International Journal of Semantic Computing
Vol. 5, No. 4 (2011) 363–381

DOI: 10.1142/S1793351X11001328

World Scientific
www.worldscientific.com

IDENTIFYING JUSTIFICATIONS IN WRITTEN DIALOGS BY CLASSIFYING TEXT AS ARGUMENTATIVE

OR BIRAN

Department of Computer Science
Columbia University, 1214 Amsterdam Avenue
New York, NY 10027, USA
orb@cs.columbia.edu

OWEN RAMBOW

Center for Computational Learning Systems
Columbia University, 475 Riverside Drive
New York, NY 10115, USA
rambow@ccls.columbia.edu

In written dialog, discourse participants need to justify claims they make, to convince the reader the claim is true and/or relevant to the discourse. This paper presents a new task (with an associated corpus), namely detecting such justifications. We investigate the nature of such justifications, and observe that the justifications themselves often contain discourse structure. We therefore develop a method to detect the existence of certain types of discourse relations, which helps us classify whether a segment is a justification or not. Our task is novel, and our work is novel in that it uses a large set of connectives (which we call indicators), and in that it uses a large set of discourse relations, without choosing among them.

Keywords: Justification; argumentation; discourse; written dialog; online media.

1. Introduction

Natural language processing has recently seen a proliferation of interest in genres other than newswire. In particular, written dialog such as email and web-based discussion forums and blogs have been attracting a lot of attention, as they show a lot of interesting uses of language which newswire and other highly monologic and purely informative genres do not. For example, consider subjective statements and attempts to justify such statements. While newswire may contain editorials, these are the exception. In contrast, in spontaneously user-generated online content, we find many subjective statements, and, partly because of the interactive nature of the medium, many attempts at justifying such statements.

In this paper, we address the problem of identifying justifications for subjective claims in interactive written dialogs. This effort is part of a larger effort to detect attempts to persuade in written dialog, which in turn is motivated from a desire to

indentify influencers among discourse participants. However, we believe that the research presented in this paper is of interest beyond the motivating application, for example, for categorizing text or passages as expositive *versus* argumentative. We do not take "justification" to mean a particular discourse relation[a] (such as the JUSTIFY relation from Rhetorical Structure Theory [1]). Instead, we use this term in a broad dialogic sense: the writer makes an utterance which conveys subjective information (in the sense of [2]) and anticipates the question "Why are you telling me that?" Put differently, she is showing the reader that she is being relevant in a Gricean sense [3], presumably in an attempt to engage the reader and have him continue reading.

Here are some examples of what we consider to be justification, taken from our two corpora of written dialog (in each category, we provide one example from each corpus). The provided categories are only for explanatory purposes, they are not part of the task we address, nor of the solution we propose.

(1) Recommendation for action, and motivation for proposed action:
Claim: I'd post an update with the new date immediately.
Justification: In case anyone makes plans between now and when you post the reminder.

Justification: Our first heading is quite long, and against our MOS, it contains most of the title of the article.
Claim: I suggest we shorten it to "Topics".

(2) Statement of like or dislike or of desires and longing, and subjective reason for this like or dislike or desire or longing.
Claim: This is a great, great record.
Justification: I'm hesitant to say that kind of thing because I'm not a critic; but it is certainly in a league with Robyn's very best work. The Venus 3 come together as a band in a way I don't think they really did on O' Tarantula, and it just touches me very deeply.

Claim: I'm having a very hard time seeing why the first would be preferable.
Justification: It seems to engage in subtle puffery, what with the "over 500 pages ... in about 60 days" statement.

(3) Statement of like or dislike or of desires and longing, and claimed objective reason for this like or dislike or desire or longing:
Claim: Song of the South should be released again.
Justification: It is not racist. Uncle Remus was a slave and the stories came from slavery days. While slavery was a horrible thing, we can't just act like it never happened.

[a] Discourse relations are relations between text spans in discourse; they have the property that two text spans which are juxtaposed and which are related by a particular discourse relation are interpreted as coherent by the reader or listener; i.e., the reader or listener understands why these two spans were juxtaposed.

Claim: It is recentism and shouldn't be there.
Justification: As much as the conflict seems significant, it actually does not appear to be that important in an historical perspective, the way we should be looking at the article.

(4) Statement of subjectively perceived fact, with a proposed objective explanation:
Claim: I don't think Wilf will die.
Justification: Wilf's going to have to kill Ten to save Donna or something, 'cause of the whole 'you've never killed a man' thing that TV woman said.

Claim: "Mythical" is a compromise, but I think it's a fairly accurate one.
Justification: there are without question many myths involving Santa Claus. Everything printed, filmed, spoken, or otherwise told about him cannot be true.

(5) A claimed general objective statement and a more specific objective statement that justifies the more general one.
Claim: But it always leads to lives and potential unfulfilled when love is thwarted . . .and depression, too.
Justification: We see a hell of a lot of that still, judging by the number of gay and bi men who are on anti-depressants and in therapy.

Claim: The brain does not function when there is no blood pressure, but nor do cells "go bye bye" in a few minutes either.
Justification: The brain just sits quietly accumulating damage that requires increasingly sophisticated technology to reverse with a good prognosis.

