Robust Pragmatics for Narrative Text

Case for Support
Alex Lascarides Maria Lapata

Part 1
Previous Research Track Record

The proposed project is concerned with the design, implementation and evaluation of a system that analyses the discourse structure of narrative text and the temporal order of its events.

Personnel

Alex Lascarides. Lascarides has over 12 years' post-doctoral experience in computational linguistics. She has attained international recognition for work in discourse processing and lexical processing. The proposed project will exploit a linguistically-principled theory of discourse structure for resolving the temporal structure of narrative texts. Lascarides, in collaboration with Nicholas Asher (Texas), has devised such a formal theory of discourse semantics known as Segmented Discourse Representation Theory (SDRT), and an accompanying formal model of pragmatics known is Discourse in Commonsense Entailment (DICE), [1, 2, 3, 4, 5, 6, 7, 18, 19, 21, 24, 25, 28]. Some of this work was developed as part of the STAGE project (SERC GR/G/22077). SDRT connects propositions, questions and requests in the discourse with rhetorical relations such as *Contrast* and *Explanation*, and it records the semantic effects that these relations have on content. DICE is a logical theory of how linguistic and pragmatic knowledge determines which rhetorical relation holds. It exploits state of the art techniques in commonsense reasoning to do this.

One distinctive feature of SDRT and DICE is that they formalise mechanisms by which compositional and lexical semantics get augmented, in a constrained manner, with further contextually determined content. The latter includes temporal relations among events, which is one of the focus areas of this project [19, 2]. In addition, SDRT and DICE have been used to resolve VP ellipsis [23], lexical sense disambiguation [2, 8, 22], bridging inferences and presuppositions [5, 6], and conversational implicatures in dialogue [4, 20]. SDRT and DICE have also been used to formalise an interface between commonsense reasoning with pragmatics and statistical preferences for word senses that are extracted from the BNC [8].

Maria Lapata. Lapata's PhD work focussed on the acquisition of meaning from corpora and the modelling of lexical semantic ambiguity in a probabilistic framework. Her thesis investigated the induction of default meaning—the dominant meaning of polysemous words in the absence of explicit contextual information—using linguistic insights to guide the acquisition process. She proposed a probabilistic model of default meaning in a Bayesian framework in which distributional information (in the form of conditional probabilities) is combined with linguistic generalisations (in the form of prior probabilities). This approach was successfully applied to polysemous verbs [13, 11], adjectives [12], and compound nouns [14].

Lapata has also applied corpus-based and and probabilistic learning techniques to computational psycholinguistics. She employed partial and statistical parsing techniques to model subcategorization preferences [15] and explored the link between semantic plausibility and lexical association [16]. This resulted in a broad coverage account of the plausibility of adjective-noun combinations, based on co-occurrence statistics and corpus-based learning techniques. The coverage of the model has been recently extended to adjective-noun combinations that do not occur in the corpus [17] using class-based and similarity-based smoothing techniques.

Lapata has also conducted research on discourse and dialogue modelling aiming at the creation of probabilistic models for the automatic identification of dialogue acts [9].

The Host Organisation

The proposed project will be carried out in collaboration between the Language Technology Group (LTG) and the Institute for Communicating and Collaborative Systems (ICCS) within the Division of Informatics at the University of Edinburgh. The work will build on earlier LTG work on the analysis of the temporal structure of discourse [10]. The LTG has also developed a series of corpus handling tools for use in data-intensive linguistic and data engineering tasks (see [26] for an overview), and tools for annotation tasks (the MATE workbench, [27]); these tools are sufficiently flexible that we can apply them here.

The proposed research requires expertise not only in corpus-based processing and language engineering but also in theoretical and computational linguistics and in probabilistic modelling. Members of ICCS do both basic and applied research in these areas, typically fusing together a mixture of methodologies from the disciplines of Cognitive Science, Artificial Intelligence, Computer Science, Linguistics, Mathematics, and Psychology. This interdisciplinary research environment is essential to the project's success.

References

[1] Asher, N. and A. Lascarides. 1994. Intentions and Information in Discourse. In *Proceedings of the 32nd Annual Meeting of the Association of Computational Linguistics*, 34–41. Las Cruces, NM.

[2] Asher, N. and A. Lascarides. 1995a. Lexical Disambiguation in a Discourse Context. *Journal of Semantics*, 12(1): 69–108.

