Fill the Gap! Analyzing Implicit Premises between Claims from Online Debates

Filip Boltužić and **Jan Šnajder**
University of Zagreb, Faculty of Electrical Engineering and Computing
Text Analysis and Knowledge Engineering Lab
Unska 3, 10000 Zagreb, Croatia
{filip.boltuzic,jan.snajder}@fer.hr

Abstract

Identifying the main claims occurring across texts is important for large-scale argumentation mining from social media. However, the claims that users make are often unclear and build on implicit knowledge, effectively introducing a gap between the claims. In this work, we study the problem of matching user claims to predefined main claims, using implicit premises to fill the gap. We build a dataset with implicit premises and analyze how human annotators fill the gaps. We then experiment with computational claim matching models that utilize these premises. We show that using manually-compiled premises improves similarity-based claim matching and that premises generalize to unseen user claims.

1 Introduction

Argumentation mining aims to extract and analyze argumentation expressed in natural language texts. It is an emerging field at the confluence of natural language processing (NLP) and computational argumentation; see (Moens, 2014; Lippi and Torroni, 2016) for a comprehensive overview.

Initial work on argumentation mining has focused on well-structured, edited text, such as legal text (Walton, 2005) or scientific publications (Jiménez-Aleixandre and Erduran, 2007). Recently, the focus has also shifted to argumentation mining from social media texts, such as online debates (Cabrio and Villata, 2012; Habernal et al., 2014; Boltužić and Šnajder, 2014), discussions on regulations (Park and Cardie, 2014), product reviews (Ghosh et al., 2014), blogs (Goudas et al., 2014), and tweets (Llewellyn et al., 2014; Bosc et al., 2016). Mining arguments from social media can uncover valuable insights into peoples' opinions; in this context, it can be thought of as a sophisticated opinion mining technique – one that seeks to uncover the reasons for opinions and patterns of reasoning. The potential applications of social media mining are numerous, especially when done on a large scale.

In comparison to argumentation mining from edited texts, there are additional challenges involved in mining arguments from social media. First, social media texts are more noisy than edited texts, which makes them less amenable to NLP techniques. Secondly, users in general are not trained in argumentation, hence the claims they make will often be unclear, ambiguous, vague, or simply poorly worded. Finally, the arguments will often lack a proper structure. This is especially true for short texts, such as microbloging posts, which mostly consist of a single claim.

When analyzing short and noisy arguments on a large scale, it becomes crucial to identify identical but differently expressed claims across texts. For example, summarizing and analyzing arguments on a controversial topic presupposes that can identify and aggregate identical claims. This task has been addressed in the literature under the name of *argument recognition* (Boltužić and Šnajder, 2014), *reason classification* (Hasan and Ng, 2014), *argument facet similarity* (Swanson et al., 2015; Misra et al., 2015), and *argument tagging* (Sobhani et al., 2015). The task can be decomposed into two subtasks: (1) identifying the main claims for a topic and (2) matching each claim expressed in text to claims identified as the main claims. The focus of this paper is on the latter.

The difficulty of the claim matching task arises from the existence of a gap between the user's claim and the main claim. Many factors contribute to the gap: linguistic variation, implied commonsense knowledge, or implicit premises from the beliefs and value judgments of the person making the

Proceedings of the 3rd Workshop on Argument Mining, pages 124–133,
Berlin, Germany, August 7-12, 2016. ©2016 Association for Computational Linguistics

User claim:	*Now it is not taxed, and those who sell it are usually criminals of some sort.*
Main claim:	*Legalized marijuana can be controlled and regulated by the government.*
Premise 1:	*If something is not taxed, criminals sell it.*
Premise 2:	*Criminals should be stopped from selling things.*
Premise 3:	*Things that are taxed are controlled and regulated by the government.*

Table 1: User claim, the matching main claim, and the implicit premises filling the gap.

claim; the latter two effectively make the argument an *enthymeme*. In Table 1, we give an example from the dataset of Hasan and Ng (2014). Here, a user claim from an online debate was manually matched to a claim previously identified as one of the main claims on the topic of marijuana legalization. Without additional premises, the user claim does not entail the main claim, but the gap may be closed by including the three listed premises.