What is striking is that all but one of these justifications, the first in (1), are not atomic discourse units, but contain argumentation themselves: in order to justify a claim, the writer is presenting an entire argument. For example, in (5), the first justification contains two parts: an empirical claim, and then evidence for that claim. The reader, however, interprets the entire passage as the justification of the original claim. Thus, we are interested in detecting argumentation in support of a given claim, such that the entire argumentation is considered the justification for the claim by the reader. As a consequence, in this paper, we are not interested in detecting a single discourse relation between the claim and the justification, for example one of those proposed by Rhetorical Structure Theory (RST) [1] or the theory underlying the Penn Discourse Treebank [4]. Instead, we are interested in justification as a type of discourse contribution, which is frequently characterized as containing argumentation. However, argumentation is not characterized by a single discourse relation; instead, it can be realized by a large number of discourse relations. As a consequence, recent work on identifying discourse relations [5− 7] is only relevant as a building block, but it is not the solution to our problem. Instead, we use a multi-step approach:

- We extract lists of indicators for a number of relations from the RST Treebank, a news corpus annotated with RST relations. For example, *because* is an indicator for the CAUSE relation.

- We extract a list of co-occurring content word pairs for each of the indicators from a large, multi-topic corpus (English Wikipedia). For example, (*chosen*, *proximity*) is a word pair for *because*.
- We use the lists of pairs to formulate features for a machine learning model and apply it to the task of identifying justifications in two corpora of online discussion.

Crucially, we do not apply these new features to both the claim and the candidate justification: we only look at the candidate justification, with the assumption that the justification frequently includes complex discourse relations. In experiments looking at both claim and candidate (which are not described in this paper) we observed that including the claim consistently adds nothing or very little (less than 0.5%) to all of our systems.

While our work is in the context of a larger project which aims at identifying persuasion, i.e., both the claim and its justification, in this paper we only report on the automatic detection of justifications, given gold-standard claims. The identification of potential claims is related to the identification of subjectivity, and thus falls in a completely different line of research. This work is an extension of [8], where we first described this task, our method and some preliminary results.

This paper is structured as follows. In Sec. 2, we provide an overview of related work. We then introduce our data and our annotation in Sec. 3. We start out by presenting fully supervised learning experiments on this corpus; these experiments provide a baseline for this paper (Sec. 4). We then investigate the use of additional features obtained through unsupervised methods (Sec. 5). We finish with a discussion (Sec. 6) and a conclusion (Sec. 7).

2. Related Work

As we explained in the introduction, our work is novel and different from other work in that we are not interested in finding discourse relations from a specific pre-defined set. In this section, we briefly review previous research that attempts to identify specific discourse relations, as we draw on the techniques developed by those researchers.

The principal idea here is the line of research started by Marcu and Echihabi [9], who use unsupervised methods to increase the recognition of implicit relations, i.e., relations not signaled by a cue word. The basic technique is simple: sentences with explicit connectives are used to train models to recognize cases of implicit relations; the underlying assumption is that implicit relations and explicitly signaled relations do not differ greatly in terms of the content of the related segments. These techniques were further developed by Blair−Goldensohn *et al.* [5, 7]. Critical assessments of this approach can be found in Sporleder and Lascarides [10] and in Pitler *et al.* [6]. The latter do an extensive study using the Penn Discourse Treebank [4], and observe that many of the meaningful word pairs learned from unannotated data involve closed-class words (which contradicts the intuition that these word pairs

represent semantic relations), and that the models derived from relations with explicit connectives do not, in fact, work very well on relations with implicit connectives. We will address this issue again in Sec. 5.2.

3. Data

We use four corpora in the different stages of our system.

The RST Treebank [11] is a subset (176,383 words) of the *Wall Street Journal* part of the Penn Treebank, annotated with discourse relations based on Rhetorical Structure Theory (RST). [1] We use the RST Treebank to extract relation indicators. We do not use the Penn Discourse Treebank directly, though we could have used it instead of the RST Treebank.

For our unsupervised word pair extraction, we use English Wikipedia.[b] We preprocessed the corpus to remove HTML tags, comments, links and text included in the tables and surrounding figures (e.g., captions and descriptions). The remaining text was lowercased and split into sentences.

Finally, we run our system on two corpora of written dialog, coming from different online media, and compare our results on the two.

The first is a corpus of 309 blog threads from LiveJournal (LJ),[c] a personal diary-style blogging service. Each thread contains an original post by the blog owner and a set of comments in a tree structure, — that is, a comment is associated with a particular previous entry, which can itself be a comment or the original post — by other LJ users as well as the owner. One important attribute of the LiveJournal corpus is that there is wide variation among the threads: the standard deviations of the entry length, both in words and in sentences, are higher than the mean; that is also the case for claims per thread and claims per entry.

The second written dialog corpus comes from Wikipedia discussion forums, or Wikipedia Talk (WT) pages. For every article on English Wikipedia, there is a corresponding Talk page where editors discuss issues related to the editing of the article. We have a total of 118 threads taken from the Talk pages of 28 articles. Each thread contains multiple posts, and posts may or may not be replies to other posts. These threads can be seen as having a tree structure as well, with the topic of discussion as the root and posts descending either from the root or from other posts. For consistency, we refer to the original post in LJ, the LJ comments and the WT posts simply as *entries*.

The threads from both sources contain annotated *claims* and their corresponding *justifications*. A justification can only be made by the same poster who made the original claim, but it may be located in a different entry. All annotated claims have justifications, and a claim may have more than one justification. In LJ, 32.4% of the justifications are in the sentence following the claim; 97.3% are in the same entry as the claim; 77.6% appear after the claim. In WT, the corresponding numbers are 22.9%, 87.7%

[b] A snapshot of all article texts as of April 8th, 2010.
[c] http://www.livejournal.com

and 79.1%. Another interesting statistic is the number of justifications given per claim in the corpus: in WT, the average is 2.2, while in LJ it is only 1.5. The variance is also higher in WT, with the standard deviation being 2.1 compared with 1.2 in LJ.

