Exploiting Scope for Shallow Discourse Parsing

Rashmi Prasad*, Aravind Joshi*, Bonnie Webber†

*University of Pennsylvania
Philadelphia, PA, USA
rjprasad,joshi@seas.upenn.edu

†University of Edinburgh
Edinburgh, Scotland
bonnie@inf.ed.ac.uk

Abstract

We present an approach to automatically identifying the arguments of discourse connectives based on data from the Penn Discourse Treebank. Of the two arguments of connectives, called Arg1 and Arg2, we focus on Arg1, which has proven more challenging to identify. Our approach employs a sentence-based representation of arguments, and distinguishes *intra-sentential connectives*, which take both their arguments in the same sentence, from *inter-sentential connectives*, whose arguments are found in different sentences. The latter are further distinguished by paragraph position into *ParaInit* connectives, which appear in a paragraph-initial sentence, and *ParaNonInit* connectives, which appear elsewhere. The paper focusses on predicting Arg1 of Inter-sentential ParaNonInit connectives, presenting a set of scope-based filters that reduce the search space for Arg1 from all the previous sentences in the paragraph to a subset of them. For cases where these filters do not uniquely identify Arg1, coreference-based heuristics are employed. Our analysis shows an absolute 3% performance improvement over the high baseline of 83.3% for identifying Arg1 of Inter-sentential ParaNonInit connectives.

1. Introduction

Recent work on discourse parsing based on the discourse-level annotations of the Penn Discourse Treebank (Prasad et al., 2008) has addressed the problem of identifying the two arguments of explicit discourse connectives (Dinesh et al., 2005; Wellner and Pustejovsky, 2007; Elwell and Baldridge, 2008; Wellner, 2009). This "shallow" discourse parsing resembles chunking at the sentence level, since it does not concern itself with building the structure of the entire text, as is the case with many prior discourse parsing methods developed within different frameworks (Marcu, 1997; Forbes et al., 2003; Polanyi et al., 2004; Baldridge et al., 2007). Explicit discourse connectives in the Penn Discourse Treebank (PDTB)[1] are expressions from well-defined syntactic classes that denote discourse relations (e.g., *cause*, *contrast*, *elaboration*) between two abstract object arguments, such as events, states, and propositions.[2] Ex. (1) and Ex. (2) show two annotation tokens for the contrastive connective *but*. (The two arguments of the connective are called Arg1 and Arg2 in the PDTB. Arg2 is the argument syntactically associated with the connective and shown in bold in the examples here. Arg1 is simply the other argument, shown in italics in the examples. Connectives are underlined.)

(1) Despite all these innovations, most of the diamonds are still found in the sand swept away by the men wielding shovels and brushes – the ignominiously named "bedrock sweepers" who toil in the wake of the excavators. *Laboring in blue and gray overalls, they are supposed to concentrate on cleaning out crevices, and not strain their eyes looking for diamonds.* <u>But</u> **should they spy one, the company will pay a bonus equal to one-third its value.** [wsj_1121]

(2) *I'm not suggesting that the producers start putting together episodes about topics like the Catholic-Jewish dispute over the Carmelite convent at Auschwitz.* That issue, like racial tensions in New York City, will have to cool down, not heat up, before it can simmer. <u>But</u> **I am suggesting that they stop requiring Mr. Mason to interrupt his classic shtik with some line about "caring for other people" that would sound shmaltzy on the lips of Miss America.** [wsj_2369]

For the discourse parsing task, identification of Arg2 is relatively trivial in that it is syntactically associated with the connective. However, as Ex. (1) and Ex. (2) show, Arg1 may or may not be adjacent to the connective, thus making the task challenging. This difference in difficulty is attested by all prior attempts at identifying the arguments of connectives. In this paper, we take Arg2 of connectives to be easily identifiable, and focus on the task of identifying the Arg1 argument.

Our approach is novel in several respects. First, rather than identifying the exact argument spans (Dinesh et al., 2005) or the "heads" of arguments (Wellner and Pustejovsky, 2007; Elwell and Baldridge, 2008), we focus on *identifying the sentences containing the arguments*. Representing arguments in this shallow way not only has empirical support but is also of practical use for many applications including extractive summarization. Second, although some prior work has argued for the value of building separate models for different syntactic classes of connectives (Elwell and Baldridge, 2008), we propose instead to *classify connectives in terms of whether the connective and its Arg1 are collocated in the same sentence or not,*

[1] http://www.seas.upenn.edu/~pdtb. The PDTB corpus is available from the Linguistic Data Consortium, catalog entry LDC2008T05.

[2] The PDTB also annotates implicit discourse relations as well as relations expressed with non-connective expressions (called Alternative Lexicalizations (AltLex)). These relation types are not within the scope of this paper.

and further, whether the connective and its Arg1 are collocated in the same paragraph or not. Our motivation for such a classification rests in the idea that *discourse relations are structured differently as one progresses from the sentence to the paragraph, which is a coherent grouping of sentences, and then from the paragraph to the entire text, which is a coherent grouping of paragraphs*. Third, using our approach, we focus on identifying one category of connectives in our classification and present an algorithm for identifying their Arg1 argument. The algorithm involves a *filter-rank-evaluate* method that combines the application of *scope-based heuristics* for filtering potential candidate arguments, followed by *coreference-based heuristics* for ranking and evaluating the remaining candidates. Our application of the algorithm shows that our approach and method is promising, showing a 3% absolute improvement over the high 83.3% baseline of selecting the immediately previous sentence as Arg1.