Previous annotation studies (Boltužić and Šnajder, 2014; Hasan and Ng, 2014; Sobhani et al., 2015) demonstrate that humans have little difficulty in matching two claims, suggesting that they are capable of filling the premise gap. However, current machine learning-based approaches to claim matching do not account for the problem of implicit premises. These approaches utilize linguistic features or rely on textual similarity and textual entailment features. From an argumentation perspective, however, these are shallow features and their capacity to bridge the gap opened by implicit premises is limited. Furthermore, existing approaches lack the explanatory power to explain why (under what premises) one claim can be matched to the other. Yet, the ability to provide such explanations is important for apprehending arguments.

In this paper, we address the problem of claim matching in the presence of gaps arising due to implicit premises. From an NLP perspective, this is a daunting task, which significantly surpasses the current state of the art. As a first step in better understanding of the task, we analyze the gap between user claims and main claims from both a data and computational perspective. We conduct two studies. The first is an annotation study, in which we analyze the gap, both qualitatively and quantitatively, in terms of how people fill it. In the second study, we focus on the computational models for claim matching with implicit premises, and gain preliminary insights into such models could benefit from the use of implicit premises.

To the best of our knowledge, this is the first work that focuses on the problem of implicit premises in argumentation mining. Besides reporting on the experimental results of the two studies, we also describe and release a new dataset with human-provided implicit premises. We believe our results may contribute to a better understanding of the premise gap between claims.

The remainder of the paper is structured as follows. In the next section, we briefly review the related work on argumentation mining. In Section 3 we describe the creation of the implicit premises dataset. We describe the results of the two studies in Section 4 and Section 5, respectively. We conclude and discuss future work in Section 6.

2 Related Work

Work related to ours comes from two broad strands of research: argumentation mining and computational argumentation. Within argumentation mining, a significant effort has been devoted to the extraction of argumentative structure from text, e.g., (Walton, 2012; Mochales and Moens, 2011; Stab and Gurevych, 2014; Habernal and Gurevych, 2016)). One way to approach this problem is to classify the text fragments into *argumentation schemes* – templates for typical arguments. Feng and Hirst (2011) note that identifying the particular argumentation scheme that an argument is using could help in reconstructing its implicit premises. As a first step towards this goal, they develop a model to classify text fragments into five most frequently used Walton's schemes (Walton et al., 2008), reaching 80–95% pairwise classification accuracy on the Araucaria dataset.

Recovering argumentative structure from social media text comes with additional challenges due to the noisiness of the text and the lack of argumentative structure. However, if the documents are sufficiently long, argumentative structure could in principle be recovered. In a recent study on social media texts, Habernal and Gurevych (2016) showed that (a slightly modified) Toulmin's argumentation model may be suitable for short documents, such as article comments or forum posts. Using sequence labeling, they identify the claim, premise, backing, rebuttal, and refutation components, achieving a token-level F1-score of 0.25.

Unlike the work cited above, in this work we do not consider argumentative structure. Rather, we focus on short (mostly single-sentence) claims, and

the task of matching a pair of claims. The task of claim matching has been tackled by Boltužić and Šnajder (2014) and Hasan and Ng (2014). The former frame the task as a supervised multi-label problem, using textual similarity- and entailment-based features. The features are designed to compare the user comments against the textual representation of main claims, allowing for a certain degree of topic independence. In contrast, Hasan and Ng frame the problem as a (joint learning) supervised classification task with lexical features, effectively making their model topic-specific.

Both approaches above are supervised and require a predefined set of main claims. Given a large-enough collection of user posts, there seem to be at least two ways in which main claims can be identified. First, they can be extracted manually. Boltužić and Šnajder (2014) use the main claims already identified as such on an online debating platform, while Hasan and Ng (2014) asked annotators to group the user comments and identify the main claims. The alternative is to use unsupervised machine learning and induce the main claims automatically. A middle-ground solution, proposed by Sobhani et al. (2015), is to first cluster the claims, and then manually map the clusters to main claims. In this work, we assume that the main claims have been identified using any of the above methods.

Claim matching is related to the well-established NLP problems: *textual entailment* (TE) and *semantic textual similarity* (STS), both often tackled as shared tasks (Dagan et al., 2006; Agirre et al., 2012). Boltužić and Šnajder (2014) explore using outputs from STS and TE in solving the claim matching problem. Cabrio and Villata (2012) use TE to determine support/attack relations between claims. Boltužić and Šnajder (2015) consider the notion of *argument similarity* between two claims. Similarly, Swanson et al. (2015) and Misra et al. (2015) consider *argument facet similarity*.