In inter-annotator agreement calculations on a subset of the LJ data (including only those claims which were marked by both annotators, with candidate justification sentences chosen in the same way they are chosen during a run of the system — see below) we observed a kappa measure of 0.69 and an f-measure between annotators of 0.75, showing that there is substantial agreement on justification once the claim is agreed upon.

4. Sentence Pair Classification: Fully Supervised Learning

The task is deciding for a pair of sentences, the first of which is marked the *claim*, whether or not the second sentence is a justification of the claim.

Our sentence pairs come from LiveJournal blog threads or Wikipedia discussion threads. We describe the exact way in which we create our data set of pairs in a later section. Here we describe the classification system we used as a baseline.

4.1. *Naive baseline*

To put things in context, we provide the results of a very naive baseline which simply chooses the sentence immediately following each claim as its justification.

4.2. *Heuristic baseline*

We achieve a better baseline performance for our task using a heuristic system with the following rules:

(1) If the claim is not in the same entry as the candidate justification, classify as NO.
(2) If the distance, in number of sentences, between the claim and the candidate justification is higher than a manually tuned threshold (4 for LJ, 3 for WT), classify as NO.
(3) Otherwise classify as YES.

This very basic system achieves high recall on LJ (more than 91% in cross-validation — see Table 2), with a significant increase in precision over an all-positive classification (29% compared with 10%). On WT, although recall suffers more (68%), it is offset by the greater gain in precision (32% compared with 9%).

4.3. *Hybrid baseline*

Because the heuristic baseline system achieves some precision with very little sacrifice of recall in LJ, we use it as a first stage in all systems in our experiments.

For simplicity of comparison and because it produces similar F-measure performance we use it similarly in all systems in the WT experiments. In preliminary experiments we found that adding the heuristic always improved performance, for both LJ and WT.

These hybrid systems first pass the data through the first two rules above; if a data point passes, then it is sent to a supervised learning classifier. Our hybrid baseline classifier operates on only two simple features: *beforeClaim*, a binary feature signifying whether the justification candidate comes before or after the claim, and *sentenceLength*, the length (in words) of the justification candidate. We found that justifications are longer on average than other sentences. We tried to match the claim to the justification in ways other than distance, by adding word overlap features with various variations including the use of stemming, *n*-grams, and WordNet for synonymy resolution. However, none of these attempts increased performance for this task.

While claims are allowed to have multiple justifications, it is rare that they have more than a few. To avoid picking too many sentences as justifications for a single claim we added a post-processing stage that looks at the pairs which share a claim and prevents all but the two (for LJ) or four (for WT) with the highest confidence from being classified as justifications. Two and four were found the be the optimal number for their respective data sets in a manual tuning.

This is the real baseline used in our experiment; the lesser two baseline results are shown for completeness. In all systems of Tables 2–4 which are described as "HB + X", HB stands for the hybrid baseline containing these two features with post-processing.

4.4. *Bag of words baseline*

Finally, we define a further baseline. We start with the hybrid baseline and add the standard Bag of Words features: we used all non-punctuation tokens which appear more than five times in the data set, each as a separate feature, for a total of 1474 in LiveJournal and 2422 in Wikipedia discussions. This baseline is referred to as "HB + bag-of-words".

5. Adding Features Obtained Through Unsupervised Mining

Particular RST relations, such as CAUSE, CONCESSION or CONTRAST may indicate argumentation.

Discovering RST relations in text is not a simple task. Some relations typically contain a connector word or phrase — such as *but* for the CONTRAST relation — but sometimes it may be implicit or replaced with a paraphrase (for example, *but* may be replaced with *on the other hand*). In online dialog especially, we expect more frequent irregularities in the usage of standard connectors. In addition, many such connectors are not reliable indicators even when present, since they tend to be common, ambiguous words. Still other relations rarely or never make use of connectors.

The idea driving our method is that some word combinations are more likely to appear as part of a relation. A simple example for contrast are antonyms — for instance, *easy* and *difficult* in the following sentence from LiveJournal:

> Its easy to flatter people, but its difficult to tell the truth and say something honest that might sound mean.

More generally, words may have a likely causal or even more subtle relationship between them. Consider the causality between *fresh* and *best* in:

> Rum tastes best when it's still relatively fresh and you can still taste the cane.

The concession indicated by *horrible* and *happened* in:

> While slavery was a horrible thing, we can't just act like it never happened.

Or the elaboration evidenced by *photography* and *sensor* (as well as other possible pairs) in:

> Canon provide an overall better photography system, from body to sensor to optics (canon Lseries lenses are something out of this world).

Crucially, the word pairs above are *content words*, which are independent of the linguistic style and even grammaticality of the text in question. We should expect such pairs to be relevant to a variety of corpora, with the reservation that domain may have much to do with the frequency of their appearance. We chose Wikipedia as the corpus from which to extract pairs in order to minimize the dominance of domain-specific pairs, since Wikipedia articles deal with a variety of topics.

