Probabilistic Head-Driven Parsing for Discourse Structure

Jason Baldridge and Alex Lascarides
School of Informatics
University of Edinburgh
2 Buccleuch Place
Edinburgh, EH8 9LW
Scotland, UK
{jbaldrid,alex}@inf.ed.ac.uk

Abstract

We describe a data-driven approach to building interpretable discourse structures for appointment scheduling dialogues. We represent discourse structures as headed trees and model them with probabilistic head-driven parsing techniques. We show that dialogue-based features regarding turn-taking and domain specific goals have a large positive impact on performance. Our best model achieves an f-score of 43.2% for labelled discourse relations and 67.9% for unlabelled ones, significantly beating a right-branching baseline that uses the most frequent relations.

1 Introduction

Achieving a model of discourse interpretation that is both robust and deep is a major challenge. Consider the dialogue in Figure 1 (the sentence numbers are from the Redwoods treebank (Oepen et al., 2002)). A robust and deep interpretation of it should resolve the anaphoric temporal description in utterance 154 to the twenty sixth of *July* in the afternoon. It should identify that time and before 3pm on the twenty-seventh as potential times to meet, while ruling out July thirtieth to August third. It should gracefully handle incomplete or ungrammatical utterances like 152 and recognise that utterances 151 and 152 have no overall effect on the time and place to meet.

According to Hobbs et al. (1993) and Asher and Lascarides (2003), a discourse structure consisting of hierarchical rhetorical connections between utterances is vital for providing a *unified model* of a wide range of anaphoric and intentional discourse phenomena, contributing to the interpretations of pronouns, temporal expressions, presuppositions and ellipses (among others), as well as influencing communicative goals. This suggests that a robust model of discourse structure could complement current robust interpretation systems, which tend to focus on only *one* aspect of the semantically ambiguous material, such as pronouns (e.g., Strübe and Müller (2003)), definite descriptions (e.g., Vieira and Poesio (2000)), or temporal expressions (e.g., Wiebe et al. (1998)). This specialization makes it hard to assess how they would perform in the context of a more comprehensive set of interpretation tasks.

149 PAM: *maybe we can get together, and, discuss, the planning, say, two hours, in the next, couple weeks,*
150 PAM: *let me know what your schedule is like.*
151 CAE: *okay, let me see.*
152 CAE: *twenty,*
153 CAE: *actually, July twenty sixth and twenty seventh looks good,*
154 CAE: *the twenty sixth afternoon,*
155 CAE: *or the twenty seventh, before three p.m., geez.*
156 CAE: *I am out of town the thirtieth through the,*
157 CAE: *the third, I am in San Francisco.*

Figure 1: A dialogue extract from Redwoods.

To date, most methods for constructing discourse structures are not robust. They typically rely on grammatical input and use symbolic methods which inevitably lack coverage. One exception is Marcu's work (Marcu, 1997, 1999) (see also Soricut and Marcu (2003) for constructing discourse structures for individual sentences). Marcu (1999) uses a decision-tree learner and shallow syntactic features

to create classifiers for discourse segmentation and for identifying rhetorical relations. Together, these amount to a model of discourse parsing. However, the results are trees of Rhetorical Structure Theory (RST) (Mann and Thompson, 1986), and the classifiers rely on well-formedness constraints on RST trees which are too restrictive (Moore and Pollack, 1992). Furthermore, RST does not offer an account of how compositional semantics gets augmented, nor does it model anaphora. It is also designed for monologue rather than dialogue, so it does not offer a precise semantics of questions or non-sentential utterances which convey propositional content (e.g., 154 and 155 in Figure 1). Another main approach to robust dialogue processing has been statistical models for identifying dialogue acts (e.g., Stolcke et al. (2000)). However, dialogue acts are properties of utterances rather than hierarchically arranged relations *between* them, so they do not provide a basis for resolving semantic underspecification generated by the grammar (Asher and Lascarides, 2003).