2. The Penn Discourse Treebank: Overview

The PDTB (Prasad et al., 2008; PDTB-Group, 2008) is to date the largest available annotated corpus of discourse relations, with two major distingishing features. First, discourse relations are low-level and annotated independently of each other, so that no commitments are made towards any particular theory of high-level discourse structure. Given that there is little agreement among discourse researchers as to the nature of high-level discourse representations, this theory-neutrality makes the corpus appealing to a broad audience. In addition, the low-level annotations also lend to greater reliability in the annotations. Second, discourse relations, when explicit, are lexically grounded, thus making several discourse processing tasks more computationally tractable.

Annotated in the PDTB are discourse relations realized explicitly by discourse connectives and alternatively lexicalized expressions, or implicitly between adjacent sentences. Each discourse relation is annotated with a sense label drawn from a hierarchical sense classification. Discourse relations and their arguments are also annotated for their attribution, to record how they are ascribed - to the writer of the text or some other individual.

In this paper, we will focus on explicit connectives and their arguments, illustrated in Ex. (1) and Ex. (2). Explicit connectives are drawn from three syntactic classes: subordinating conjunctions, coordinating conjunctions, and discourse adverbials. Arguments of connectives can appear anywhere in the text, and they can span single clauses, mutliple clauses, or multiple sentences, as well as nominalizations that refer to abstract objects. There are a total of 40600 tokens of discourse relations annotated in PDTB, 18459 (45%) of which are explicit connectives. The distribution of the location of Arg1 of explicit connectives, reported in Prasad et al. (2008), shows that the majority (61%) of explicit connectives have Arg1 in the same sentence as the connective, with 30% of Arg1s in the sentence immediately preceding the connective, and 9% of Arg1s in some non-adjacent sentence.

3. Span and Location of Arguments

Before developing algorithms and models for identifying the arguments of connectives in the PDTB, two important issues arise. The first is due to the fact that arguments don't necessarily span a single clause or a single sentence. They can also span multiple clauses, multiple sentences, and even noun phrases and verb phrases. Furthermore, arguments within sentences can be discontinuous so that they don't necessarily project a single constituent in the syntax. This wide variation in the syntactic possibilities of argument spans makes the task of their identification quite challenging. Prior work has either tackled the problem of identifying exact argument spans, or circumvented the problem in some way. Dinesh et al. (2005) attempt to identify the exact argument spans of subordinating conjunctions. Since both arguments of this class of connectives are invariably in the same sentence, they develop a tree-subtraction algorithm to identify both arguments. They tackle the problem further by examining the sources of errors and suggesting improvements to the tree-subtraction heuristics. Wellner and Pustejovsky (2007), Wellner (2009) and Elwell and Baldridge (2008), on the other hand, attempt to identify the arguments of all classes of connectives using MaxEnt rankers and CRFs, but circumvent the problem of an exact argument match by assuming a "head-based" representation of arguments. This allows them to handle the full variation found in the syntax of the arguments.

The second issue is due to the fact that different types of connectives might be subject to different types of constraints in discourse. Thus, it is useful to separate connectives into distinct classes based on the methods being followed or on the view one adopts of how connectives differ from each other. For example, the tree-subtraction algorithm in Dinesh et al. (2005) developed specifically for subordinating conjunctions would not be able to handle coordinating conjunctions or discourse adverbials because it operates directly on syntactic trees where the argument (syntactic) dependencies for the three classes of connectives are quite different. In Elwell and Baldridge (2008), where arguments for all connectives are identified, separate models are developed for connectives grouped into their three syntactic classes. Importantly, this kind of classification of connectives was shown to improve on the results of Wellner and Pustejovsky (2007), where no such classification was made and a single model was used for all connectives.

Our own approach, which also considers these issues, differs from prior work, as discussed below.

3.1. Sentence-based Representation of Arguments

Like Wellner and Pustejovsky (2007), Elwell and Baldridge (2008), and Wellner (2009), we also circumvent the problem of exact argument identification. However, we do this by representing arguments in terms of the *sentences containing them* rather than their heads. There are empirical and practical reasons for this. First, experiments on the syntactic distributions of arguments show that *cases where an argument of a connective is a subordinate or embedded clause instead of the main clause of a sentence are in fact very rare* (Lee et al., 2008). Thus, identifying sentences would not only be equivalent to a head-based approach if

the sentence was equated with the head of the main clause, but it would also preclude other candidate arguments in the same sentence that would be mostly spurious and unnecessarily complicate (or even bias) the search problem. Secondly, some applications such as extractive summarization extract complete sentences for inclusion in summaries. Thus, it is useful to explore as a first approximation how well a sentence-based representation of arguments would fare in the discourse parsing task.