The problem of implicit information has also been tackled in the computational argumentation community. Work closest to ours is that of Wyner et al. (2010), who address the task of inferring implicit premises from user discussions. They annotate implicit premises in *Attempto Controlled English* (Fuchs et al., 2008), define propositional logic axioms with annotated premises, and extract and explain policy stances in discussions. In our work, we focus on the NLP approach and work with implicit premises in textual form.

Topic	# claim pairs	# main claims
Marijuana (MA)	125	10
GayRights (GR)	125	9
Abortion (AB)	125	12
Obama (OB)	125	16

Table 2: Dataset summary.

3 Data and Annotation

The starting point of our study is the dataset of Hasan and Ng (2014). The dataset contains user posts from a two-side online debate platform on four topics: "Marijuana" (MA), "Gay rights" (GR), "Abortion" (AB), and "Obama" (OB). Each post is assigned a stance label (*pro* or *con*), provided by the author of the post. Furthermore, each post is split up into sentences and each sentence is manually labeled with a single claim from a predefined set of main claims, different for each topic. Note that all sentences in the dataset are matched against exactly one main claim. Hasan and Ng (2014) report substantial levels of inter-annotator agreement (between 0.61 and 0.67, depending on the topic).

Our annotation task extends this dataset. We formulate the task as a "fill-the-gap" task. Given a pair of previously matched claims (a user claim and a main claim), we ask the annotators to provide the premises that bridge the gap between the two claims. No further instructions were given to the annotators; we hoped that they would resort to common-sense reasoning and effectively reconstruct the deductive steps needed to entail the main claim from the user claim. The annotators were also free to abstain from filling the gap, if they felt that the claims cannot be matched; we refer to such pairs as *Non-matching*. If no implicit premises are required to bridge the gap (the two claims are paraphrases of each other), then the claim pair is annotated as *Directly linked*.

We hired three annotators to annotate each pair of claims. The order of claim pairs was randomized for each annotator. We annotated 125 claims pairs for each topic, yielding a total of 500 gap-filling premise sets. Table 2 summarizes the dataset statistics. An excerpt from the dataset is given in Table 3. We make the dataset freely available.[1]

[1] Available under the CC BY-SA-NC license from `http://takelab.fer.hr/argpremises`

Claim pair	Annotation
User claim: *Obama supports the Bush tax cuts. He did not try to end them in any way.* **Main claim:** *Obama destroyed our economy.*	**P1:** *Obama continued with the Bush tax cuts.* **P2:** *The Bush tax cuts destroyed our economy.*
User claim: *What if the child is born and there is so many difficulties that the child will not be able to succeed in life?* **Main claim:** *A fetus is not a human yet, so it's okay to abort.*	Non-matching
User claim: *Technically speaking, a fetus is not a human yet.* **Main claim:** *A fetus is not a human yet, so it's okay to abort.*	Directly linked

Table 3: Examples of annotated claim pairs.

4 Study I: Implicit Premises

The aim of the first study is to analyze how people fill the gap between the user's claim and the corresponding main claim. We focus on three research questions. The first concerns the variability of the gap: to what extent do different people fill the gap in different ways, and to what extent the gaps differ across topics. Secondly, we wish to characterize the gap in terms of the types of premises used to fill it. The third question is how the gap relates to the more general (but less precise) notion of textual similarity between claims, which has been used for claim matching in prior work.

4.1 Setup and Assumptions

To answer the above questions, we analyze and compare the gap-filling premise sets in the dataset of implicit premises from Section 3. We note that, by doing so, we inherit the setup used by Hasan and Ng (2014). This seems to raise three issues.

First, the main claim to which the user claim has been matched to need not be the correct one. In such cases, it would obviously be nonsensical to attempt to fill the gap. We remedy this by asking our annotators to abstain from filling the gap if they felt the two claims do not match. Moreover, considering that the agreement on the claim matching task on this dataset was substantial (Hasan and Ng, 2014), we expect this to rarely be the case.

The second issue concerns the granularity of the main claims. Boltužić and Šnajder (2015) note that the level of claim granularity is to a certain extent arbitrary. We speculate that, on average, the more general the main claims are, the fewer the number of main claims for a given topic and the bigger the gaps between the user-provided and main claims.

	A1	A2	A3	Avg.
Avg. # premises	3.6	2.6	2.0	2.7 ± 0.7
Avg. # words	26.7	23.7	18.6	23.0 ± 3.4
Non-matching (%)	1.2	3.6	14.5	6.4 ± 5.8

Table 4: Gap-filling parameters for the three annotators.

