "pro-choice"[3] and "pro-life"[4].

Example 3:

> *"Everybody kisses your ass when you're hot. If you're not hot, they don't even call. So it's always good to stay hot."*
>
> *"He thinks he's hot stuff. And I hate people that think they're hot stuff, and they're nothing."*

While Donald Trump believes that it is good to "stay hot", he also says that he hates "people that think they're hot stuff". But, is he talking about everyone who thinks is "hot stuff" or the ones that think that, but, in fact, they are not? If we consider the first case, then it would not make sense supporting "to stay hot" and hate those who think they are. On the other hand, the second scenario does not create conflicts in recommending to "stay hot" and hating people that think they are more popular than they actually are.

Example 4:

> *"And I win, I win, I always win. In the end I always win, whether it's in golf, whether it's in tennis, whether it's in life, I just always win. And I tell people I always win, because I do."*
>
> *"I want to win, and I'm not happy about not winning."*

If Trump always wins, it is contradictory to consider the case where he doesn't. Yet, the second quote might not be considering the possibility of losing, but rather highlighting how eager he is to win.

While verifying the used quotes, we could also find paraphrases or similar phrases, also said by Donald Trump, and used them to extend the dataset. For example, when searching for *"Here's a man that not only got elected, I think he's doing a really good job."*, we managed to find *"I think that he's really doing a nice job in terms of representation of this country. And he represents such a large part of the country."* and *"Well, I really like him. I think that he's working very hard."*. These similar instances were not only used to increase the

[3]Favour the legal right of a woman to choose whether or not she will have an abortion.

[4]Opposing abortion and euthanasia.

positive examples, but also to generate negative examples, because a pair of equivalent texts cannot be contradictory. Therefore, if there was a pair of documents $\langle D_1, D_2 \rangle$ known to be contradictory, and later we would find a third document D_1' which meaning and content is similar (both documents expressing the same idea) to D_1, then we would generate a new positive example $\langle D_1', D_2 \rangle$ and a new negative example $\langle D_1, D_1' \rangle$.

Besides the negative examples formed from paraphrases that we found as we looked for evidence, we also resort to the platform Factbase that provides the entire corpus of Donald Trump's public, and unedited, statements and recordings. The transcribed information is linked directly to the originating source. For this process, we filtered transcripts by keywords, like "gun control", or just opened random transcripts. Then, we extracted, from those selected transcripts, sentences where Trump would repeat the same idea.

To analyse the distribution of topics through positive and negative examples, we used the Latent Dirichlet Allocation (LDA) algorithm, for topic modeling, through python's library Scikit learn.

Before generating the topics, we consider the unique instances of all dataset examples (we don't use duplicate documents), and perform data cleaning (remove backslashes, commas and semicolons) and tokenization.

Then, to create the document word matrix, the LDA model main input, we use CountVectorizer. We configured it to ignore terms that appear in more than 95% of the documents (max_df=0.95) and terms that appear in less than 2 documents (min_df=2), to remove built-in english stopwords (stop_words='english'), to convert all words to lowercase (lowercase=True), and to impose that a word has to contain numbers and/or alphabets, of at least length 3, in order to be qualified as a word (token_pattern='[a-zA-Z0-9]{3, }').

When building the LDA model, we set the number of topics to 15 (n_components=15), the maximum learning iterations to 5 (max_iter=5), the learning method to online (learning_method='online'), the learning offset/tau_0 to 50 (learning_offset=50.), and the seed used by the random number generator to 0 (random_state=0).

Table 3 shows the obtained topics and the distribution of each topic for all positive examples (pair of two documents) and for all negative examples. The LDA model performance is out of the scope and its the denomination of each of the 15 obtained topics is not that relevant. We just want to explore the difference in frequency of topic occurrence (Diff.) between the two classes.

According to Table 3, the three most frequent topics in positive examples (pairs of contradictions) are "pro", "think" and "dont", whereas for the negative examples are "oil", "dont" and "years". The top five of most unbalanced topic distributions between the two classes are "pro", "oil", "great", "think" and "people". These values may have an impact on the experimental results, as we consider the possibility of having the model predicting based on the input topic, instead of predicting based on the relation of contradiction between two given documents.

It is important to remember that there might be other factors influencing and creating bias in this dataset, since it was not built by trained annotators.