5.1. *Extracting indicators*

We extracted a list of indicators from the RST Treebank. Unlike the PDTB, which has a list of indicators that are used (explicitly or implicitly) for each relation, the RST Treebank simply specifies that two or more spans of text have a particular relation between them. We aim to automatically create a list of the most relevant n-grams for each relation, and choose our indicators from among the top candidates. Using this method we are able to find indicators that may not appear in manual lists.

Specifically, our method works as follows.

We first choose n relations which we view as relevant for our task. We chose relations which relate to increasing the reader's willingness to accept a claim. RST [1] distinguishes presentational relations from subject-matter relations; presentational relations are defined in terms of changes in the reader's strength of belief, desire, or intention, while the latter are defined in terms of making the reader entertain a new proposition, such as causality. Basically, we are interested in presentational relations. However, as [12] point out, subject matter relations can co-exist with presentational relations: claiming a causal relation between two events may well be the best way of convincing the reader that the caused (or the causing)

event did indeed happen. Thus, we choose among both presentational and subject-matter relations those which are most likely to be usable in an attempt to make the reader accept a previously made claim.

For our experiments, we originally chose 14 relations. The RST Treebank uses a superset of a subset of the original relations proposed by RST. Specifically, MOTIVATION and JUSTIFY are not in the RST Treebank — we would have used them if they were. We excluded mainly subject-matter relations, specifically relations which are purely semantic such as MANNER, MEANS, or TEMPORAL-SAME-TIME; topic- and structure-related relations such as LIST, SUMMARY or TOPIC-SHIFT; and BACKGROUND, the only presentational relation we excluded, since its effect is to increase the reader's *ability* to understand the presented material, not necessarily his or her inclination to do so. During experiments, we discarded two of these, ATTRIBUTION and RHETORICAL-QUESTION, since they had no effect on the results, and were left with the 12 relations shown in Table 1.

After choosing our relations, we create a set of documents $D = \{d_1, \ldots, d_n\}$, where each document d_i contains all the text from the RST Treebank participating in relation i. The two spans of text participating in a relation (identified as such by the corpus) are retained as a single line.

We compute the top n-grams with a variant of tf-idf. We do the following for unigrams, bigrams, trigrams and 4-grams:

(1) Extract all n-grams from all documents.
(2) Compute idf for each n-gram in the usual way.
(3) Compute for each n-gram j in each document d_i the tf variant $\text{tf}*_{ij} = \frac{l_{ij}}{\sum_k l_{ik}}$ where l_{ik} is the number of lines in d_i in which the n-gram k appears at least once. The intuition for this altered measure is that since each line corresponds to one instance of the relation, an n-gram appearing multiple times in the same line would be overweighted with the standard measure.
(4) Create a list of n-grams for each document sorted by tf*-idf.

Table 1. The relations for which indicators were extracted, with the number of indicators and a few samples.

Relation	Nb	Sample indicators
analogy	15	as a, just as, comes from the same
antithesis	18	although, even while, on the other hand
cause	14	because, as a result, which in turn
concession	19	despite, regardless of, even if
consequence	15	because, largely because of, as a result of
contrast	8	but the, on the other hand, but it is the
evidence	7	attests, this year, according to
example	9	including, for instance, among the
explanation-argumentative	7	because, in addition, to comment on the
purpose	30	trying to, in order to, so as to see
reason	13	because, because it is, to find a way
result	23	resulting, because of, as a result of

We delete all *n*-grams below a certain tf*-idf score. We used 0.004 as the cutoff value in all experiments. Some filtering was needed as it was not feasible to go over the entire lists (in the next stage below), and this was casually observed as a reasonable cutoff.

Finally, we manually went over the lists and deleted *n*-grams that seemed irrelevant, ambiguous or domain-specific. Many *n*-grams that appear even at the very top for some relations are clearly not relevant, mostly because of the relatively narrow domain of the RST Treebank. For example, the highest-ranking trigram for the EVIDENCE relation is *as a result*, which is reasonable; the next down the list, however, is *in New York* — clearly a product of the particular corpus. This manual culling only took place once, and the resulting list is publicly available.[d] It can be used to extract pairs from any corpus in an unsupervised way, as explained in Sec. 5.2.

At the end of this process, we are left with 69 indicators in total, some of which are shared between multiple relations. Table 1 shows the number of indicators and a few samples for each relation.

5.2. *Extracting word pairs*

Having finalized the list of indicators, we use it to extract word pairs from English Wikipedia. We split the corpus into sentences, remove sentences longer than 50 words,[e] and for each indicator in our list, extract a list of word pairs occurring in sentences in which the indicator occurs. We extract two lists of word pairs:

- The *sides* list, where the first word must occur to the left of the indicator and the second must occur to the right of the connector. This set contains 447,149,688 pairs.
- The *anywhere* list, where the words may occur anywhere in the sentence but the first word must occur earlier. This set contains 1,017,190,824 pairs.

The words participating in the indicator itself are not considered for either of the lists. Stoplisted words (using the list of [13]) are also not considered. When stop words are allowed to participate in the pairs performance decreases: we include the results for a system which uses a set of pairs extracted without using a stoplist in Table 2, for comparison with our better-performing systems. Interestingly, [6] report the opposite — that removing stop words hurts their performance. This difference in the contribution of using a stoplist can be explained by how we use the features; an explanation for this is given in the next section.

We collect frequency information — that is, how many times each word pair appears in the corpus. Pairs which appear less than 20 times are removed from the lists to reduce noise, but the frequency of the remaining pairs is not used in subsequent steps. After this filtering, the size of the *sides* list is 334,925 pairs and the size of the *anywhere* list is 719,439 pairs.