Here, we present the first probabilistic approach to parsing the discourse structure of dialogue. We use dialogues from Redwoods' appointment scheduling domain and adapt head-driven generative parsing strategies from sentential parsing (e.g., Collins (2003)) for discourse parsing. The discourse structures we build conform to Segmented Discourse Representation Theory (SDRT) (Asher and Lascarides, 2003). SDRT provides a precise dynamic semantic interpretation for its discourse structures which augments the conventional semantic representations that are built by most grammars. We thus view the task of learning a model of SDRT-style discourse structures as one step towards achieving the goal of robust and precise semantic interpretations.

We describe SDRT in the context of our domain in Section 2. Section 3 discusses how we encode and annotate discourse structures as headed trees for our domain. Section 4 provides background on probabilistic head-driven parsing models, and Section 5 describes how we adapt the approach for discourse and gives four models for discourse parsing. We report results in Section 6, which show the importance of dialogue-based features on performance. Our best model performs far better than a baseline that uses the most frequent rhetorical relations and right-branching segmentation.

h_0 : *Request-Elab*$(149, 150) \wedge$
Plan-Elab$(150, h_1)$
h_1 : *Elaboration*$(153, h_2) \wedge$
Continuation$(153, 156) \wedge$
Continuation$(156, 157)$
h_2 : *Alternation*$(154, 155)$

Figure 2: The SDRS for the dialogue in Figure 1.

2 Segmented Discourse Representation Theory

SDRT extends prior work in dynamic semantics (e.g., van Eijk and Kamp (1997)) via logical forms that feature rhetorical relations. The logical forms consist of *speech act discourse referents* which label content (either of a clause or of text segments). Rhetorical relations such as *Explanation* relate these referents. The resulting structures are called *segmented discourse representation structures* or SDRSs. An SDRS for the dialogue in Figure 1 is given in Figure 2; we have used the numbers of the elementary utterances from Redwoods as the speech act discourse referents but have omitted their labelled logical forms. Note that utterances 151 and 152, which do not contribute to the truth conditions of the dialogue, are absent – we return to this shortly.

There are several things to note about this SDRS. First, SDRT's dynamic semantics of rhetorical relations imposes constraints on the contents of its arguments. For example, *Plan-Elab*$(150, h1)$ (standing for *Plan-Elaboration*) means that $h1$ provides information from which the speaker of 150 can elaborate a plan to achieve their communicative goal (to meet for two hours in the next couple of weeks). The relation *Plan-Elab* contrasts with *Plan-Correction*, which would relate the utterances in dialogue (1):

(1) a. A: *Can we meet at the weekend?*
b. B: *I'm afraid I'm busy then.*

Plan-Correction holds when the content of the second utterance in the relation indicates that its communicative goals conflict with those of the first one. In this case, A indicates he wants to meet next weekend, and B indicates that he does not (note that *then* resolves to the weekend). Utterances (1ab) would also be related with *IQAP* (*Indirect Question Answer*

Pair): this means that (1b) provides sufficient information for the questioner A to infer a direct answer to his question (Asher and Lascarides, 2003).

The relation *Elaboration*$(153, h_2)$ in Figure 2 means that the segment 154 to 155 resolves to a proposition which elaborates part of the content of the proposition 153. Therefore *the twenty sixth* in 154 resolves to the twenty sixth of *July*—any other interpretation contradicts the truth conditional consequences of *Elaboration*. *Alternation*$(154, 155)$ has truth conditions similar to (dynamic) disjunction. *Continuation*$(156, 157)$ means that 156 and 157 have a common topic (here, this amounts to a proposition about when CAE is unavailable to meet).

The second thing to note about Figure 2 is how one rhetorical relation can outscope another: this creates a hierarchical segmentation of the discourse. For example, the second argument to the *Elaboration* relation is the label h_2 of the *Alternation*-segment relating 154 to 155. Due to the semantics of *Elaboration* and *Alternation*, this ensures that the dialogue entails that one of 154 or 155 is true, but it does not entail 154, nor 155.