Table 3: The number of instances of each topic that appear in positive examples (contradictions) and negative examples, and the difference in frequency of topic occurrence (Diff.) between these two classes.

Topics	Positive examples	Negative examples	Diff.
great	24	3	**21**
oil	14	**43**	**-29**
win	20	10	10
love	11	6	5
like	14	21	-7
people	31	13	**18**
dont	**39**	**39**	0
jobs	3	10	-7
think	**39**	19	**20**
pro	**41**	7	**34**
penalty	16	5	11
going	18	4	14
cancer	4	8	-4
thinker	8	2	6
years	6	**22**	-16

MultiNLIGovernment

The Multi-Genre Natural Language Inference (MultiNLI)[5] corpus [1] addresses the coverage limitation faced by other Natural Language Inference (NLI) datasets, in terms of variety of meanings, expressed in English. It is one of the largest corpora for NLI tasks, and includes ten distinct genres of written and spoken English. The wide range of styles, degrees of formality, and topics introduce greater linguistic difficulty and diversity, and make this corpus a benchmark for cross-genre domain adaptation. Moreover, the MultiNLI dataset allows to evaluate a model's ability to generate sentence representations in unfamiliar domains (cross-domain transfer learning). These characteristics and objectives meet our purpose too, as we want to mitigate whether we can take advantage of unknown and different, but still similar, document relations for our specific task of detecting contradictions.

Nine of the genres were extracted from the second release of the Open American National Corpus (OANC):

- **FACE-TO-FACE** genre uses transcriptions from the Charlotte Narrative and Conversation Collection of two-sided conversations.
- **GOVERNMENT** genre uses reports, speeches, letters, and press releases from public domain government websites.

[5] https://www.nyu.edu/projects/bowman/multinli/

- **LETTERS** genre uses letters from the Indiana Center for Intercultural Communication of Philanthropic Fundraising Discourse.
- **9/11** genre resorts to the public report from the National Commission on Terrorist Attacks Upon the United States.
- **OUP** genre uses five non-fiction works on the textile industry and child development, published by the Oxford University Press.
- **SLATE** genre uses popular culture articles from the archives of Slate Magazine.
- **TELEPHONE** genre uses transcriptions from University of Pennsylvania's Linguistic Data Consortium Switchboard corpus of two-sided telephone conversations.
- **TRAVEL** genre uses travel guides published by Berlitz Publishing.
- **VERBATIM** genre uses short posts about linguistics for non-specialists from the Verbatim archives.

The tenth genre, **FICTION**, uses several freely available works of contemporary fiction.

MultiNLIGovernment is the name we gave to the subset of MultiNLI corpus that only includes the examples of "government" genre. MultiNLI corpus has three possible labels:

- **Entailment**: relation between two sentences, a premise and a hypothesis, where the hypothesis is necessarily true or appropriate whenever the premise is true.
- **Contradiction**: relation between two sentences, a premise and a hypothesis, where the hypothesis is necessarily false or inappropriate whenever the premise is true.
- **Neutral**: relation between two sentences, a premise and a hypothesis, where none of the above conditions (entailment and contradiction) are applicable.

Since our task is to detect contradictions, using only two labels (0 and 1, respectively *not contradiction* and *contradiction*), we consider the labels "entailment" and "neutral" to be negative examples (not contradictions).

To built the dataset we used the JSON Lines format of the corpus. We filter the objects of 'genre' 'government' because, as said before, we are considering a political domain and the "government" type belongs to that field. To create an input pair, we use object's values for 'sentence1', 'sentence2' and 'gold_label'. Gold-label is the label used for classification. In the validation process of the MultiNLI corpus, when an example does not receive a three-vote consensus on any label, the golden-label is '-'. In this case, we consider it to be a negative example.

We use the python's library Pandas to use its data structure DataFrame. We create three data frames, for test, validation and train sets. We split 70% of the obtained data for training, 10% for validation, and 20% for testing, then we save each set in a tab separated text format ('test.tsv', 'dev.tsv' and 'train.tsv').

At the end, we got a total of 79,350 examples, 26,418 positives and 52,932 negatives (Table 2).