[d] At http://www.cs.columbia.edu/~orb

[e] Sentences longer than 50 words constitute only 2.7% of all Wikipedia sentences. Longer sentences are likely to be syntactically complex and thus too noisy for this method.

Although this method misses cases where the indicator is implicit and, for the *sides* list, cases where the indicator is sentence-initial, the abundance of data still allows the collection of large sets of word pairs.

5.2.1. *Similar indicators*

It might at first seem surprising that we use pairs from two related indicators, such as *because* and *because it is*. Why would it be useful to use both indicators? Clearly, the instances in the corpus of *because it is* are a subset of the instances of *because*. However, the word pair frequencies corresponding to each are very different. For example, two of the top pairs we found for *because* are (*education*, *act*) and (*road*, *traffic*), but they do not appear at all in the list for *because it is*; some top pairs there are (*population*, *threatened*) and (*wide*, *unlikely*) which occur for *because* but are not in the top 10,000. (Note that in this work we do not make direct use of frequencies except for cutting off pairs that appear less than 20 times around the indicator.) Upon inspection, we see that *because* is an indicator in the following discourse relations: CAUSE, CONSEQUENCE, EVIDENCE, EXPLANATION-ARGUMENTATIVE, REASON, RESULT. In contrast, *because it is* is an indicator only for REASON. These relations are defined differently by RST (or the RST Treebank). If the intensional definition differs, we expect the realization in language to differ as well. For example, CAUSE, RESULT or CONSEQUENCE will typically relate the depiction of two events, and *because it is* probably does not introduce the description of an event, but of a state (e.g., *I will not drive to the airport because it is unlikely that John will arrive*). Thus, indicator *because it is* is associated with a much narrower discourse meaning than indicator *because*; this is evidenced by the rather different sets of word pairs extracted using the two indicators. Therefore, the indicator *because it is* it can be useful in addition to the broad indicator *because*.

5.3. *Using the information*

Our task is defined as follows: given a pair of sentences, the first of which is defined to be the *claim*, determine whether the second sentence is a justification of the first sentence. In our experiments, we used a supervised classifier to decide this question. We describe the classifier and preparation in more detail in the next section; this section describes the various ways in which we formulated machine learning features from the data (indicators and word pairs) extracted in the previous sections.

We tried several approaches for using the extracted indicators and pairs.

5.3.1. *Indicators as lexical features*

In this simple approach, we used the indicators themselves as binary lexical features. These features did not improve on the baseline, perhaps because phrases which are good indicators are too rare in the data while common phrases are not very good indicators.

5.3.2. *Word pair features*

Here we used the extracted pairs from the two sets, *sides* and *anywhere*, to build features. In order to avoid a sparse feature space we took advantage of the natural structure of the pair lists — namely, the fact that each pair is associated with one or more indicators. Utilizing this fact, we use only 69 word pair conjunction features in our experiments. We created three different sets, each containing 69 features, which we use separately (that is, we experiment with systems that use only one of the sets), and which represent three levels of strictness in identifying a pair. Each feature ϕ_j is associated with a set of word pairs P_j corresponding to indicator j, and the variations are

$$\phi_j[\text{unigrams}] = \begin{cases} 1 & \text{if the candidate sentence contains either} \\ & \quad \text{of the words of any pair } p \in P_j \\ 0 & \text{otherwise} \end{cases}$$

$$\phi_j[\text{unordered}] = \begin{cases} 1 & \text{if the candidate sentence contains both} \\ & \quad \text{words of any pair } p \in P_j \\ 0 & \text{otherwise} \end{cases}$$

$$\phi_j[\text{ordered}] = \begin{cases} 1 & \text{if the candidate sentence contains both words} \\ & \quad \text{of any pair } p \in P_j, \text{ in their original order} \\ 0 & \text{otherwise.} \end{cases}$$

Going back to our initially surprising find that including stopwords in the pairs hurts performance, we attribute the difference to the fact that our features are not traditional sparse lexical features, but a relatively small number of lexical set features, and adding frequently occuring pairs to these sets renders the features virtually identical. The lexical set features rely on the fact that the member pairs are infrequent. Given the assumption that they share similar meaning in terms of the classification task, the union of the pairs becomes a meaningful and reasonably common indicator.

5.4. *Making it concrete — an example*

To make our description more concrete, consider the annotated claim and justification sample (4) from Sec. 1, repeated here for convenience as (6):

(6) **Claim:** I don't think Wilf will die.
Justification: Wilf's going to have to kill Ten to save Donna or something, 'cause of the whole 'you've never killed a man' thing that TV woman said.

Our system correctly identifies the justification for this claim. Specifically, the process following the stages explained earlier is as follows.

The RST Treebank contains spans of text which are annotated with the relations PURPOSE, CAUSE and REASON. Within these, our method described in Sec. 5.1 found the *n*-grams *because* for both CAUSE and REASON, and *in order to* for PURPOSE with high tf*-idf scores. Both of these *n*-grams passed our manual culling and ended in

our list of indicators. Note that neither indicator occurs verbatim in our example; instead, there is the purpose *to*, which has many meanings in English, and the contraction *'cause*. However, we do not rely on the occurrence of indicators for the classification.