Finally, we note that each gap was not filled by the same person who identified the main claim, which in turn is not the original author of the claim. Therefore, it may well be that the original author would have chosen a different main claim, and that she would commit to a different set of premises than those ascribed to by our annotators.

Considering the above, we acknowledge that we cannot analyze the *genuine* implicit premises of the claim's author. However, under the assumption that the main claim has been correctly identified, there is a gap that can be filled with *sensible* premises. Depending on how appropriate the chosen main claim was, this gap will be larger or smaller.

4.2 Variability in Gap Filling

We are interested in gauging the variability of gap filling across the annotators and topics. To this end, we calculate the following quantitative parameters: the average number of premises, the average number of words in premises, and the proportion of non-matched claim pairs.

Table 4 shows that there is a substantial variance in these parameters for the three annotators. The average number of premises per gap is 2.7 and the average number of words per gap is about 23, yielding the average length of about 9 words per premise. We also computed the word overlap between the three annotators: 8.51, 7.67, and 5.93 for annotator pairs A1-A2, A1-A3, and A2-A3, respectively. This indicates that, on average, the premise sets overlap in just 32% of the words. The annotators A1 and A2 have a higher word overlap and use more words to fill the gap. Also, A1 and A2 managed to fill the gap for more cases than A3, who much more often desisted from filling the gap. An example where A1 used more premises than A3 is shown in Table 5.

Table 6 shows the gap-filling parameters across topics. Here the picture is more balanced. The least number of premises and the least number of words per gap are used for the AB topic. The GR

User claim:	*It would be loads of empathy and joy for about 6 hours, then irrational, stimulant-induced paranoia. If we can expect the former to bring about peace on Earth, the latter would surely bring about WWIII.*
Main claim:	*Legalization of marijuana causes crime.*
A1 Premise 1:	*Marijuana is a stimulant.*
A1 Premise 2:	*The use of marijuana induces paranoia.*
A1 Premise 3:	*Paranoia causes war.*
A1 Premise 4:	*War causes aggression.*
A1 Premise 5:	*Aggression is a crime.*
A1 Premise 6:	*"WWIII" stands for the Third World War.*
A3 Premise 1:	*Marijuana leads to irrational paranoia which can lead to commiting a crime.*

Table 5: User claim, the matching main claim, and the implicit premise(s) filling the gap provided by two different annotators.

	Topic				
	MA	GR	AB	OB	Avg.
Avg. # premises	2.8	2.8	2.5	2.8	2.7 ± 0.1
Avg. # words	23.6	24.9	19.1	23.4	22.8 ± 2.2
Non-matching (%)	5.9	6.8	4.6	4.3	5.4 ± 1.0

Table 6: Gap-filling parameters for the four topics.

topic contained the most (about 7%) claim pairs for which the annotators desisted from filing the gap.

4.3 Gap Characterization

We next make a preliminary inquiry into the of nature of the gap. To this end, we characterize the gap in terms of the individual premises that are used to fill it. At this point we do not look at the relations between the premises (the argumentative structure); we leave this for future work.

Our analysis is based on a simple ad-hoc typology of premises, organized along three dimensions: premise type (fact, value, or policy), complexity (atomic, implication, or complex), and acceptance (universal or claim-specific). The intuition behind the latter is that some premises convey general truths or widely accepted beliefs, while others are specific to the claim being made, and embraced only by the supporters of the claim in question.

We (the two authors) manually classified 50 premises from the MA topic into the above categories and averaged the proportions. The kappa-agreement is 0.42, 0.62, and 0.53 for the premise type, complexity, and acceptance, respectively. Factual premises account for the large majority (85%) of cases, value premises for 9%, and policy premises for 6%. Most of the gap-filling premises are atomic (77%), while implication and other complex types constitute 16% and 7% of cases, respectively. In terms of acceptance, premises are well-balanced: universal and claim-specific premises account for 62% and 38% of cases, respectively.

We suspect that the kind of the analysis we did above might be relevant for determining the overall strength of an argument (Park and Cardie, 2014). An interesting venue for future work would be to carry out a more systematic analysis of premise acceptance using the complete dataset, dissected across claims and topics, and possibly based on surveying a larger group of people.

4.4 Semantic Similarity between Claims

Previous work addressed claim matching as a semantic textual similarity task (Swanson et al., 2015; Misra et al., 2015; Boltužić and Šnajder, 2015). It is therefore worth investigating how the notion of semantic similarity relates to the gap between two claims. We hypothesize that the textual similarity between two claims will be negatively affected by the size of the gap. Thus, even though the claims are matching, if the gap is too big, similarity will not be high enough to indicate the match.