Source Task Domains

MultiNLI-

This dataset follows the same procedure as the one described for the MultiNLIGovernment dataset. However, since we aim at exploring a model's learning performance when giving data containing different relations from the ones we are using as target task domains (contradictions in a political domain), we use all the examples of the MultiNLI corpus, except the ones of "government" genre. We also split the data in two sets, 80% for training ('train.tsv') and 20% for testing ('test.tsv'). In this case, we got a total of 333,352 examples, 110,938 positives and 222,414 negatives (Table 2).

We decided to ignore the examples of "government" genre because we want to explore the behaviour of different document relations. Since we are going to use the political domain as a case study, we remove the "government" genre to only include genres that are not closely related to the political field.

US2016

US2016[6][2] is the largest corpus of annotated dialogical argumentation[7]. It comprises transcripts, collected from The American Presidency Project, of televised debates leading up to the 2016 presidential election in the United States of America: the first Republican primary debate on 6 August 2015 in Cleveland, Ohio; the first Democrat primary debate on 13 October 2015 in Las Vegas, Nevada; and the first general election debate between Hillary Clinton and Donald Trump on 26 September 2016 in Hempstead, New York. Therefore, the domain is argumentation in political debate. US2016 also includes online reactions, from Reddit, towards the three presented debates. Anyone who is a registered user in this social media platform can make posts, which leads to a greater diversity in language used, due to having people contributing from varying backgrounds, nationalities and education levels. Thus, for the online reactions, it is expected a mixed argumentative quality (rhetorical efficacy, and dialectical and logical fallaciousness) and many less well-crafted and well-signalled examples.

We are again facing cross-genre data, as we have both televised election debates and social media discussions. The US2016 corpus is a set of "argument maps" that are the result of the text annotation. It is organized in sub-corpora related to either the television debated transcripts (US2016tv) or Reddit threads (US2016reddit), for each of the three candidate debates preceding the 2016 US presidential elections (US2016R1, US2016D1 and US2016G1): US2016R1tv, US2016R1reddit, US2016D1tv, US2016D1reddit, US2016G1tv, and US2016G1reddit.

The data annotation format is based on Inference Anchoring Theory (IAT) [3]. IAT adheres to the extended Argument Interchanged Format (AIF+) standard which is a graph-based ontology that facilitates the representation of arguments. For the annotation, we have the

[6] http://www.corpora.aifdb.org/US2016

[7] Argumentation is reasoning in discourse to support a contested point of view. To resolve disagreements, arguments can be used and the reason supporting them can be tested.

following concepts:

- **Locution**: speaker identification followed by an argumentative discourse unit (ADU) which is a segmented transcribed text, that has a discrete argumentative function (right top and bottom boxes in Figure 1).
- **Transitions**: functional relation between locutions, representing the dialogue protocol (right middle box in Figure 1).
- **Illocutions**: the intended communicative function of a locution or of a transition between two locutions (middle column of boxes in Figure 1), and can be agreeing, arguing, asserting, challenging, disagreeing, questioning, restating and default illocution.
- **Proposition**: propositional content reconstructed from a locution (left top and bottom boxes in Figure 1).
- **Inference**: relation between two propositions where one supplies a reason for accepting the other (premise of an argument supporting its conclusion).
- **Conflict**: relation between two propositions where one is incompatible with the other (left middle box in Figure 1).
- **Rephrase**: relation between two propositions where one is meant to be a reformulation of another proposition.

Figure 1 shows an example of disagreement and how the argumentation is anchored in the structure of the dialogue. The blue boxes on the right are locutions and the ones on the left are the correspondent propositions.

Since we are proposing a binary classification model, we only use two labels (0 and 1, respectively, *not contradiction* and *contradiction*). In this dataset, we see both inference and rephrase relations as negative examples (not contradictions) and the relation of conflict as a positive example. A relation of conflict is linked to a disagreement illocution. A disagreement occurs when two interlocutors dispute the acceptability of a standpoint (an opinion, a belief, a proposal). They can, then, give arguments in order to resolve the disagreement while testing the reasons supporting their arguments. Hence, we are talking about the case when two people share different opinions which we see as a contradiction when considering both points of view as true.

To build our dataset we use the entire US2016 corpus in JSON format. The JSON includes a list of nodes and a list of edges. Each node is an object that has 'nodeID', 'text', 'type' and 'timestamp'. Below we present the seven possible node types:

- **L -** Locutions, excerpts from the used transcripts. In this case, the node text is the extracted snippet.
- **TA -** A transitions (link between locutions). In this case, the node text is "Default Transition".