Finally, observe that SDRT allows for a situation where an utterance connects to more than one subsequent utterance, as shown here with *Elaboration*$(153, h_2) \wedge$ *Continuation*$(153, 156)$. In fact, SDRT also allows two utterances to be related by multiple relations (see (1)) and it allows an utterance to rhetorically connect to multiple utterances in the context. These three features of SDRT capture the fact that an utterance can make more than one illocutionary contribution to the discourse. An example of the latter kind of structure is given in (2):

(2) a. A: *Shall we meet on Wednesday?*
b. A: *How about one pm?*
c. B: *Would one thirty be OK with you?*

The SDRS for this dialogue would feature the relations *Plan-Correction*$(2b, 2c)$, *IQAP*$(2b, 2c)$ and *Q-Elab*$(2a, 2c)$. *Q-Elab*, or *Question-Elaboration*, always takes a question as its second argument; any answers to the question must elaborate a plan to achieve the communicative goal underlying the first argument to the relation. From a logical perspective, recognising *Plan-Correction*$(2b, 2c)$ and *Q-Elab*$(2a, 2c)$ are co-dependent.

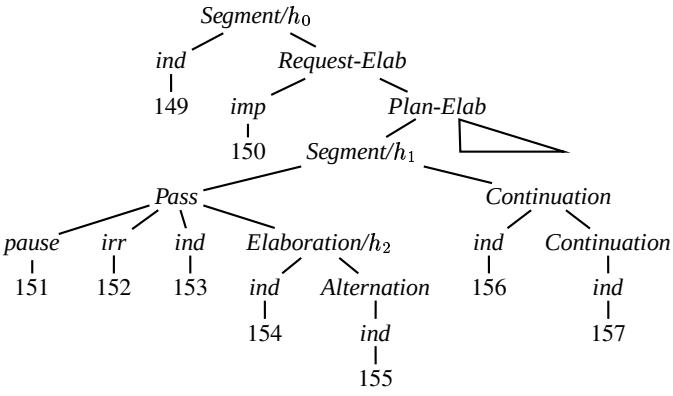

Figure 3: The discourse structure for the dialogue from Figure 1 in tree form.

3 Augmenting the Redwoods treebank with discourse structures

Our starting point is to create training material for probabilistic discourse parsers. For this, we have augmented dialogues from the Redwoods Treebank (Oepen et al., 2002) with their analyses within a fragment of SDRT (Baldridge and Lascarides, 2005). This is a very different effort from that being pursued for the Penn Discourse Treebank (Miltsakaki et al., 2004), which uses discourse connectives rather than abstract rhetorical relations like those in SDRT in order to provide theory neutral annotations. Our goal is instead to leverage the power of the semantics provided by SDRT's relations, and in particular to do so for dialogue as opposed to monologue.

Because the SDRS-representation scheme, as shown in Figure 2, uses graph structures that do not conform to tree constraints, it cannot be combined directly with statistical techniques from sentential parsing. We have therefore designed a headed tree encoding of SDRSs, which can be straightforwardly modeled with standard parsing techniques and from which SDRSs can be recovered.

For instance, the tree for the dialogue in Figure 1 is given in Figure 3. The SDRS in Figure 2 is recovered automatically from it. In this tree, utterances are leaves which are immediately dominated by their tag, indicating either the sentence mood (*ind*icative, *int*errogative or *imp*erative) or that it is *irr*elevant, a *pause* or a pleasantry (e.g., *hello*), annotated as *pls*. Each non-terminal node has a unique head daughter: this is either a *Segment* node, *Pass* node, or a

leaf utterance tagged with its sentence mood. Non-terminal nodes may in addition have any number of daughter *irr*, *pause* and *pls* nodes, and an additional daughter labelled with a rhetorical relation.