In Wikipedia, we found many cases of the words *kill* and *save* appearing in a sentence with *in order to* between them, so the pair (*kill*, *save*) made it into our pairs list for the indicator *in order to*. Similarly, the word pair (*kill*, *killed*) made it into the pairs list for the indicator *because*, since the indicator appeared in Wikipedia between the two words. These pairs are part of the *sides* list since the words in this case were on both sides of the indicator, and trivially also the *anywhere* list which is a superset. The list features called *in order to* and *because* fire whenever the candidate sentence contains both words in the pairs (*kill*, *save*) and (*kill*, *killed*), respectively (in order or not, dependeing on the experiment), as well as other pairs found in a similar way. The classifier learned that these two features are good enough indicators to classify the sentence as a justification. The length of the sentence in this case is not exceptional, and no pairs for other indicators were found, but these two features were enough to make this judgement.

Note that we identify two relations in the sentence: the PURPOSE relation indicated by *to*, and the REASON or CAUSE relation indicated by *'cause*. The complexity of the sentence in terms of the number of discourse relations suggests that it contains argumentation, and our system will correspondingly have higher confidence in this sample because of the multiple positive features.

5.5. *Experiments*

We performed experiments on two corpora:

(1) An annotated corpus of 309 LiveJournal blog threads. Out of these, we reserved 40 threads for the test set and used 269 for training.
(2) An annotated corpus of 118 Wikipedia discussion threads. Out of these, we reserved 15 threads for the test set and used 103 for training.

We provide results on a 10-fold cross validation of the training set (which is what we used for development) as well as on the unseen test set for each corpus.

For each corpus, to build our data set we take each claim and produce from it a number of data instances, each including the claim and a candidate justification sentence. Candidate justifications are all sentences which belong to an entry that is either equal to or subsequent to the entry containing the claim, and which was authored by the same poster who made the claim.[f] Positive points are those containing the actual annotated justifications, while the rest are negative. Using this method, we arrive at 6636 training instances and 756 test instances in LJ, and at 19117 training instances and 1696 test instances in WT. We made sure the

[f]Although annotators were allowed to place justifications in an earlier entry than that containing the claim, in practice no such cases exist in the corpus.

Table 2. Precision, recall and F-measure obtained by the system in various experiments on the LiveJournal cross validation and test set. The best score at each column as well as the results for the best system (judged by F-measure) are highlighted.

System	CV P	CV R	CV F	Test P	Test R	Test F
next sentence	**46.35**	32.44	38.17	41.67	40	40.82
heuristic baseline	28.97	**91.04**	43.95	27.16	**88.35**	41.55
hybrid baseline (HB)	41.52	54.68	47.2	31.72	45.63	40.69
HB + bag-of-words	41.37	48.57	44.68	37.5	43.69	40.36
HB + indicators	41.52	54.68	47.2	31.72	45.63	40.69
HB + unigrams	42.12	56.5	48.26	35.38	46	40
HB + *anywhere*, no ordering	35.61	20.9	26.34	34.92	17.46	23.28
HB + *anywhere* with ordering	38.17	19.81	26.08	41.67	19.84	26.88
HB + *sides*, no ordering	42.93	61.6	50.6	**42.64**	53.4	**47.41**
HB + *sides* with ordering	42.97	61.24	50.5	41.86	52.43	46.55
HB + indicators + *sides*, no ordering	**43.12**	**61.81**	**50.8**	**41.86**	**52.43**	**46.55**
HB + indicators + *sides-no-stoplist*, no ordering	42.07	58.18	48.83	37.12	47.57	41.7

share of positive points in the training set is approximately equal to that of the test set in both corpora. In the LJ data sets, approximately 10% of the points are positive. In the WT data sets it is 9%.

For each corpus, we trained a Naive Bayes classifier on combinations of the features described in the previous section, using a 10-fold stratified cross validation as the development set. Table 2 shows the results of the LJ experiment, and Table 3 shows the results of the WT experiment. We found the results to be statistically significant using paired permutation tests on key system combinations — in particular, the best performing system for each corpus against all systems which use other pair lists or no pair features at all.

Table 3. Precision, recall and F-measure obtained by the system in various experiments on the Wikipedia Discussions cross validation and test set. The best score at each column as well as the results for the best system (judged by F-measure) are highlighted.

System	CV P	CV R	CV F	Test P	Test R	Test F
next sentence	33.03	22.94	27.08	34.19	23.87	28.12
heuristic baseline	31.75	**67.84**	43.26	31.67	**68.47**	43.3
hybrid baseline (HB)	36.22	35.4	35.8	37.68	35.14	36.36
HB + bag-of-words	43.76	42.64	43.19	**47.41**	49.55	48.46
HB + indicators	36.55	35.75	36.15	37.44	34.23	35.76
HB + unigrams	42.73	43.28	43.01	42.92	45.05	43.96
HB + *anywhere*, no ordering	47.25	21.38	29.44	40.86	17.12	24.13
HB + *anywhere* with ordering	**47.4**	20.1	28.23	38.46	15.77	22.36
HB + *sides*, no ordering	40.83	61.53	49.09	41.77	61.71	49.82
HB + *sides* with ordering	41.22	61.76	49.44	41.69	62.16	49.91
HB + indicators + *sides* with ordering	**41.26**	**61.82**	**49.49**	**41.69**	**62.16**	**49.91**
HB + indicators + *sides-no-stoplist* with ordering	44.55	18.71	26.36	44.09	18.47	26.03

Table 4. Precision, recall and F-measure obtained by the system in various experiments on the combined data set cross validation and test set. The best score at each column as well as the results for the best system (judged by F-measure) are highlighted.