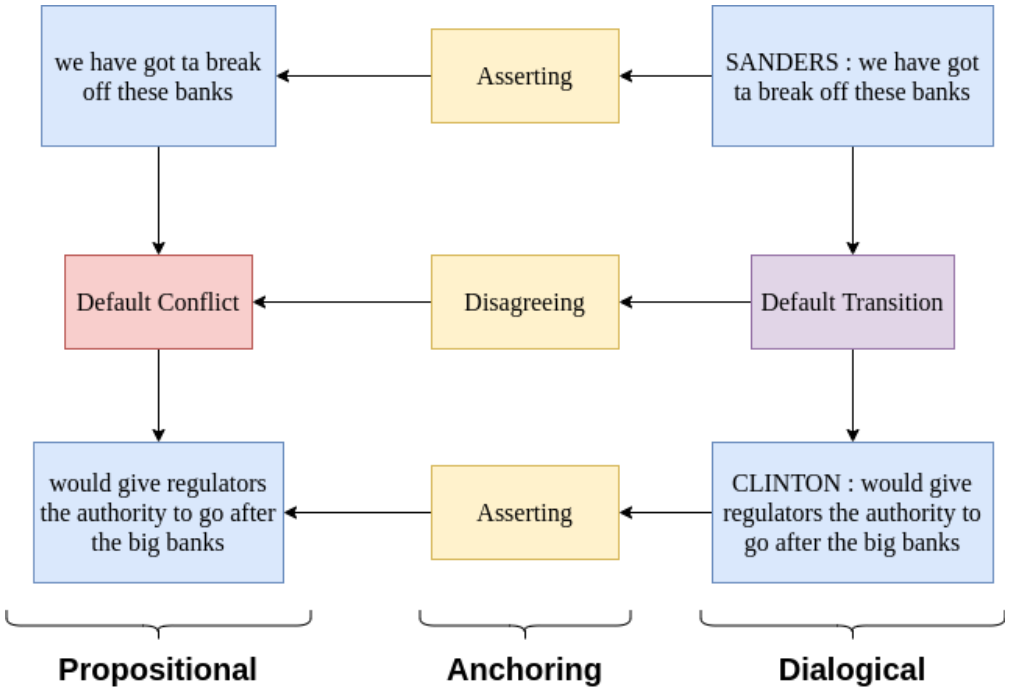

Figure 1: Diagrammatic visualisation of an example of disagreement showing how the propositional reasoning on the left is anchored in the dialogical realisation of the argument on the right. This example was taken from US2016 corpus and is available online at http://www.aifdb.org/argview/10439.

- **YA -** Illocutions that link locutions to propositions. In this case, the node text can be "Agreeing", "Arguing", "Asserting", "Challenging", "Default Illocuting", "Disagreeing", "Restating", and "Questioning".
- **I -** Proposition. In this case, the node text is a processed locution.
- **RA -** Relation of inference which is a link between two propositions where one gives a reason for the other to be accepted. In this case, the node text is "Default Inference".
- **CA -** Relation of conflict which is a link between two propositions where one is an incompatible alternative to another. In this case, the node text is "Default Conflict".
- **MA -** Relation of rephrase which is a link between two propositions where one reformulates the other. In this case, the node text is "Default Rephrase".

An edge represents the connection between two nodes. It has an 'edgeID', 'fromID' (id of the source node), 'toID' (id of the destination node), and 'formID' (which is always "null"). Therefore, to form an input pair, we get the 'nodeID' from a node of type CA. That node would be the red box from Figure 1. Then, we need two edges, one that has the CA node as 'fromID' and other that has it as 'toID'. There will be two distinct edges with the CA node as 'toID' because, besides the proposition node (blue box in Figure 1 upper left corner), there is always an illocution node (yellow boxes in the middle of Figure 1) anchoring the

propositional reasoning to the dialogical act (linking the boxes on the right side to the boxes on the left side, in Figure 1).

From the edges coming and leaving the CA node, we get the proposition nodes (type I). The text of those two nodes are the sentences of the input and, in this case, the label will be 1 (contradiction). For the negative examples (label 0), we follow the same procedure, but resorting to relation nodes of type RA and MA.