To verify this, we compare the semantic similarity score between each pair of claims against its gap size, characterized by the number of premises required to fill the gap, averaged across the three annotators. To obtain a reliable estimate of semantic similarity between claims, instead of computing the similarity automatically, we rely on human-annotated similarity judgments. We set up a crowdsourcing task and asked the workers to judge the similarity between 846 claim pairs for the MA topic. The task was formulated as a question "*Are two claims talking about the same thing?*", and judgments were made on a scale from 1 ("not similar") to 6 ("very similar"). Each pair of claims received five judgments, which we averaged to obtain the gold-similarity score. The average standard deviation is 1.2, indicating good agreement.

The Pearson correlation coefficient between the similarity score and the number of premises filling the gap for annotators A1, A2, and A3 is -0.30, -0.28, and -0.14, respectively. The correlation between the similarity score and the number of premises averaged across the annotators is -0.22 ($p<0.0001$). We conclude that there is a statistically significant, albeit weak negative relationship between semantic similarity and gap size.

5 Study II: Claim Matching Model

In this section we focus on claim matching models with implicit premises. In the previous section, we demonstrated that the degree of similarity between matched claims varies and is negatively correlated with the number of gap-filling premises. This result directly suggests that the similarity scores for matched claims could be increased by reducing the size of the gap. Furthermore, we expect that the size of the gap can be effectively reduced by including premises in the similarity computation.

Motivated by these insights, we conduct a preliminary study on the use of implicit premises in claim matching. The study is also motivated by our long-term goal to develop efficient models for recognizing main claims in social media texts. Given a user's claim, the task is to find the main claim from a predefined set of claims to which the user's claim matches the best. We address three research questions: (1) whether and how the use of implicit premises improves claim matching, (2) how well do the implicit premises generalize, and (3) could the implicit premises be retrieved automatically.

5.1 Experimental Setup

The claim matching task can be approached in a supervised or unsupervised manner. We focus here on the latter, based on semantic similarity between the claims and the premises. We think unsupervised claim matching provides a more straightforward and explicit way of incorporating the implicit premises. Furthermore, the unsupervised approach better corresponds to the very idea of argumentation, where claims and premises are compared to each other and combined to derive other claims.

Dataset. We use the implicit premise dataset from Section 3, consisting of 125 claim pairs for each of the four topics. We use the gap-filling premise sets from annotator A1, who on average has provided the largest number of implicit premises. We refer to this dataset as the *development set*. In addition, we sample and additional *test set* consisting of 125 pairs for each topic from the dataset of Hasan and Ng (2014); for claim pairs from this set we have no implicit premises.

Semantic similarity. We adopt the distributional semantics approach (Turney and Pantel, 2010) to computing semantic textual similarity. We rely on distributed representation based on the neural network skip-gram model of Mikolov et al. (2013a).[2] We represent the texts of the claims and the premises by summing up the distributional vectors of their individual words, as the semantic composition of short phrases via simple vector addition has been shown to work well (Mikolov et al., 2013b). We measure claim similarity using cosine distance between two vectors.

Inspired by (Cabrio et al., 2013; Boltužić and Šnajder, 2014), we also attempted to model claim matching using textual entailment. However, our results, obtained using the *Excitement Open Platform* (Padó et al., 2015), were considerably worse than that of distributional similarity models, hence we do not consider them further in this paper.

Baselines. We employ two baselines. First, an unsupervised baseline, which simply computes the similarity between the user claim and main claim vectors without using the implicit premises. Each user claim is matched to the most similar main claim. The other is a supervised baseline, which uses a support vector machine (SVM) classifier with an RBF kernel, trained on the user comments, to predict the label corresponding to the main claim. We train and evaluate the model using a nested 5×3 cross-validation, separately for each topic. The hyperparameters C and γ are optimized using grid search. We use the well-known LibSVM implementation (Chang and Lin, 2011).

Premise sets and combination with claims. To obtain a single combined representation of a premise set, we simply concatenate the premises together before computing the distributional vector representation. We do the same when combining the premises with either of the claims. This is exemplified in Table 7. In what follows, we denote the user claim, the main claim, and the gap-filling premise set with $\mathrm{U_i}$, $\mathrm{M_j}$, and $\mathrm{P_{ij}}$, respectively.