The notion of headedness has no status in the semantics of SDRSs themselves. The heads of these discourse trees are not like verbal heads with subcategorization requirements in syntax; here, they are nothing more than the left argument of a rhetorical relation, like 154 in *Alternation*$(154, 155)$. Nonetheless, defining one of the arguments of rhetorical relations as a head serves two main purposes. First, it enables a fully deterministic algorithm for recovering SDRSs from these trees. Second, it is also crucial for creating probabilistic head-driven parsing models for discourse structure.

Segment and *Pass* are non-rhetorical node types. The former explicitly groups multiple utterances. The latter allows its head daughter to enter into relations with segments higher in the tree. This allows us to represent situations where an utterance attaches to more than one subsequent utterance, such as 153 in dialogue (1). Annotators manually annotate the rhetorical relation, *Segment* and *Pass* nodes and determine their daughters. They also tag the individual utterances with one of the three sentence moods or *irr*, *pause* or *pls*. The labels for segments (e.g., h_0 and h_1 in Figure 3) are added automatically. Non-veridical relations such as *Alternation* also introduce segment labels on their parents; e.g., h_2 in Figure 3.

The SDRS is automatically recovered from this tree representation as follows. First, each relation node generates a rhetorical connection in the SDRS: its first argument is the discourse referent of its parent's head daughter, and the second is the discourse referent of the node itself (which unless stated otherwise is its head daughter's discourse referent). For example, the structure in Figure 3 yields *Request-Elab*$(149, 150)$, *Alternation*$(154, 155)$ and *Elaboration*$(153, h_2)$. The labels for the relations in the SDRS—which determine segmentation—must also be recovered. This is easily done: any node which has a segment label introduces an *outscopes* relation between that and the discourse referents of the node's daughters. This produces, for example, *outscopes*$(h_0, 149)$, *outscopes*$(h_1, 153)$ and *outscopes*$(h_2, 154)$. It is straightforward to determine the labels of *all* the rhetorical relations from these conditions. Utterances such as 151 and 152, which are attached with *pause* and *irr* to indicate that they have no overall truth conditional effect on the dialogue, are ignored when constructing the SDRS, so SDRT does not assign these terms any semantics. Overall, this algorithm generates the SDRS in Figure 2 from the tree in Figure 3.

Thus far, 70 dialogues have been annotated and reviewed to create our gold-standard corpus. On average, these dialogues have 237.5 words, 31.5 utterances, and 8.9 speaker turns. In all, there are 30 different rhetorical relations in the inventory for this annotation task, and 6 types of tags for the utterances themselves: *ind*, *int*, *imp*, *pause*, *irr* and *pls*.

Finally, we annotated all 6,000 utterances in the Verbmobil portion of Redwoods with the following: whether the time mentioned (if there is one) is a good time to meet (e.g., *I'm free then* or *Shall we meet at 2pm?*) or a bad time to meet (e.g., *I'm busy then* or *Let's avoid meeting at the weekend*). These are used as features in our model of discourse structure (see Section 5). We use these so as to minimise using directly detailed features from the utterances themselves (e.g. the fact that the utterance contains the word *free* or *busy*, or that it contains a negation), which would lead to sparse data problems given the size of our training corpus. We ultimately aim to *learn* good-time and bad-time from sentence-level features extracted from the 6,000 Redwoods analyses, but we leave this to future work.

4 Generative parsing models

There is a significant body of work on probabilistic parsing, especially that dealing with the English sentences found in the annotated Penn Treebank. One of the most important developments in this work is that of Collins (2003). Collins created several lexicalised head-driven generative parsing models that incorporate varying levels of structural information, such as distance features, the complement/adjunct distinction, subcategorization and gaps. These models are attractive for constructing our discourse trees, which contain heads that establish non-local dependencies in a manner similar to that in syntactic parsing. Also, the co-dependent tasks of determining segmentation and choosing the rhetorical connections are both heavily influenced by the content of

the utterances/segments which are being considered, and lexicalisation allows the model to probabilistically relate such utterances/segments very directly.