Regarding the count of propositional relations, this corpus has 2830 inference relations, 942 Conflict relations and 764 Rephrase relations. In our dataset we keep a balanced ratio of positive and negative examples by using 941 conflict relations (one of the conflict relations had an error since it was missing a node linking to the CA node) and getting a total of 941 examples from both inference and rephrase relations. We give priority to the examples of rephrase because in this relation it is more clear that the two sentences do not conflict, since one is basically paraphrasing the other. However, we ended up not using all the 764 rephrase instances because there were ones that were too small and simple to matter, like the following examples:

1. "CHINA. Mexico" and "Mexico. CHINA"
2. "flat tax" and "I 've advocated a proportional tax system"
3. "X for TRUMP's family" and "X"
4. "Wrong. Wrong wrong" and "Wrong"

Thus, for the negative examples, we only consider pairs in which each sentence has at least four words.

Finally, we shuffle all obtained examples and split them in two sets, 80% for training ('train.tsv') and 20% for validation ('dev.tsv').

ArgumentativeMicrotext

The Argumentative Microtext Corpus[8] is the result of argumentation mining which involves capturing the different aspects of the argumentation structure of a text (central claim, supporting reasons, possible objections, counters to the objections). Thus, the argumentation structure of a text is a graph representation, depicting the argumentative relation between the propositions.

The corpus provides short texts which are responses to trigger questions and is divided in two parts. The first part [4] has 122 texts: 89 texts collected in a controlled text generation experiment based on a list of controversial questions[9], and 23 texts written by Andreas Peldszus, as a "proof of concept" for the idea, and with the purpose of teaching and testing students argumentative analysis. The second part [5] was produced by a crowdsourcing experiment, also based on a list of trigger questions[10], resulting in 171 more texts.

[8] http://angcl.ling.uni-potsdam.de/resources/argmicro.html
[9] https://github.com/peldszus/arg-microtexts/blob/master/topics_triggers.md
[10] https://github.com/discourse-lab/arg-microtexts-part2/blob/master/topics_triggers.md

The annotation scheme is based on the idea of modeling the argumentation as a hypothetical discussion between the proponent, who presents and defends its claims, and the opponent, who question and criticizes them. However, each microtext of this corpus only has one author that not only gives reasons in favour of the main claim, but may also take counter-arguments into consideration. Figure 2 shows the schematic diagram of one of the corpus microtexts. The nodes represent propositions extracted from text segments (the grey boxes). The shape of the nodes indicates the role of the correspondent proposition: round nodes are in favour of the claim and square nodes against it. The arrowhead, circle-head and square-head edges represent, respectively, a supporting move, an attacking move of rebuttal (challenging the acceptability of a proposition), and an attacking move of undercutter (challenging the acceptability of an inference between two propositions). In the example in Figure 2, the fourth segment rebuts the first segment, and this rebutting move is undercut by the fifth segment.

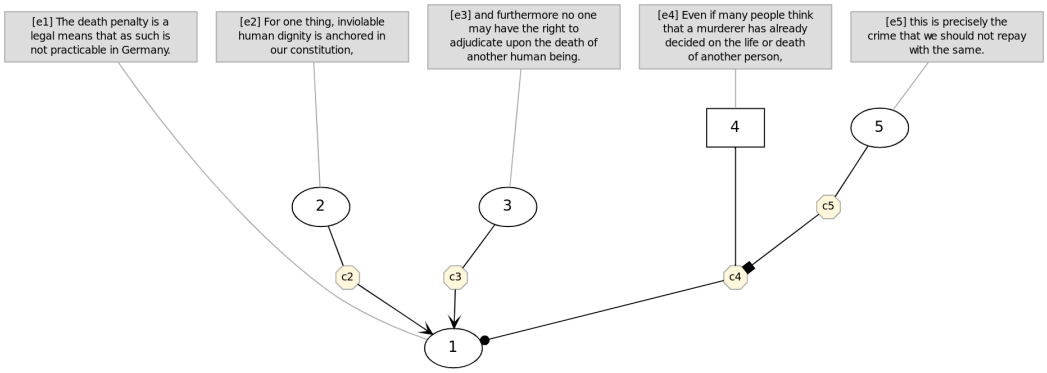

Figure 2: Microtext and argumentation graph. This example was taken from the first part of the Argumentative Microtext Corpus and is available online at https://github.com/peldszus/arg-microtexts/blob/master/corpus/en/micro_b006.pdf