5.2 Matching with Implicit Premises

To answer the first research question – whether using premise sets can help in matching claims – we use gold-annotated premise sets and combine these with either the main claim or the user claim. The main idea is that, by combining the premises with a claim, we encode the information conveyed by the premises into the claim, hopefully making the two claims more similar at the textual level.

We consider four models: the unsupervised baseline, denoted "$\mathrm{U_i} \leftrightarrow \mathrm{M_j}$", the supervised baseline,

[2] We use the pre-trained vectors available at `https://code.google.com/p/word2vec/`

Type	Text content
U_i	*Marijuana has so many benefits for sick people.*
M_j	*Marijuana is used as a medicine for its positive effects.*
P_{ij}	*Marijuana helps sick people. Sick people use marijuana.*
U_i+P_{ij}	*Marijuana has so many benefits for sick people. Marijuana helps sick people. Sick people use marijuana.*
M_j+P_{ij}	*Marijuana is used as a medicine for its positive effects. Marijuana helps sick people. Sick people use marijuana.*

Table 7: Combination of premise sets and claims.

	Topic				
Model	MA	GR	AB	OB	Avg.
$U_i \leftrightarrow M_j$	7.39	12.52	24.59	10.87	13.84
$U_i \leftrightarrow M_j$ (S)	35.26	27.81	33.30	20.92	29.32
U_i+$P_{ij} \leftrightarrow M_j$	22.73	**46.03**	47.22	21.41	34.35
$U_i \leftrightarrow M_j$+P_{ij}	**48.05**	28.23	**49.34**	**64.11**	**47.43**

Table 8: Performance of claim matching baselines and oracle performance of the claim matching models utilizing implicit premises from annotator A1 (macro-averaged F1-score).

denoted "$U_i \leftrightarrow M_j$ (S)", the model in which the premises are combined with the user claim, denoted "U_i+$P_{ij} \leftrightarrow M_j$", and the model in which the premises are combined with the main claim, denoted "$U_i \leftrightarrow M_j$+P_{ij}". The latter two predict the main claim as the one that maximizes the similarity between two claims, after one of the claims is combined with the premises. The U_i+$P_{ij} \leftrightarrow M_j$ model considers all pairs of the user claim U_i and the gold-annotated premise sets P_{i*} for that user claim. In contrast, the $U_i \leftrightarrow M_j$+P_{ij} model considers all pairings of the main claim M_j and the gold-annotated premise sets P_{*j} for that main claim. In effect, this model tries to fill the gap using different premise sets linked to the given main claim. In this oracle setup, we always use the gold-annotated premise set for the main claim.

In Table 8, we show the claim matching results in terms of the macro-averaged F1-score on the development set. Results demonstrate that using the implicit premises helps in selecting the most similar main claim, as the models with added implicit premises outperform the unsupervised baseline by 20.5 and 33.6 points of F1-score. Furthermore, the model that combines the premises with the main claim considerably outperforms the two baselines and the model that combines the premises with the user claim. An exception is the GR topic, on which the the latter model works best. Our analysis revealed this to be due to the presence of very general (i.e., lexically non-discriminative) premises in some of the premise sets (e.g., *"Straight people have the right to marry"*), which makes the corresponding main claim more similar to user claims. Another interesting observation is the very good performance on the OB topic. This is because only one of the 16 main claims contains the word *Obama*, also making it more similar to user claims. However, after the premise sets get combined with all the main claims, this difference diminishes and the matching performance improves.

We obtained the above results using premises compiled by annotator A1. To see how model performance is influenced by the differences in premise sets, we re-run the same experiment with the best-performing $U_i \leftrightarrow M_j$+P_{ij} model, this time using the premises compiled by annotators A2 and A3. Although we obtained a lower macro-averaged F1-score (33.97 for A2 and 32.91 for A3), the model still outperforms both baselines. On the other hand, this suggest that the performance very much depends on the quality of the premises.

The claim matching problem bears resemblance with query matching in information retrieval. A common way to address the lexical gap between the queries and the documents is to perform query expansion (Voorhees, 1994). We hypothesize that human-compiled premises are more useful for claim matching than standard query expansion. To verify this, we replicate setups U_i+$P_{ij} \leftrightarrow M_j$ and $U_i \leftrightarrow M_j$+P_{ij}, but instead of premise sets, use (1) WordNet synsets and (2) top k distributionally most similar words (using word vectors from Section 5.1 and k=$\{1, 3, 5, 7, 9\}$) to expand the user or the main claim. We obtained no improvement over the baselines, suggesting that the lexical information in the premises is indeed specific.