3.2. Classifying Connectives by Arg1 Location

Like Elwell and Baldridge (2008), we believe that different connectives are subject to different constraints. However, we don't believe that the syntax of connectives fully captures these constraints, specifically for the purpose of identifying their arguments. In particular, *connectives from different syntactic classes share properties with respect to their Arg1 location*. For example, the arguments of both subordinating conjunctions (e.g., *because*, *when*) and sentence-medial coordinating conjunctions (e.g., *and*, *but*, *or*) will both be found in the same sentence as the connective (and identified through tree-subtraction heuristics (Dinesh et al., 2005)). The same is not true of sentence-initial coordinating conjunctions, whose Arg1s are located in a different sentence. This in fact led Wellner (2009) to treat sentence-initial *But* as a discourse adverbial, and sentence-medial *but* as a coordinating conjunction. And finally, Arg1s of discourse adverbials are located mostly in different sentences, but they can also be found in the same sentence as the connective.

We therefore propose that explicit connectives should be classified in terms of their expected sentence collocation with their Arg1s, that is, those that take both their arguments in the same sentence (henceforth, *intra-sentential connectives*), and those whose arguments are found in different sentences (henceforth, *inter-sentential connectives*). This kind of partitioning has also been explored in Webber (2009). We note that this classification can be made reliably based partly on their syntactic class (for subordinating conjunctions) and partly on their sentential position (for coordinating conjunctions). Since syntactic class and sentential position alone are not reliable for distinguishing discourse adverbials, however, we experimented with a simple binary classifier to classify intra-sentential discourse adverbials from inter-sentential ones. We used as features the connective head, connective position and syntactic path from the connective to the root of the sentence. Our preliminary results show that the discourse adverbials can be classified for their two types of Arg1 location with high accuracy (93%, with an 86% baseline for Arg1 being located in a different sentence).

In addition, we propose that inter-sentential connectives should be further partitioned into two classes based on their paragraph position — those that appear in a paragraph-initial sentence (henceforth, *ParaInit* connectives) and those that appear elsewhere in the paragraph (henceforth, *ParaNonInit* connectives). This is because of the suggested role of paragraphs as the high-level "building blocks" of a discourse, with each paragraph defining a particular "local focus" within the overall topic of the text and exhibiting a coherent organization of the sentences around that focus (Hearst, 1997). Thus, we hypothesize that ParaNonInit connectives and their Arg1s would more likely be collocated in the same paragraph, while the Arg1 of a ParaInit connective can only occur in a previous paragraph. This is indeed what we find in the PDTB: 98% (4301/4373) of the ParaNonInit connectives are collocated with their Arg1 in the same paragraph. Ex. (1) and Ex. (2) illustrate such connectives.

We also hypothesize that ParaInit connectives would pose a more significant challenge than ParaNonInit connectives, since new paragraphs are often motivated by a new focus that may be linked to some topic or entity mentioned *anywhere* in the prior text. This hypothesis is also confirmed in the PDTB: While 91% (3962/4373) of the time, Arg1 of ParaNonInit connectives is the previous sentence, this is true only 49% (1110/2243) of the time for Arg1 of ParaInit connectives. In addition, although the Arg1 of ParaInit connectives is in the immediately preceding paragraph 79% (1767/2243) of the time, it is also not the case that the first sentence of this paragraph is selected more often as Arg1. Looking at the cases where such paragraphs contain more than one sentence (1348/2243), we find that this turns out to be even less frequent, i.e., 36% (166/1348). Ex. (3) illustrates a case where the Arg1 of ParaInit *In addition* is located medially in the previous paragraph.

(3) Countries in the region also are beginning to consider a framework for closer economic and political ties. *The economic and foreign ministers of 12 Asian and Pacific nations will meet in Australia next week to discuss global trade issues as well as regional matters such as transportation and telecommunications.* Participants will include the U.S., Australia, Canada, Japan, South Korea and New Zealand as well as the six members of the Association of Southeast Asian Nations – Thailand, Malaysia, Singapore, Indonesia, the Philippines and Brunei.

In addition, **the U.S. this year offered its own plan for cooperation around the Pacific rim in a major speech by Secretary of State James Baker, following up a proposal made in January by Australian Prime Minister Bob Hawke.** [wsj_0043]

These distributions confirm our hypothesis that the problem of identifying Arg1 of ParaInit connectives is much harder than for ParaNonInit connectives, and confirms the value of partitioning the inter-sentential connectives into ParaInit and ParaNonInit classes. Overall, our approach for classifying connectives captures the idea that *discourse relations are structured differently as one progresses from the sentence to the paragraph, which is a coherent grouping of sentences, and then from the paragraph to the entire text, which is a coherent grouping of paragraphs.*