Probabilistic Context Free Grammars (PCFGs) determine the conditional probability of a right-hand side of a rule given the left-hand side, $\mathcal{P}(RHS|LHS)$. Collins instead decomposes the calculation of such probabilities by first generating a head and then generating its left and right modifiers independently. In a supervised setting, doing this gathers a much larger set of rules from a set of labelled data than a standard PCFG, which learns only rules that are directly observed.[1]

The decomposition of a rule begins by noting that rules in a lexicalised PCFG have the form:

$$P(h) \rightarrow L_n(l_n) \ldots L_1(l_1) H(h) R_1(r_1) \ldots R_m(r_m)$$

where h is the head word, $H(h)$ is the label of the head constituent, $P(h)$ is its parent, and $L_i(l_i)$ and $R_i(r_i)$ are the n left and m right modifiers, respectively. It is also necessary to include $STOP$ symbols L_{n+1} and R_{m+1} on either side to allow the Markov process to properly model the sequences of modifiers. By assuming these modifiers are generated independently of each other but are dependent on the head and its parent, the probability of such expansions can be calculated as follows (where $\mathcal{P}_h$, $\mathcal{P}_l$ and $\mathcal{P}_r$ are the probabilities for the head, left-modifiers and right-modifiers respectively):

$$\begin{aligned}
&\mathcal{P}(L_n(l_n) \ldots L_1(l_1) H(h) R_1(r_1) \ldots R_m(r_m) | P(h)) = \\
&\quad \mathcal{P}_h(H|P(h)) \\
&\quad \times \prod_{i=1 \ldots n+1} \mathcal{P}_l(L_i(l_i)|P(h), H) \\
&\quad \times \prod_{i=1 \ldots m+1} \mathcal{P}_r(R_i(r_i)|P(h), H)
\end{aligned}$$

This provides the simplest of models. More conditioning information can of course be added from any structure which has already been generated. For example, Collins' model 1 adds a distance feature that indicates whether the head and modifier it is generating are adjacent and whether a verb is in the string between the head and the modifier.

[1] A similar effect can be achieved by converting n-ary trees to binary form.

5 Discourse parsing models

In Section 3, we outlined how SDRSs can be represented as headed trees. This allows us to create parsing models for discourse that are directly inspired by those described in the previous section. These models are well suited for our discourse parsing task. They are lexicalised, so there is a clear place in the discourse model for incorporating features from utterances: simply replace lexical heads with whole utterances, and exploit features from those utterances in discourse parsing in the same manner as lexical features are used in sentential parsing.

Discourse trees contain a much wider variety of kinds of information than syntactic trees. The leaves of these trees are sentences with full syntactic and semantic analyses, rather than words. Furthermore, each dialogue has two speakers, and speaker style can change dramatically from dialogue to dialogue. Nonetheless, the task is also more constrained in that there are fewer overall constituent labels, there are only a few labels which can act as heads, and trees are essentially binary branching apart from constituents containing ignorable utterances.

The basic features we use are very similar to those for the syntactic parsing model given in the previous section. The feature P is the parent label that is the starting point for generating the head and its modifiers. H is the label of the head constituent. The tag t is also used, except that rather than being a part-of-speech, it is either a sentence mood label (*ind*, *int*, or *imp*) or an ignorable label (*irr*, *pls*, or *pause*). The word feature w in our model is the first discourse cue phrase present in the utterance.[2] In the absence of a cue phrase, w is the empty string. The distance feature Δ is true if the modifier being generated is adjacent to the head and false otherwise. To incorporate a larger context into the conditioning information, we also utilize a feature HCR, which encodes the child relation of a node's head.

We have two features that are particular to dialogue. The first ST, indicates whether the head utterance of a segment starts a turn or not. The other, TC, encodes the number of turn changes within a segment with one of the values 0, 1, or ≥ 2.

Finally, we use the good/bad-time annotations discussed in Section 3 for a feature TM indicating

[2] We obtained our list of cue phrases from Oates (2001).