To build our dataset, we resort to the corpus XML format. The XML representing a microtext graph has elementary discourse unit (EDU) elements and argumentative discourse units (ADU) elements, which are EDUs that serve as independent arguments to the argumentation. The EDU element's content is character data (CDATA) presenting a text segment, and the ADU element has an attribute 'type' that says if the text segment supports ("type="pro"") or refutes/attacks ("type="opp"") the main claim. The XML also has edge elements with four attributes: "id", "src" (element from where the edge is leaving), "trg" (edge destination element), and "type" (type of link between two XML elements). We are interested in four edge types:

- **seg -** an edge of this type connects an EDU element to its correspondent ADU element.
- **sup -** an edge of this type represents a relation of support, connecting an ADU element to another ADU element, with the objective of increasing the credibility of the second ("trg" element) by providing a reason ("scr" element) for accepting it.

- **reb -** an edge of this type represents a rebutter (attack between propositions), connecting an ADU element to another ADU element, using the first ("src" element) to refute or weaken the force of the second ADU ("trg" element).

- **und -** an edge of this type represents an undercutter (attack to the relation between propositions), connecting an ADU element to an edge element, using the first ("src" element) to challenge the second ("trg" element).

In this scenario, we will use the support relations as negative examples and the attack relations as positive examples, because a supporting statement aims to increase the strength of the argument and a attacking statement aims to refute the target.

Regarding rebutters, if the two elements of this relation have other elements directly supporting them, we concatenate those text segments to give more context to the document used in the input pair. In the example from Figure 3, where the trigger question is "Should the statutory retirement age remain at 63 years in the future?", text segments two and three support the first, and the fourth text segment rebuts the first, so, here, the input pair with label 1 (contradiction), would be ⟨ *"The implementation of retirement at 63 is no longer socially sustainable, as the population in Germany has, viewed demographically, a disproportionate number of old people, and constantly declining birth rates are being recorded." ; "Admittedly the number of immigrants is constantly rising in Germany"*⟩. The first document of this input pair talks about the struggle of retiring people at the age of 63 because the elderly population in Germany is increasing and the birth rate is decreasing. By doing so, Germany would lose many employees, ending up missing workers. The second document of the pair attacks the first by stating that the number of immigrants is rising which can help covering the lack of employees created due to the possibility of retiring at 63. Thus, the issue of the first document would no longer be a problem. That is why we see this as a contradiction, because if the counter argument holds, then the first claim loses its strength and/or meaning. Still, it is not truly a contradiction, since it is not certain that the fact expressed in the second statement is a solution for the problem presented in the first statement, hence both statements might co-exist (be simultaneously true).

Regarding undercutters, since the target element ("trg") is a relation (an edge) between two elements, from that relation we get the two linked elements/nodes. Then, we only use the ones that are of a different type ("pro" or "opp") as compared to the undecutter's source element ("src"). In the example from Figure 3, the fifth text segment undercuts the relation of attack between segments one and four. Since the fourth text segment is of proponent type ("type="pro"") ant the fifth text segment is of opponent type ("type="opp""), the input pair with label 1 would be ⟨ *"Admittedly the number of immigrants is constantly rising in Germany" ; "but without sufficient, well-qualified junior employees there is hardly a possibility for adequate pension financing."*⟩. The second document of this input pair shows that, although the number of immigrants is increasing, we shouldn't take it for granted since number is not the only factor, the employees qualification and skills is equally important. Even though these two statement are of different types (one supports the main claim and other is against it), the contradiction is barely understandable. It can be explained through the fact that, besides the lack of context, while a rebutter can be seen as an argument for