4. Identifying Arguments of Connectives

Given our approach for representing arguments and classifying connectives as described in Section 3., the remainder of this paper is focussed on identifying the Arg1 sentence of *Inter-sentential ParaNoninit connectives*. This class primarily comprises inter-sentential discourse adverbials,

but also contains sentence-initial coordinating conjunctions and rare occurrences of subordinating conjunctions. We explored several heuristics for identifying the Arg1 sentences of these connectives and developed an algorithm based on these heuristics. One heuristic involved the use of coreference chains, for which we used the OntoNotes-2.0 coreference annotations (Weischedel et al., 2007),[3]. The source corpus for OntoNotes-2.0 partially overlaps the source corpus for the PDTB, namely the Wall Street Journal (WSJ) corpus, and it is this overlapping portion (598 WSJ texts) that we used in order to take advantage of the OntoNotes coreference annotation. For the connectives of interest here, the overlapping portion yielded 743 tokens of connectives along with their arguments.[4]

The rest of this section first describes our algorithm, which consists of a component for filtering potential candidate Arg1 sentences (Arg1Ss) and a component for ranking and evaluating the candidate Arg1Ss. We then present the results we obtained from a manual application of the algorithm.[5]

4.1. Filtering Potential Candidate Arg1 Sentences

The problem of identifying the Arg1S of a connective starts with the creation of a set of potential candidate sentences. For any Inter-sentential ParaNonInit connective, this consists of all and only the sentences appearing prior to the connective's sentence within the same paragraph. After this set is created, some candidates are filtered out according to the criteria described below.

4.1.1. Connectives in Opaque Direct Speech Segments

We define *direct speech segments* (DS segments) as segments containing one or more sentences appearing as quoted speech within quotation marks, with the speaker source, or speech attribution (Prasad et al., 2007), explicitly specified at most once for all the included sentences. Thus, in Ex. (4), there are two direct speech segments, DS1 and DS2, shown with square brackets and subscripts.[6] DS2 illustrates a common property of WSJ texts — having direct speech sentences distributed across multiple sentences, when all such sentences are enclosed within a single beginning and end quotation, and the atttribution, if explicit, is associated with any one of the sentences. In DS2, the attribution ("he predicts") is associated with its final sentence.

(4) Butch McCarty, who sells oil-field equipment for Davis Tool Co., is also busy. A native of the area, he is back now after riding the oil-field boom to the top, then surviving the bust running an Oklahoma City convenience store. $[_{DS1}$ "First year I came back there wasn't any work," he says.$]_{DS1}$ $[_{DS2}$ *"I think it's on the way back now.* But **it won't be a boom again.** No major booms, no major setbacks," he predicts.$]_{DS2}$ [wsj_0725]

Because quoted speech in WSJ-style texts can assume quite complex forms, it is necessary to define how to determine the boundaries of DS-segments. One question that arises is whether to treat adjacent segments like DS1 and DS2 in Ex. (4) as a single segment or as two distinct segments. For the task of discourse parsing, we adopted the more restrictive strategy of *associating a DS-segment with at most one explicit mention of attribution*. Thus, although DS1 and DS2 are adjacent in the text in Ex. (4), and both have the same attribution (i.e., Butch McCarty, referred to with the pronoun "he" in both segments), they are nevertheless treated as two distinct segments.

DS segments need not have explicit attributions, as can be seen for the DS1 segment in Ex. (5). Importantly, note that although the attribution for DS1 is inferred from the previous sentence where the quoted speech is indirect, these two sentences are not grouped together into a single segment.

(5) Corporations and museums are among the serious buyers, giving greater market stability, says Robert Persky of the Photograph Collector. $[_{DS1}$ "When I see prints going into the hands of institutions, I know they aren't going to come back on the market."$]_{DS1}$ Most in demand: classic photographs by masters such as Stieglitz and Man Ray. [wsj_0120]

In research related to Centering Theory (Grosz et al., 1995), it has been argued that the referential mechanisms of a discourse appearing in a direct speech segment should be determined independently of the text surrounding it (Kameyama, 1998). We extend this idea to discourse relations as well. In particular, we hypothesize that *DS segments close off the scope for the interpretation of discourse connectives*. For our Arg1 identification algorithm, this means that for connectives appearing within DS segments, sentences that do not also appear in the same DS segment as the connective are filtered out from the potential candidate set. Thus, for the connective *But* in Ex. (4), although the potential candidate set contains the previous four sentences in the paragraph, all but the immediately previous sentence are filtered out.[7]

Note that the DS segment filter only applies to connectives in non-initial sentences of such segments. Connectives that appear in initial sentences of such segments are treated like connectives appearing in non-DS segments.

4.1.2. Connectives in Opaque Parenthesized Segments

Just like DS segments, parenthesized segments, identifiable by enclosing parentheses, are often distributed over multi-

[3] LDC Catalog Entry LDC2008T04.

[4] The actual number of connectives appearing in the overlapping portion of PDTB and OntoNotes is actually greater than 743. However, for this paper, we have ignored connectives whose Arg1 spans multiple sentences, as well as the few connectives whose Arg1 sentence was not in the same paragraph as the connective. We have also excluded connectives which appeared in the second sentence of the paragraph, since these would trivially select the immediately previous sentence, i.e., the first sentence of the paragraph, as Arg1.