	Head features							Modifier features								
	P	*t*	*w*	*HCR*	*ST*	*TC*	*TM*	*P*	*t*	*w*	*H*	Δ	*HCR*	*ST*	*TC*	*TM*
Model 1	✓	✓	✓					✓	✓	✓	✓	✓				
Model 2	✓	✓	✓	✓	✓			✓	✓	✓	✓	✓	✓	✓		
Model 3	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	
Model 4	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Figure 4: The features active for determining the head and modifier probabilities in each of the four models.

one of the following values for the head utterance of a segment: *good_time*, *bad_time*, *neither*, or *both*.

With these features, we create the four models given in Figure 4. As example feature values, consider the *Segment* node labelled h_1 in Figure 3. Here, the features have as values: P=*Segment*, H=*Pass*, t=*ind* (the tag of utterance 153), w=*Actually* (see 153 in Figure 1), HCR=*Elaboration*, ST=*false*, TC=0, and TM=*good_time*.

As is standard, linear interpolation with back-off levels of decreasing specificity is used for smoothing. Weights for the levels are determined as in Collins (2003).

6 Results

For our experiments, we use a standard chart parsing algorithm with beam search that allows a maximum of 500 edges per cell. The figure of merit for the cut-off combines the probability of an edge with the prior probability of its label, head and head tag. Hypothesized trees that do not conform to some simple discourse tree constraints are also pruned.[3]

The parser is given the elementary discourse units as defined in the corpus. These units correspond directly to the utterances already defined in Redwoods and we can thus easily access their complete syntactic analyses directly from the treebank.

The parser is also given the correct utterance moods to start with. This is akin to getting the correct part-of-speech tags in syntactic parsing. We do this since we are using the parser for semi-automated annotation. Tagging moods for a new discourse is a very quick and reliable task for the human. With them the parser can produce the more complex hierarchical structure more accurately than if it had to guess them – with the potential to dramatically reduce the time to annotate the discourse structures of further dialogues. Later, we will create a sentence mood tagger that presents an n-best list for the parser to start with, from the tag set *ind, int, imp, irr, pause,* and *pls*.

Models are evaluated by using a leave-one-out strategy, in which each dialogue is parsed after training on all the others. We measure labelled and unlabelled performance with both the standard PARSEVAL metric for comparing spans in trees and a relation-based metric that compares the SDRS's produced by the trees. The latter gives a more direct indication of the accuracy of the actual discourse logical form, but we include the former to show performance using a more standard measure. Scores are globally determined rather than averaged over all individual dialogues.

For the relations metric, the relations from the derived discourse tree for the test dialogue are extracted; then, the overlap with relations from the corresponding gold standard tree is measured. For labelled performance, the model is awarded a point for a span or relation which has the correct discourse relation label and both arguments are correct. For unlabelled, only the arguments need to be correct.[4]

Figure 5 provides the f-scores[5] of the various models and compares them against those of a baseline model and annotators. All differences between models are significant, using a pair-wise t-test at 99.5% confidence, except that between the baseline and Model 2 for unlabelled relations.

The baseline model is based on the most frequent way of attaching the current utterance to its dia-

[3] E.g., nodes can have at most one child with a relation label.

[4] This is a much stricter measure than one which measures relations between a head and its dependents in syntax because it requires two *segments* rather than two heads to be related correctly. For example, Model 4's labelled and unlabelled relation f-scores using segments are 43.2% and 67.9%, respectively; on a head-to-head basis, they rise to 50.4% and 81.8%.

[5] The f-score is calculated as $\frac{2 \times precision \times recall}{precision + recall}$.

Model	PARSEVAL		Relations	
	Lab.	Unlab.	Lab.	Unlab.
Baseline	14.7	33.8	7.4	53.3
Model 1	22.7	42.2	23.1	47.0
Model 2	30.1	51.1	31.0	54.3
Model 3	39.4	62.8	39.4	64.4
Model 4	46.3	69.2	43.2	67.9
Inter-annotator	53.7	76.5	50.3	73.0
Annotator-gold	75.9	88.0	75.3	84.0

Figure 5: Model performance.

logue context. The baseline is informed by the gold-standard utterance moods. For this corpus, this results in a baseline which is a right-branching structure, where the relation *Plan-Elaboration* is used if the utterance is indicative, *Question-Elaboration* if it is interrogative, and *Request-Elaboration* if it is imperative. The baseline also appropriately handles ignorable utterances (i.e, those with the mood labels irrelevant, pause, or pleasantry).