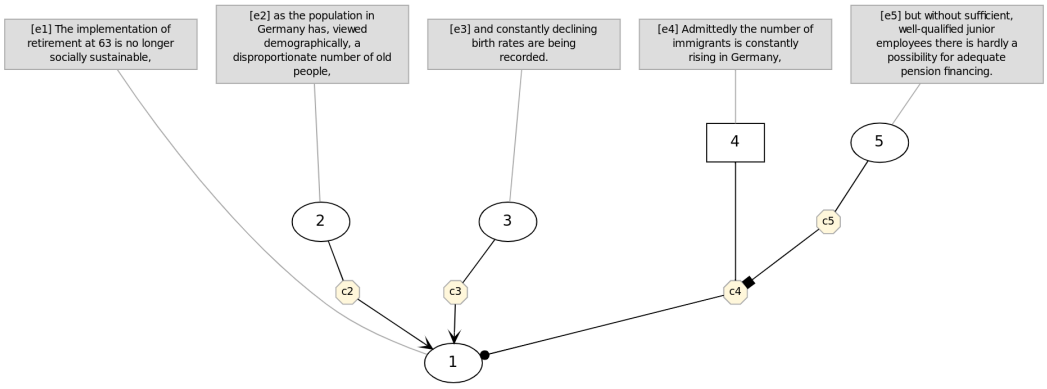

Figure 3: Microtext and argumentation graph. This example was taken from the first part of the Argumentative Microtext Corpus and is available online at https://github.com/peldszus/arg-microtexts/blob/master/corpus/en/micro_k017.pdf

the negation of the proposition under attack, an undercutter does not challenge the validity of a proposition, but challenges the acceptability of an inference between two propositions.

After gathering the input pairs (a total of 1133 examples, 403 positive and 730 negative), we shuffle them and split them in two sets, 80% for training ('train.tsv') and 20% for validation ('dev.tsv').

ArgumentEssays

Argument Annotated Essays[11][6] is a corpus of 402 persuasive essays annotated with discourse-level argumentation structures modeled as a connected tree. The major claim (author's standpoint) is the root node and is usually found in the essay's introduction, that describes the controversial topic. The arguments, presented in individual paragraphs of an essay, can support or attack the major claim. One argument consists of a claim (central component) and, at least, one premise (reason of the argument). Each claim has a stance that can be either "for" (supporting argument) or "against" (attacking argument) the major claim. A premise can be used to justify a claim (relation of support) or to refute it (relation of attack).

In contrast to the previous corpus annotation scheme (described in Section ArgumentativeMicrotext) that splits a microtext/answer in different text segments, but does not process them, this corpus annotation has stricter argument component boundary rules, such as ignoring "shell language", phrases like "Another reason is that" or "I am strongly convinced", that are not relevant for the argument's content.

The corpus has, for each essay, a ".txt" file, that is the entire and unchanged essay, and an annotation file (e.g., "essay001.ann") which contains:

[11] https://www.informatik.tu-darmstadt.de/ukp/research_6/data/argumentation_mining_1/argument_annotated_essays_version_2/index.en.jsp

- **Entities -** They can be "MajorClaim", "Claim" or "Premise". One line representing an entity has the entity id, the entity tag, character positions in the essay ".txt" file where the entity starts and ends, and the entity content, as shown in the following example: *T1 MajorClaim 503 575 we should attach more importance to cooperation during primary education*

- **Relations -** They can assume the values "supports" or "attacks". One line representing a relation has the relation id, the relation value, the relation's source argument/entity id (that can only be of a premise or of a claim), and the relation's target argument/entity id (that can be of a premise, claim or major claim), as shown in the following example: *R1 supports Arg1:T4 Arg2:T3*

- **Attributes -** They are the claim's stance and can be "For" or "Against". One line representing an attribute has the attribute id, the tag "Stance", the claim id, and the value, as shown in the following example: *A2 Stance T7 Against*

When building our dataset, we consider the positive examples to be the pairs of major claim and claim with stance "Against", and the pairs of two entities that share a relation of attack. On the other hand, the negative examples will be the pairs of major claim and claim of stance "For", and the pairs of two entities that share a relation of support. After gathering the input pairs (a total of 6673 examples, 715 positive and 5958 negative), we shuffle them and split them in two sets, 80% for training ('train.tsv') and 20% for validation ('dev.tsv').

W2E

In our research we wanted to confirm that a successful transfer learning is caused by the relevance of document relationships depicted in a source dataset, and not only by the increase of training examples. Therefore, we collected data from W2E[12] [7]. With this dataset we expect to achieve negative transfer learning as the documents relationships are not related with the ones in our target task datasets. While the target task domains present contradiction in a political domain, W2E dataset was designed for topic detection and tracking.