[5] Although we began this work with the goal of automating the algorithm, we faced some challenges in automatically detecting direct speech segments, which our algorithm requires (Section 4.1.1.). We plan to tackle this task again in future work.

[6] In all examples henceforth, we show all sentences from the beginning of the paragraph upto the connective's sentence.

[7] This represents the case of the filter yielding a single candidate with which Arg1 is trivially (and correctly) identified. However, not all cases are like this, and further heuristics may need to be applied after the application of this filter, to decide between remaining candidates.

ple sentences, as shown in Ex. (6).

(6) When Anne Volokh and her family immigrated to the U.S. 14 years ago, they started life in Los Angeles with only $400. They'd actually left the Soviet Union with $480, but during a stop in Italy Ms. Volokh dropped $80 on a black velvet suit. Not surprisingly, she quickly adapted to the American way. Three months after she arrived in L.A. she spent $120 she didn't have for a hat. ("A turban," she specifies, *"though it wasn't the time for that 14 years ago.* But **I loved turbans."**) [wsj_1367]

Like DS segments, we hypothesize that *parenthesized segments also close off the scope for the interpretation of connectives*, and the set of potential candidates for connectives appearing in such segments are filtered to exclude sentences that belong outside the connectives' parenthesized segment. Thus, the set of five candidate sentences for the connective *But* in Ex. (6) is reduced to just the one prior sentence that appears in the same parenthesized segment as the connective. Also, as for DS segments, this filter applies only to connectives in non-initial sentences of such segments.

4.1.3. Connectives Outside Opaque Zones

For connectives that appear in non-opaque zones, or in the initial sentences of opaque zones, all prior sentences appearing in other opaque zones are excluded from the set of potential candidates. This filter was implemented as a natural extension of the two previous filters, in that the opacity of DS segments and parenthesized segments also renders them unavailable for the interpretation of connectives outside those segments. But a more direct motivation for this filter comes from the hypothesis that DS segments and parenthesized segments in discourse are most often used to present elaborations, digressions, or background information. Structurally, they create embedded segments in the discourse that connectives outside these segments are unlikely to take as arguments. Ex. (7) and Ex. (8) show the application of this filter. In Ex. (7), the set of two potential candidates for the connective *But* is reduced to one after filtering out the immediately previous sentence which constitutes a DS segment. Likewise, in Ex. (8), the set of two potential candidates for the connective *though* is filtered to exclude the immediately previous sentence which constitutes a parenthesized segment.

(7) *Big buyers like Procter & Gamble say there are other spots on the globe, and in India, where the seed could be grown.* "It's not a crop that can't be doubled or tripled," says Mr. Krishnamurthy. But **no one has made a serious effort to transplant the crop.** [wsj_0515]

(8) *If all of this seems a little stale, it's redeemed in part by some tricky plot twists: The usual suspects are found to be guilty, then not guilty, then guilty – but of a different crime.* (In last week's rape case, for example, the girl turns out to have been a victim of incest, and the biggest villains are the politicians who exploit the case.) **Most of all** though, **the show is redeemed by the character of Mancuso.** [wsj_1397]

4.1.4. Exclusions Beyond Contrastive Sentences

Different classes of connectives are compatible with different discourse structures. Here, we explored the role of contrastive connectives in defining rhetorical zones in the discourse. In particular, we hypothesized that a sentence-initial contrastive connective like *but* or *however* introduces a new rhetorical zone into a paragraph that limits the argument possibilities of subsequent connectives. For our algorithm, this means that sentences prior to a sentence-initial contrastive connective are barred from the candidate set of a subsequent connective.

As an example, consider the connective *So* in Ex. (9). It has four sentences in its potential candidate set. However, since the immediately preceding sentence starts with the contrastive connective (*But*), all earlier ones are filtered out from this set. The same filtering is done for the connective *but* in the last sentence of this example (annotation not shown). The contrast filter will again exclude sentences appearing before the prior contrastive sentence, although here the filter would yield two sentences in the candidate set instead of the one candidate obtained for *so*.

(9) Which brings up the worst and meanest ghost of all – the ghost of the shah of Iran. When the shah died, President Carter was so scared that the shah's ghost would blame him for shoving him out to make way for the ayatollah that he declared the Carter Doctrine. Mr. Carter said he would go to war to stop anyone from trying to grab Iran. *But that ghost wouldn't settle for words, he wanted money and people – lots.* So **Mr. Carter formed three new Army divisions and gave them to a new bureaucracy in Tampa called the Rapid Deployment Force.** But that ghost wasn't fooled; he knew the RDF was neither rapid nor deployable nor a force – even though it cost $8 billion or $10 billion a year. [wsj_2112]

4.1.5. Interaction of Filters

Except for the contrast filter, all the other three filters are mutually exclusive for any given connective. The contrast filter, on the other hand, is applied on the result of each of these three filters. Thus, in each case, the resulting candidate set may be further reduced on application of the contrast filter. An example of such an interaction is shown in Ex. (10) for the connective *Still*. Here, the filter to exclude opaque segments (Section 4.1.3.) applies on the initial set of potential candidates and excludes the parenthesized sentence from the set. Subsequent to this, the application of the contrast filter identifies the second sentence of the paragraph as a contrastive sentence and thus excludes the paragraph's first sentence from the candidate set.