The baseline performs poorly on labelled relations (7.4%), but is more competitive on unlabelled ones (53.3%). The main reason for this is that it takes no segmentation risks. It simply relates every non-ignorable utterance to the previous one, which is indeed a typical configuration with common content-level relations like *Continuation*. The generative models take risks that allow them to correctly identify more complex segments – at the cost of missing some of these easier cases.

Considering instead the PARSEVAL scores for the baseline, the labelled performance is much higher (14.7%) and the unlabelled is much lower (33.8%) than for relations. The difference in labelled performance is due to the fact that the intentional-level relations used in the baseline often have arguments that are multi-utterance segments in the gold standard. These are penalized in the relations comparison, but the spans used in PARSEVAL are blind to them. On the other hand, the unlabelled score drops considerably – this is due to poor performance on dialogues whose gold standard analyses do not have a primarily right-branching structure.

Model 1 performs most poorly of all the models. It is significantly better than the baseline on labelled relations, but significantly worse on unlabelled relations. All its features are derived from the structure of the trees, so it gets no clues from speaker turns or the semantic content of utterances.

Model 2 brings turns and larger context via the ST and HCR features, respectively. This improves segmentation over Model 1 considerably, so that the model matches the baseline on unlabelled relations and beats it significantly on labelled relations.

The inclusion of the TC feature in Model 3 brings large (and significant) improvements over Model 2. Essentially, this feature has the effect of penalizing hypothesized content-level segments that span several turns. This leads to better overall segmentation.

Finally, Model 4 incorporates the domain-based TM feature that summarizes some of the semantic content of utterances. This extra information improves the determination of labelled relations. For example, it is especially useful in distinguishing a *Plan-Correction* from a *Plan-Elaboration*.

The overall trend of differences between PARSEVAL and relations scoring show that PARSEVAL is tougher on overall segmentation and relations scoring is tougher on whether a model got the right arguments for each labelled relation. It is the latter that ultimately matters for the discourse structures produced by the parser to be useful; nonetheless, the PARSEVAL scores do show that each model progressively improves on capturing the trees themselves, and that even Model 1 – as a syntactic model – is far superior to the baseline for capturing the overall form of the trees.

We also compare our best model against two upperbounds: (1) inter-annotator agreement on ten dialogues that were annotated independently and (2) the best annotator against the gold standard agreed upon after the independent annotation phase. For the first, the labelled/unlabelled relations f-scores are 50.3%/73.0% and for the latter, they are 75.3%/84.0%—this is similar to the performance on other discourse annotation projects, e.g., Carlson et al. (2001). On the same ten dialogues, Model 4 achieves 42.3%/64.9%.

It is hard to compare these models with Marcu's (1999) rhetorical parsing model. Unlike Marcu, we did not use a variety of corpora, have a smaller training corpus, are analysing dialogues as opposed to monologues, have a larger class of rhetorical relations, and obtain the elementary discourse units

from the Redwoods annotations rather than estimating them. Even so, it is interesting that the scores reported in Marcu (1999) for labelled and unlabelled relations are similar to our scores for Model 4.

7 Conclusion

In this paper, we have shown how the complex task of creating structures for SDRT can be adapted to a standard probabilistic parsing task. This is achieved via a headed tree representation from which SDRSs can be recovered. This enables us to directly apply well-known probabilistic parsing algorithms and use features inspired by them. Our results show that using dialogue-based features are a major factor in improving the performance of the models, both in terms of determining segmentation appropriately and choosing the right relations to connect them.

There is clearly a great deal of room for improvement, even with our best model. Even so, that model performed sufficiently well for use in semi-automated annotation: when correcting the model's output on ten dialogues, one annotator took 30 seconds per utterance, compared to 39 for another annotator working on the same dialogues with no aid.