System	CV P	CV R	CV F	Test P	Test R	Test F
next sentence	36.87	25.64	30.25	39.67	29.54	33.86
heuristic baseline	31.89	**73.36**	44.46	31.27	**73.23**	43.83
hybrid baseline (HB)	36.59	40.99	38.66	36.77	40.62	38.6
HB + bag-of-words	**43.77**	42.67	43.21	**45.43**	48.92	47.11
HB + indicators	33.41	42.47	37.4	33.41	42.46	37.4
HB + unigrams	42.79	42.35	42.57	42.18	44	43.07
HB + *anywhere*, no ordering	41.21	21.51	28.27	37.87	19.69	25.91
HB + *anywhere* with ordering	41.69	20.91	27.85	38.82	20.31	26.67
HB + *sides*, no ordering	**41.66**	**60.3**	**49.27**	**41.79**	**60.31**	**49.37**
HB + *sides* with ordering	39.1	61.26	47.74	39.09	60.62	47.53
HB + indicators + *sides*, no ordering	39.06	61.54	47.79	38.02	59.08	46.27
HB + *sides-no-stoplist*, no ordering	40.21	21.79	28.27	41.4	23.69	30.14

In the next experiment, we created a mixed data set: the training set was the combined LJ and WT training sets, and the test set the combined LJ and WT test sets. The best configuration for the baseline system in this experiment was found to be a limit of 3 for the distance of the candidate from the claim (as in the WT system) and a maximum number of justifications per claim of 3 (in between the LJ and WT systems). Again, we report the results on a 10-fold cross validation of the training set (making sure that the ratio of LJ threads to WT threads in each fold is similar) and on the test set. The results of this experiment are shown in Table 4.

In addition, we tested the performance of each of the best systems for the two data sets on the opposite data set. That is, we trained the best LJ system (HB + indicators + *sides*, no ordering) on the full LJ set, including all training and test threads, and tested its performance on the full WT set; conversely, we trained the best WT system (HB + indicators + *sides* with ordering) on the full WT set and tested it on the full LJ set. The results are shown in Table 5.

To put things in context, we also performed another experiment in which we evaluated on single sentences (as opposed to sentence-pairs). Here the task is simply to decide whether or not a sentence is a justification, for *any* claim. The sizes of the data sets in this experiment are 8508 for training and 1197 for the test set in LJ, and 9274 for training and 1112 for the test set in WT. Here we again used a Naive Bayes classifier, this time with only word pairs as features (the same features participating in the best system for the data set — *sides* with ordering for WT, *sides* with no ordering for LJ and the combined set). The heuristic rules could not be applied in this experiment, as they relate to a specific claim. The baseline is simply a greedy all-positive classification.

The results are shown in Table 6. They can be interpreted as the raw gain achieved by the pair features, since they only operate to identify justification

Table 5. Precision, recall and F-measure obtained by training the best system for each data set and testing on the opposite data set.

System	P	R	F
Best LJ system tested on WT	43.45	45.33	44.37
Best WT system tested on LJ	38.29	60.02	46.76

Table 6. Precision, recall and F-measure obtained for the single-sentence classification experiment. The baseline classifies all points as positive.

Data	System	CV P	CV R	CV F	Test P	Test R	Test F
LJ	baseline	11.66	100	20.89	14.75	100	25.71
	sides, no ordering	30.88	48.85	37.84	30.30	40	34.48
WT	baseline	19.14	100	32.13	21.22	100	35.01
	sides with ordering	23.27	91.72	37.13	25.03	91.95	39.35
LJ + WT	baseline	14.54	100	25.39	15.02	100	26.12
	sides, no ordering	18.81	87.5	30.97	19.18	88.63	31.54

sentences independent of claim in our main experiment. The heuristic rules together with the *beforeClaim* feature are the parts relating a sentence to the particular claim in the pair.

6. Discussion

Our results show that combinations of content words can be used to predict the presence of the justification-related RST relations, and that these are helpful in identifying justifications in online discussions. This finding suggests that justification segments in the context of a dialog make use of particular rhetorical tools. Specifically, our experiments also show that it is not merely the presence of certain words that is indicative of justifcation, but specifically the presence of a discourse relation, as evidenced by the inferior results for the unigram and "anywhere" pairs (as well as the bag-of-words baseline) in Tables 2–4.

The increase in performance in the pair decision task when compared to the single sentence task suggests that the presence of a claim and its location relative to the justification candidate are important and can significantly boost performance, even when only rudimentary methods are used.

Regarding indicators, it is interesting that we were able to find content word pairs with indicator phrases that would not in all cases be traditionally regarded as connectives; still, the location of the words with regards to the indicators is important — performance was dramatically increased when the content words came from opposing sides of the indicator.

In LJ, our best systems do not perform quite as well on the test set as they do on the cross-validation, but we can attribute that to the high degree of variation of the

corpus. Some variance between sets of threads is expected, and the difference in results is not so high as to warrant suspicions of overfitting. In particular, the relative contributions of the different components of our system are similar. It is interesting that in the WT and combined data set experiments, performance on the test data was very similar to the performance on the cross validation. This suggests that WT (which is a larger corpus, and accounts for approximately three quarters of the combined data set) is a fairly uniform corpus from which we can derive a robust model.