(10) We had great success in Somalia. But then it turned out that President Siad Barrah was not at all a nice person and the Navy pointed out that *the base he promised us in Berbera had silted up about a hundred years ago and anyway was 1,244 miles from the mouth of the Gulf.* (But who's counting.) Still, **Berbera was the best we could get, so we stay in bed with President Barrah.** [wsj_2112]

This example also illustrates the important fact that the application of the filters is *ordered* in that the contrast filter is applied only after the other three filters have been applied. Thus, it cannot be the case that the contrastive sentence within the parenthesized segment leads to the exclusion of all the prior sentences from the candidate set. In this case,

(11d): [Fed] > (11b): [Fed]
(11d): [the mutinous Fed member] > (11b): [A " senior Fed official "]
(11d): [the chairman 's] > (11c): [Chairman Greenspan]
(11d): [the chairman 's decision *PRO* to remain silent] > (11a): [Mr. Greenspan 's decision *PRO* to keep quiet]

Figure 1: Coreference Chains for Ex. (11)

CorefA: If the entity mention in the connective's sentence has a pronominal form, Arg1 is the first sentence linked via the coreference chain for this entity. As long as such entities are present in the entity set, this rule will always yield an Arg1S on the very first evaluation.

CorefB: If the entity mention in the connective's sentence has a non-pronominal form, then Arg1 is the first sentence that has a non-pronominal mention of the entity in the coreference chain for that entity. This rule may fail to yield an Arg1S since there may not be any non-pronominal mentions in the coreference chains.

Figure 2: Evaluation Rules for Coreference Chains

this would have led to the exclusion of the correct Arg1S from the candidate set of the connective.

4.2. Ranking Candidate Arg1 Sentences Using Coreference

The filtering process as described in Section 4.1. yields one or more candidates in the set of potential candidates. If there is only one, this is simply selected as Arg1S. Otherwise, a coreference-based decision procedure is implemented to rank the multiple candidate Arg1Ss. The close interaction of discourse structure and coreference has been proposed and studied by several researchers, although it is somewhat of a "chicken and egg" problem. While some argue that anaphora resolution is dependent on discourse structure (Cristea et al., 2000; Asher and Lascarides, 2003), arguments for the reverse dependence have also been made (Seretan and Cristea, 2002). In our approach, we take the latter position. Furthermore, from a practical point of view, exploring the role of coreference for discourse structuring is more reasonable rather than the other way around, since the state of the technology in coreference resolution is currently ahead of that in discourse parsing.

Before using the coreference information in our algorithm, we manually examined the annotations provided by OntoNotes for the paragraphs in which our connectives appeared. Since our dataset is small and because the goal of this study was to explore the importance of various discourse features rather than build a state-of-the-art discourse parsing system, we wanted to ensure high reliability for the input features. We found that we needed to augment the coreference annotation along the following lines: (a) correcting annotations which were in error given the OntoNotes guidelines; (b) adding annotations that were missing, given the OntoNotes guidelines; (c) annotating certain bare plurals that corefer to "specific" classes of entities; and (d) annotating set-instance anaphoric relations.

We applied the following coreference-based ranking of candidate Arg1Ss on our augmented version of OntoNotes (henceforth called OntoNotes+). For each connective, the set of entities mentioned in its own sentence were extracted from OntoNotes+ and ordered according to their string order within the sentence. Then, for each entity in the set, backward-looking coreference chains were created over the sentences remaining in the candidate set for that connective. If an entity was mentioned more than once in a sentence, only its first string-wise mention was recorded. If no coreference chains were retrieved for an entity, it was discarded from the entity set. We illustrate this with Ex. (11), where the target connective *And* is in the very last sentence. Four entities were identified in this sentence, shown in Fig. (1) - vertical order reflecting the string order of their mentions. Since none of the potential candidates are excluded by the filters, all are available for coreference chains, which yields non-empty chains for all entities. In the coreference chains shown in the figure, > indicates a coreference link between entities in different sentences. For example, the first chain in the figure shows that "Fed" in (11d) corefers with "Fed" in (11b).

(11) a. Mr. Greenspan's decision to keep quiet also prompted a near-mutiny within the Fed's ranks.

b. *A "senior Fed official" spoke on Saturday after the market swoon to both the Washington Post and the New York Times, saying the Fed was prepared to provide as much credit as the markets needed.*

c. The statement angered Chairman Greenspan, but it was greeted with applause by the Bush administration and the financial markets.

d. And, **while the mutinous Fed member hasn't gone public, some Fed governors, most notably Vice Chairman Manuel Johnson, are known to have disagreed with the chairman's decision to remain silent.**

For identifying Arg1, coreference chains are evaluated in the given order according to the two mutually exclusive rules shown in Fig. 2. The first chain that a rule succeeds on is selected and the search terminated. If the evaluation fails overall, Arg1S is resolved by default to the sentence immediately preceding the connective. The coreference evaluation rules are partially inspired by Centering Theory constraints on the realization of anaphoric expressions in local discourse segments (Grosz et al., 1995).