In future work, we intend to exploit an existing implementation of SDRT's semantics (Schlangen and Lascarides, 2002), which adopts theorem proving to infer resolutions of temporal anaphora and communicative goals from SDRSs for scheduling dialogues. This additional semantic content can in turn be added (semi-automatically) to a training corpus. This will provide further features for learning discourse structure and opportunities for learning anaphora and goal information directly.

Acknowledgments

This work was supported by Edinburgh-Stanford Link R36763, ROSIE project. Thanks to Mirella Lapata and Miles Osborne for comments.

References

N. Asher and A. Lascarides. *Logics of Conversation*. Cambridge University Press, 2003.

J. Baldridge and A. Lascarides. Annotating discourse structures for robust semantic interpretation. In *Proceedings of the 6th International Workshop on Computational Semantics*, Tilburg, The Netherlands, 2005.

L. Carlson, D. Marcu, and M. Okurowski. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Proceedings of the 2nd SIGDIAL Workshop on Discourse and Dialogue, Eurospeech*, 2001.

M. Collins. Head-driven statistical models for natural language parsing. *Computational Linguistics*, 29(4):589–638, 2003.

J. R. Hobbs, M. Stickel, D. Appelt, and P. Martin. Interpretation as abduction. *Artificial Intelligence*, 63(1–2):69–142, 1993.

W. C. Mann and S. A. Thompson. Rhetorical structure theory: Description and construction of text structures. In G. Kempen, editor, *Natural Language Generation: New Results in Artificial Intelligence*, pages 279–300. 1986.

D. Marcu. The rhetorical parsing of unrestricted natural language texts. In *Proceedings of ACL/EACL*, pages 96–103, Somerset, New Jersey, 1997.

D. Marcu. A decision-based approach to rhetorical parsing. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL99)*, pages 365–372, Maryland, 1999.

E. Miltsakaki, R. Prasad, A. Joshi, and B. Webber. The Penn Discourse TreeBank. In *Proceedings of the Language Resources and Evaluation Conference*, Lisbon, Portugal, 2004.

J. D. Moore and M. E. Pollack. A problem for RST: The need for multi-level discourse analysis. *Computational Linguistics*, 18(4):537–544, 1992.

S. Oates. Generating multiple discourse markers in text. Master's thesis, ITRI, University of Brighton, 2001.

S. Oepen, E. Callahan, C. Manning, and K. Toutanova. LinGO Redwoods—a rich and dynamic treebank for HPSG. In *Proceedings of the* LREC *parsing workshop: Beyond* PARSEVAL*, towards improved evaluation measures for parsing systems*, pages 17–22, Las Palmas, 2002.

D. Schlangen and A. Lascarides. Resolving fragments using discourse information. In *Proceedings of the 6th International Workshop on the Semantics and Pragmatics of Dialogue (Edilog)*, Edinburgh, 2002.

R. Soricut and D. Marcu. Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of Human Language Technology and North American Association for Computational Linguistics*, Edmonton, Canada, 2003.

A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, D. Jurafsky R. Bates, P. Taylor, R. Martin, C. van Ess-Dykema, and M. Meteer. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–374, 2000.

M. Strübe and C. Müller. A machine learning approach to pronoun resolution in spoken dialogue. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL2003)*, pages 168–175, 2003.

J. van Eijk and H. Kamp. Representing discourse in context. In J. van Benthem and A. ter Meulen, editors, *Handbook of Logic and Linguistics*, pages 179–237. Elsevier, 1997.

R. Vieira and M. Poesio. Processing definite descriptions in corpora. In *Corpus-based and computational approaches to anaphora*. UCL Press, 2000.

J. M. Wiebe, T. P. O'Hara, T. Ohrstrom-Sandgren, and K. J. McKeever. An empirical approach to temporal reference resolution. *Journal of Artificial Intelligence Research*, 9:247–293, 1998.