5.3 Premise Generalization

From a practical perspective, we are interested to what extent the premises generalize, i.e., whether it is possible to reuse the premises compiled for the main claims, but different user claims. We choose the best-performing model from the previous section ($U_i \leftrightarrow M_j$+P_{ij}), and apply this model and the baseline models on the test set. This means that the model uses the premise sets P_{ij} for pairs of claims

Model	Topic				Avg.
	MA	GR	AB	OB	
$U_k \leftrightarrow M_j$	9.60	19.68	27.70	12.39	17.35
$U_k \leftrightarrow M_j$ (S)	29.01	**29.39**	21.09	18.22	24.43
$U_k \leftrightarrow M_j$+P_{ij}	**30.63**	23.00	**32.72**	**23.87**	**27.55**

Table 9: Performance of claim matching baselines and the models utilizing the implicit premises on the test set (macro-averaged F1-score).

U_i and M_j from the training set, and the hope is that the same premise sets will be useful for unseen user claims U_k. Results are shown in Table 9. The model again outperforms the baselines, except on the GR topic. The performance improvement varies across topics: the average improvement over the unsupervised and supervised baselines is 10.2 and 3.12 points of F1-score, respectively. This result suggests that the premises that fill the gap generalize to a certain extent, and thus can be reused for unseen user claims.

5.4 Premise Retrieval

In a realistic setting, we would not have at our disposal the implicit premises for each main claim, but try to generate or retrieve them automatically. We preliminary investigate the feasibility of this option with our third research question – could the implicit premises be retrieved automatically?

To retrieve the premise set P and then perform claim matching, we use a simple heuristic: given a user claim as input, we choose N premises most similar to the user claim, and then combine them with the user claim. We next compute the similarity between the premise-augmented claim vector and all the main claims. If the average similarity to main claims has increased, we increment N and repeat the procedure, otherwise we stop. The main idea is to retrieve as many premises as needed to bring the user claim "closer" to the main claims. We run this with N ranging from 1 to 5. In cases when combining the user claim with additional premises makes the claim less similar to the main claims, no combination takes place.

We consider two setups: one in which the pool of premises to retrieve from comes from the topic in question (within-topic), and the other in which the premises from all four topics are considered (cross-topic). Results are shown in Table 10. We evaluate on both the development set the test set, as well as within-topic (WT) and cross-topic (XT) premise retrieval. Results suggest that our simple method for within-topic premise retrieval improves claim matching over the baseline for all topics except the OB topic. On the other hand, results on the test set indicate that the model does not generalize well, as it does not outperform the baseline.

Model	Topic				Avg.
	MA	GR	AB	OB	
$U_i \leftrightarrow M_j$	7.39	12.52	24.59	**10.87**	13.84
U_i+P $\leftrightarrow M_j$ (WT)	**8.95**	**19.54**	**29.32**	7.30	**16.28**
U_i+P $\leftrightarrow M_j$ (XT)	8.56	19.01	28.73	7.07	15.84
$U_k \leftrightarrow M_j$	**9.60**	**19.68**	**27.70**	12.39	**17.35**
$U_k \leftrightarrow M_j$ (XT)	5.69	17.75	15.38	**12.43**	12.82

Table 10: Performance of the claim matching model with premise retrieval on the dev. set (upper part) and test set (lower part); macro-avg. F1-score.

6 Conclusion

We addressed the problem of matching user claims to main claims. Implicit premises introduce a gap between two claims. This gap is easily filled by humans, but difficult to bridge for natural language processing methods.

In the first study, we compiled a dataset of implicit premises between matched claims from online debates. We showed that there is a considerable variation in the way how human annotators fill the gaps with premises, and that they use premises of various types. We also showed that the similarity between claims, as judged by humans, negatively correlates with the size of the gap, expressed in the number of premises needed to fill it.

In the second study, we experimented with computational models for claim matching. We showed that using gap-filling premises effectively reduces the similarity gap between claims and improves claim matching performance. We also showed that premise sets generalize to a certain extent, i.e., we can improve claim matching on unseen user claims. Finally, we made a preliminary attempt to retrieve automatically the gap-filling premises.

This paper is a preliminary study of implicit premises and their relevance for argumentation mining. For future work, we want to further study the types of implicit premises, as well as relationships between them. We also intend to experiment with more sophisticated premise retrieval models.

Acknowledgments. We thank the reviewers for their many insightful comments and suggestions.

References

Eneko Agirre, Mona Diab, Daniel Cer, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 task 6: A pilot on semantic textual similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 385–393. Association for Computational Linguistics.