With respect to Ex. (11), we find that the CorefB rule needs to be applied, since the first entity mention in the connec-

tive's sentence ("Fed") has a non-pronominal form. Tracing its coreference chain leads to the identification of (11b) as Arg1, which is the correct resolution in this case.

Ex. (12) illustrates a case where CorefB again applies but sentence (12b) in the coreference chain is rejected as Arg1S because it only contains pronominal mentions of the entity. The single coreference chain for Ex. (12) is shown in Fig. 3, with links to the two sentences, (12b) and (12a). Although (12b) is the closer candidate, it is rejected because the entity is mentioned with a pronominal form. The CorefB rule therefore moves back along the chain and finds that the entity is mentioned with a non-pronominal form in sentence (12a), which is therefore (correctly) selected as Arg1.

(12) a. *The framers hardly discussed the appropriations clause at the Constitutional Convention of 1787, according to Madison's notes.*

b. To the extent they did, their concern was to ensure fiscal accountability.

c. Moreover, **the framers believed that the nation needed a unitary executive with the independence and resources to perform the executive functions that the Confederation Congress had performed poorly under the Articles of Confederation.** [wsj_0112]

(12c): [the framers] > (12b): [they] > (12a): [The framers]

Figure 3: Coreference Chains for Ex. (12)

5. Results and Discussion

The first author applied the filtering, ranking and evaluation heuristics to the full set of 743 tokens in the overlapping portion of PDTB and OntoNotes. Coreference chains came from the augmented OntoNotes+, as described earlier. The baseline for comparison was selection of the immediately previous sentence as Arg1, which was 83.3% (619/743) in our data set.[8] We acheived an overall accuracy of 86.3%, a 3% improvement over the baseline.

Since the data set is very unbalanced and highly skewed towards the baseline, we also created a confusion matrix for the results, to look at how the algorithm performed for cases with immediately previous sentence Arg1s (IPS) versus cases with Arg1s in non-adjacent sentences (NAPS). The confusion matrix shown in Table 1 shows that a significant proportion of both IPS (88%) as well as NAPS Arg1s (79%) are correctly identified.

	IPS-P	NAPS-P	Total
IPS-A	543 (88%)	76	619
NAPS-A	26	98 (79%)	124

Table 1: Confusion Matrix. Rows show actual classes ("-A"). Columns show predicted classes ("-P")

[8]Note that because we are using a subset of the PDTB in our analysis, the baseline in our data set is different from the baseline over the entire corpus, which is 91% (see Section 3.1.).

There are two forms of error. The first type (12%) reflect what we believe might be errors and misannotations in the PDTB annotation, which is to be expected as part of annotation noise in any corpora. In Ex. (13), for instance, both sentences prior to the connective *And* are available as candidates for coreference ranking since none of the filters apply, and there is a single coreference chain from the connective's sentence, with a single link between "Fed" in the last sentence and "Fed" in the first sentence. Although the CorefB rule would use this information to select the first sentence as Arg1, it is the second sentence that is annotated as Arg1 in PDTB. However, a closer look at the annotation shows that the second sentence is an elaboration of the first sentence, which ought to have made the first sentence a more appropriate choice for Arg1.

(13) The Fed chairman's caution was apparent again on the Monday morning after the market's plunge, when the central bank took only modest steps to aid the markets. *A surprisingly small amount of reserves was added to the banking system.* And, **by the end of that week, the key federal funds interest rate, which is largely controlled by the Fed, had settled at 8.75%, barely changed from the level of just under 9% that prevailed the previous week.** [wsj_0598]

The remaining errors occurred because the algorithm simply failed to work for the particular case. In Ex. (14), for instance, none of the filters lead to exclusion of any Arg1 candidates. Further, no coreference chains are found for the connective's sentence, as a result of which the algorithm defaults to the immediately previous sentence as Arg1. However, the correct Arg1 is two sentences away, as shown in the example.

(14) *Thousands of East Germans fled to Czechoslovakia after the East Berlin government lifted travel restrictions.* The ban on cross-border movement was imposed last month after a massive exodus of emigres to West Germany. Also, **a Communist official for the first time said the future of the Berlin Wall could be open to discussion.** [wsj_0174]

This suggests that additional heuristics or modifications to the heuristics might be needed to account for the full set of cases. We believe that the coreference ranking and evaluation, in particular, needs further investigation. We also believe that we need a more sophisticated account and, hence, annotation of coreference and anaphoric relations in general. Although OntoNotes overcomes some limitations of previously coreference annotated corpora (e.g., MUC-6, MUC-7, and ACE corpora) by annotating reference to events, for example, there are arguably further gaps to be filled for a proper treatment of anaphoric relations in discourse (Poesio and Artstein, 2008).