W2E is a Worldwide-Event benchmark dataset for topic detection and tracking, containing 207,722 news articles written in English, from 52 mass media channels (e.g., CNN, BBC, Fox News, etc.). The news articles cover a large set of 4,501 popular events, within the entire year of 2016, each belonging to one out of 10 categories: "Sport", "Science and technology", "Politics and elections", "Law and crime", "International relations", "Health and medicine", "Disasters and accidents", "Business and economy", "Arts and culture", "Armed conflicts and attacks". To select the events, the authors resorted to Wikipedia's Current Event portal[13] (WCEP) which contains short summaries of events. They chose the year of 2016 because of the variety of popular long-run stories in that period, such as the US presidential election, UK's European Union membership referendum, Middle East wars, the Summer Olympics, and disasters in North America.

[12] https://sites.google.com/site/w2edataset/
[13] https://en.wikipedia.org/wiki/Portal:Current_events

Table 4: Distribution of text categories in an input pair for each class. The short form for each category is "S" for "Sport", "ST" for "Science and technology", "PE" for "Politics and elections", and "HM" for "Health and medicine".

Class	Categories in a pair	Number of examples
Positive	S-ST	400
	S-PE	400
	S-HM	400
	ST-PE	400
	ST-HM	400
	PE-HM	400
Negative	S-S	600
	ST-ST	600
	PE-PE	600
	HM-HM	600

Since different news sources were used, W2E includes distinct views of the same event. Different news articles regarding the same issue belong to the same topic which is assigned to a more generic category. There is a total of 2,015 topics. W2E provides, for each topic, the topic's information (id, category, and description) and its correspondent events (date of the event, event's summary, and search query). The fields in each line are tab-separated. Next, we have an example of a topic and its events:

"

*TOPIC-7 Disasters and accidents *** Kollam temple fire*

2016-04-10 A fire occurs at a Hindu temple in the Kollam district ...

Kollam Kerala India Hindu temple fire

2016-04-11 Five workers from the company that supplied fireworks to the Puttingal Temple ...

worker fireworks Puttingal Temple dead

"

Thus, TOPIC-7 belongs to the "Disasters and accidents" category, its description is "Kollam temple fire", and it has two events (one happening on 2016-04-10, and the other on 2016-04-11).

When building our dataset, we consider the positive examples to be a pair of summaries of two news articles from topics assigned to distinct categories, and from topics assigned to the same category for negative examples. We only extracted data from four categories that seemed more distant regarding the content context ("Sport", "Science and technology", "Politics and elections", "Health and medicine"). For both classes, we have the same distribution for each category, as depicted in Table 4.

After gathering the input pairs (a total of 4800 examples, 2400 positive and 2400 negative), we shuffle them and split them in two sets, 80% for training ('train.tsv') and 20% for validation ('dev.tsv').

References

[1] Adina Williams, Nikita Nangia, e Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. Em *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, páginas 1112–1122. Association for Computational Linguistics, 2018.

[2] Jacobus Visser, Barbara Konat, Rory Duthie, Marcin Koszowy, Katarzyna Budzynska, e Chris Reed. Argumentation in the 2016 us presidential elections: Annotated corpora of television debates and social media reaction. *Language Resources and Evaluation*, 54:123–154, March 2019.

[3] Katarzyna Budzynska e Chris Reed. Whence inference. *University of Dundee Technical Report*, 2011.

[4] Andreas Peldszus e Manfred Stede. An annotated corpus of argumentative microtexts. Em *Argumentation and Reasoned Action: Proceedings of the 1st European Conference on Argumentation, Lisbon 2015 / Vol. 2*, páginas 801–815, London, 2016. College Publications.

[5] Maria Skeppstedt, Andreas Peldszus, e Manfred Stede. More or less controlled elicitation of argumentative text: Enlarging a microtext corpus via crowdsourcing. Em *Proceedings of the 5th Workshop on Argument Mining*, páginas 155–163, Brussels, Belgium, November 2018. Association for Computational Linguistics.

[6] Christian Stab e Iryna Gurevych. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659, September 2017.

[7] Tuan-Anh Hoang, Khoi Duy Vo, e Wolfgang Nejdl. W2e: A worldwide-event benchmark dataset for topic detection and tracking. Em *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, página 1847–1850, New York, NY, USA, 2018. Association for Computing Machinery.