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

Filip Boltužić and Jan Šnajder. 2015. Identifying prominent arguments in online debates using semantic textual similarity. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 110–115. Association for Computational Linguistics.

Tom Bosc, Elena Cabrio, and Serena Villata. 2016. DART: A dataset of arguments and their relations on Twitter. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association.

Elena Cabrio and Serena Villata. 2012. Combining textual entailment and argumentation theory for supporting online debates interactions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 208–212. Association for Computational Linguistics.

Elena Cabrio, Serena Villata, and Fabien Gandon. 2013. A support framework for argumentative discussions management in the web. In *The Semantic Web: Semantics and Big Data*, pages 412–426. Springer.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pages 177–190. Springer.

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

Norbert E Fuchs, Kaarel Kaljurand, and Tobias Kuhn. 2008. Attempto controlled english for knowledge representation. In *Reasoning Web*, pages 104–124. Springer.

Debanjan Ghosh, Smaranda Muresan, Nina Wacholder, Mark Aakhus, and Matthew Mitsui. 2014. Analyzing argumentative discourse units in online interactions. In *Proceedings of the First Workshop on Argumentation Mining*, pages 39–48. Association for Computational Linguistics.

Theodosis Goudas, Christos Louizos, Georgios Petasis, and Vangelis Karkaletsis. 2014. Argument extraction from news, blogs, and social media. In *Hellenic Conference on Artificial Intelligence*, pages 287–299. Springer.

Ivan Habernal and Iryna Gurevych. 2016. Argumentation mining in user-generated web discourse. *arXiv preprint arXiv:1601.02403*.

Ivan Habernal, Judith Eckle-Kohler, and Iryna Gurevych. 2014. Argumentation mining on the web from information seeking perspective. In *Proceedings of the Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing*.

Kazi Saidul Hasan and Vincent Ng. 2014. Why are you taking this stance? Identifying and classifying reasons in ideological debates. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 751–762. Association for Computational Linguistics.

María Pilar Jiménez-Aleixandre and Sibel Erduran. 2007. Argumentation in science education: An overview. In *Argumentation in Science Education*, pages 3–27. Springer.

Marco Lippi and Paolo Torroni. 2016. Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology (TOIT)*, 16(2):10.

Clare Llewellyn, Claire Grover, Jon Oberlander, and Ewan Klein. 2014. Re-using an argument corpus to aid in the curation of social media collections. In *Proceedings of the Nineth International Conference on Language Resources and Evaluation (LREC 2014)*, pages 462–468. European Language Resources Association.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *Proceedings of ICLR*, Scottsdale, AZ, USA.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Amita Misra, Pranav Anand, JEF Tree, and MA Walker. 2015. Using summarization to discover argument facets in online idealogical dialog. In *Proceedings of*

the 2015 Annual Conference of the North American Chapter of the ACL, pages 430–440. Association for Computational Linguistics.

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

Marie-Francine Moens. 2014. Argumentation mining: Where are we now, where do we want to be and how do we get there? In *Post-proceedings of the forum for information retrieval evaluation (FIRE 2013)*.

Sebastian Padó, Tae-Gil Noh, Asher Stern, Rui Wang, and Roberto Zanoli. 2015. Design and realization of a modular architecture for textual entailment. *Natural Language Engineering*, 21(02):167–200.

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

Parinaz Sobhani, Diana Inkpen, and Stan Matwin. 2015. From argumentation mining to stance classification. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 67–77. Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1501–1510.

Reid Swanson, Brian Ecker, and Marilyn Walker. 2015. Argument mining: Extracting arguments from online dialogue. In *Proceedings of 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 2015)*, pages 217–227. Association for Computational Linguistics.

Peter D Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37(1):141–188.

Ellen M Voorhees. 1994. Query expansion using lexical-semantic relations. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 61–69. Springer-Verlag New York, Inc.

Douglas Walton, Christopher Reed, and Fabrizio Macagno. 2008. *Argumentation schemes*. Cambridge University Press.

Douglas Walton. 2005. *Argumentation methods for artificial intelligence in law*. Springer.

Douglas Walton. 2012. Using argumentation schemes for argument extraction: A bottom-up method. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 6(3):33–61.

Adam Wyner, Tom van Engers, and Anthony Hunter. 2010. Working on the argument pipeline: Through flow issues between natural language argument, instantiated arguments, and argumentation frameworks. In *Proceedings of the Workshop on Computational Models of Natural Argument*.