6. Conclusion and Future Work

In the context of work in shallow discourse parsing for identifying connectives and their arguments based on the Penn Discourse Treebank corpus, the first important decision is how to represent them. In contrast to previous approaches, we have proposed a novel approach that represents arguments as the sentences containing them, and

classifies connectives in terms of their expected collocation with their arguments in sentences and paragraphs. In addition to being practically useful, our approach is also theoretically and empirically well-founded. Following our approach, we have also developed a heuristics-based method for identifying the arguments of connectives, focussing on more challenging Arg1, which can be arbitrarily far from its connective. Our heuristics capture well-founded scope constraints and coreference preferences in discourse for the interpretation of discourse relations. Compared to a high baseline of selecting the immediately previous sentence as Arg1, our manual application of the algorithm on the data showed an absolute 3% improvement, showing that the proposed approach and method holds promise. However, the error analysis shows room for improvement, since most errors are due to failure of the algorithm. We believe that the coreference heuristics need further investigation, which will be carried out shortly. Also planned is the automation of our method, where the difficulty to date has been automatically detecting direct speech segments, as the scope-based heuristics require.

Acknowledgements

This work was partially supported by NSF grants EIA-05-63063, and IIS-07-05671. We would like to thank Partha Pratim Talukdar for participating in early experiments towards this work, and Geraud Campion and Shai Nir for help with the data preparation.

7. References

Nicholas Asher and Alex Lascarides. 2003. *Logics of conversation*. Cambridge University Press.

Jason Baldridge, Nicholas Asher, and Julie Hunter. 2007. Annotation for and robust parsing of discourse structure on unrestricted texts. *Zeitschrift fur Sprachwissenschaft*, 26:213–239.

Dan Cristea, Nancy Ide, Daniel Marcu, and Valentin Tablan. 2000. An empirical investigation of the relation between discourse structure and co-reference. In *Proceedings of the 18th conference on Computational linguistics*, Morristown, NJ, USA. Association for Computational Linguistics.

Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2005. Attribution and the (non)-alignment of syntactic and discourse arguments of connectives. In *Proceedings of the ACL Workshop on Frontiers in Corpus Annotation II: Pie in the Sky*, Ann Arbor, Michigan.

Robert Elwell and Jason Baldridge. 2008. Discourse connective argument identification with connective specific rankers. In *Proceedings of ICSC-2008*.

Kate Forbes, Eleni Miltsakaki, Rashmi Prasad, Anoop Sarkar, Aravind Joshi, and Bonnie Webber. 2003. D-LTAG system: Discourse parsing with a lexicalized tree adjoining grammar. *Journal of Logic, Language and Information*, 12(3).

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

Marti A. Hearst. 1997. TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64.

Megumi Kameyama. 1998. Intrasentential centering: A case study. In M.A. Walker, A.K. Joshi, and E.F. Prince, editors, *Centering Theory in Discourse*, pages 89–112. Oxford University Press, Oxford, U.K.

Alan Lee, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2008. Departures from tree structures in discourse: Shared arguments in the Penn Discourse Treebank. In *Proceedings of the Constraints in Discourse III Workshop*, Potsdam, Germany.

Daniel Marcu. 1997. The rhetorical parsing of natural language texts. In *Proceedings of the Meeting of the Association for Computational Linguistics*.

Massimo Poesio and Ron Artstein. 2008. Anaphoric annotation in the ARRAU corpus. In *Proceedings of 6th International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco.

Livia Polanyi, Chris Culy, Martin H. van den Berg, Gian Lorenzo Thione, and David Ahn. 2004. A rule based approach to discourse parsing. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*, Cambridge, MA.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Aravind Joshi, and Bonnie Webber. 2007. Attribution and its annotation in the Penn Discourse TreeBank. *Traitement Automatique des Langues, Special Issue on Computational Approaches to Document and Discourse*, 47(2).

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of 6th International Conference on Language Resources and Evaluation (LREC 2008)*.

PDTB-Group. 2008. The Penn Discourse TreeBank 2.0 Annotation Manual. Technical Report IRCS-08-01, Institute for Research in Cognitive Science, University of Pennsylvania.

Violeta Seretan and Dan Cristea. 2002. The use of referential constrains in structuring discourse. In *Proceedings of The Third International Conference on Language Resources and Evaluation (LREC 2002)*, Las Palmas, Spain.

Bonnie Webber. 2009. Genre distinctions for discourse in the Penn TreeBank. In *Proceedings of the Association for Computational Linguistics*, Singapore.

Ralph Weischedel, Sameer Pradhan, Lance Ramshaw, Martha Palmer, Nianwen Xue, Mitchell Marcus, Ann Taylor, Craig Greenberg, Eduard Hovy, Robert Belvin, and Ann Houston. 2007. Ontonotes release 2.0. Technical report, Linguistic Data Consortium.

Ben Wellner and James Pustejovsky. 2007. Automatically identifiying the arguments of discourse connectives. In *Proceedings of EMNLP-CoNLL*.

Ben Wellner. 2009. *Sequence Models and Re-ranking Methods for Discourse Parsing*. Ph.D. thesis, Brandeis University, Boston, MA.