Another interesting observation is that while the hybrid baseline does very well compared to the heuristic baseline in LJ, it is not the case in WT. It may be attributable to the sentence length and claim entry features being better indicators in LJ, but note that while in both data sets the recall of the hybrid baseline is significantly lower, the increase in precision is much higher in LJ than in WT. One possible explanation is that the beam search, which in LJ limits the number of justifications per claim to 2, is important in increasing precision. In WT the limit is 4 justifications per claim, which does not have the same effect. However, it does provide the best performance for this baseline in WT; that is, using a lower number reduces recall significantly. In WT, the variation in number of justifications per claim is higher than in LJ, as participants more often offer many justifications for their claims.

While using the same system for both corpora reduces performance, it is encouraging that the decrease is relatively small. Training on one set exclusively and testing on the other is hard, as can be expected, but F-measure is only 4—5 points lower than the single-corpus systems. This shows that whether a sentence is a justification or not has more to do with the sentence itself than with the corpus it comes from. The decrease in performance is even lower in the case of the combined data set: the performance of the best system on this set is comparable to the performance of the best systems for each individual corpus. This suggests that while there are differences between the genres structurally and in the frequency with which word pairs are used, the semantics of the word pairs are consistent across both. If the words were used differently, we would expect more noise and lower performance in the combined data set.

Finally, comparing the three systems which were best-performing on each of the data sets, it seems that while the *sides* list is clearly superior to *anywhere*, it is less important whether we enforce the original order. What this suggests is that the original order actually is maintained most of the time. Using the indicators themselves as additional features contributes a little bit to performance when evaluating on only one corpus, but hurts in the combined data set. It may be that the indicators are used differently in the two corpora, and when trained on both there is simply more noise. In WT, which is a platform dedicated to debate, these connectives possibly correspond to discourse relations more often than in LJ, which includes more narratives and short contributions.

7. Conclusion

In this paper, we have addressed the issue of detecting justifications in written dialogs. This is a new task. We address the problem not by characterizing the link between the claim and its justification, but primarily by characterizing the justification itself as argumentative. We hypothesize that a text segment is argumentative if it contains at least one discourse relation which aims to increase the reader's willingness to accept a proposition. We build on previously developed techniques for mining characterizations of discourse relations from unannotated text [9], but we replace the notion of connective by a new notion of "indicator" of a relation, which is not tied to a specific morphological or syntactic class. We show that this technique improves the recognition of justifications as claims. In other work, which we are planning to publish in the near future, we have shown that our components helps in detecting persuasion, which in turn helps in identifying discourse participants as influencers. In that work, we also show that identifying argumentative discourse on its own (without relation to a claim) also improves the detection of influencers.

We will make both the argumentation classifier and the underlying data available to other researchers. Please contact the authors for information.

Acknowledgments

This work has been supported by the Intelligence Advanced Research Projects Activity (IARPA) via Army Research Laboratory (ARL) contract number W911NF-09-C-0141. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

References

[1] W. C. Mann and S. A. Thompson, Rhetorical structure theory: Toward a functional theory of text organization, *Text* **8**(3) (1988) 243–281.

[2] J. Wiebe, T. Wilson and C. Cardie, Annotating expressions of opinions and emotions in language ann, *Language Resources and Evaluation* **39**(2/3) (2005) 164–210.

[3] H. P. Grice, Logic and conversation, in *Syntax and Semantics*, Vol. 3, eds. P. Cole and J. Morgan (Academic Press, New York, 1975).

[4] R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi and B. Webber, The Penn Discourse Treebank 2.0, in *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, 2008.

[5] S. Blair-Goldensohn, K. McKeown and O. Rambow, Building and refining rhetorical-semantic relation models, in *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, Rochester, New York, Association for Computational Linguistics, 2007, pp. 428–435. [Online]. Available: http://www.aclweb.org/anthology/N/N07/N07-1054.

[6] E. Pitler, A. Louis and A. Nenkova, Automatic sense prediction for implicit discourse relations in text, in *Proceedings of the Joint Conference of the 47th Annual Meeting of*

the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, Suntec, Singapore, Association for Computational Linguistics, August 2009, pp. 683–691. [Online]. Available: http://www.aclweb.org/anthology/P/P09/P09–1077.

[7] Z. M. Zhou, M. Lan, Z. Y. Niu, Y. Xu and J. Su, The effects of discourse connectives prediction on implicit discourse relation recognition, in *Proceedings of the SIGDIAL 2010 Conference*, Tokyo, Japan, Association for Computational Linguistics, September 2010, pp. 139–146. [Online]. Available: http://www.aclweb.org/anthology/W/W10/W10–4326.

[8] O. Biran and O. Rambow, Identifying justifications in written dialog, in *Proceedings of the Fifth IEEE International Conference on Semantic Computing*, 2011.

[9] D. Marcu and A. Echihabi, An unsupervised approach to recognizing discourse relations, Philadelphia, PA, 2002.

[10] C. Sporleder and A. Lascarides, Using automatically labelled examples to classify rhetorical relations: An assessment, *Natural Language Engineering* **14**(3) (2008) 369–416.

[11] L. Carlson, D. Marcu and M. E. Okurowski, Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory, in *Current Directions in Discourse and Dialogue*, eds. J. van Kuppevelt and R. Smith (Kluwer Academic Publishers, 2003).

[12] M. Moser and J. D. Moore, Toward a synthesis of two accounts of discourse structure, **22**(3) (1996).

[13] C. Fox, A stop list for general text, SIGIR Forum Vol. 24, September 1989, 19–21 [Online]. Available: http://doi.acm.org/10.1145/378881.378888.