[3] Asher, N. and A. Lascarides. 1995b. Metaphor in Discourse. In *Proceedings of the* AAAI *Spring Symposium Series: Representation and Acquisition of Lexical Knowledge*, 3–7. Stanford, CA.

[4] Asher, N. and A. Lascarides. 1998b. Questions in Dialogue. *Linguistics and Philosophy* 21(2): 237–309.

[5] Asher, N. and A. Lascarides. 1998b. Bridging. *Journal of Semantics* 15(1): 83–113.

[6] Asher, N. and A. Lascarides. 1998c. The Semantics and Pragmatics of Presupposition. *Journal of Semantics*, 15: 239–299.

[7] Asher, N. and A. Lascarides. in press. The Semantics and Pragmatics of Metaphor. Pustejovsky, J. and F. Busa (eds.) *Semantic Parameters and Lexical Universals*. Cambridge, MA: The MIT Press.

[8] Copestake, A. and A. Lascarides. 1997. Integrating Symbolic and Statistical Representations: The Lexicon Pragmatics Interface. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, 136–143. Madrid, Spain.

[9] Finke, M., M. Lapata, *et al*. 1998. Clarity: Inferring Discourse Structure from Speech. In *Proceedings of the AAAI Spring Symposium on Applying Machine Learning to Discourse Processing*. Stanford, CA.

[10] Hitzeman, J., M. Moens, and C. Grover. 1995. Algorithms for Analysing the Temporal Structure of Discourse. In *Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics*, 253–260. Dublin, Ireland.

[11] Lapata, M. and C. Brew. 1999. Using Subcategorization to Resolve Verb Class Ambiguity. In *Proceedings of Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 266–274. College Park, MD.

[12] Lapata, M. 2001. A Corpus-based Account of Regular Polysemy: The Case of Context-sensitive Adjectives. In *Proceedings of the 2nd North American Annual Meeting of the Association for Computational Linguistics*. Pittsburgh, PA.

[13] Lapata, M. 1999. Acquiring Lexical Generalizations from Corpora: A Case Study for Diathesis Alternations. In *Proceedings of the 37th American Annual Meeting of the Association for Computational Linguistics*, 397–404. College Park, MD.

[14] Lapata, M. 2000. The Automatic Interpretation of Nominalizations. In *Proceedings of the Proceedings of the 17th National Conference on Artificial Intelligence*, 716–721. Austin, TX.

[15] Lapata, M., F. Keller, and S. Schulte im Walde. 2001. Verb Frame Frequency as a Predictor of Verb Bias. To appear in *Journal of Psycholinguistic Research*.

[16] Lapata, M. S. McDonald, and F. Keller. 2001. Determinants of Adjective-Noun Plausibility. In *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics*, 30–36. Bergen Norway.

[17] Lapata, M., F. Keller, and S. McDonald. 2001. Validating Smoothing Techniques: The Case of Adjective-Noun Plausibility. Submitted.

[18] Lascarides, A. and Asher, N. 1991. Discourse Relations and Defeasible Knowledge. In *Proceedings to the 29th Annual Meeting of the Association of Computational Linguistics*, 55–63. Berkeley, CA.

[19] Lascarides, A. and N. Asher. 1993. Temporal Interpretation, Discourse Relations and Commonsense Entailment. *Linguistics and Philosophy*, 16(5): 437–493.

[20] Lascarides, A. and N. Asher. 1999. Cognitive States, Discourse Structure and the Content of Dialogue. In *Proceedings from Amstelogue 1999*. 1–12. Amsterdam, Holland.

[21] Lascarides, A., N. Asher, and J. Oberlander. 1992. Inferring Discourse Relations in Context. In *Proceedings of the 30th Annual Meeting of the Association of Computational Linguistics*, 1–8. Newark, DE.

[22] Lascarides, A. and A. Copestake. 1998. Pragmatics and Word Meaning. *Journal of Linguistics*, 34(2): 387–414.

[23] Lascarides, A., A. Copestake, and E. J. Briscoe. 1996. Ambiguity and Coherence. *Journal of Semantics*, 13(1): 41–65.

[24] Lascarides, A. and J. Oberlander. 1992. Abducing Temporal Discourse. In Dale, R., Hovy, E., Rösner, D. and Stock, O. (eds.) *Aspects of Automated Natural Language Generation*, 167–182. Berlin: Springer Verlag.

[25] Lascarides, A. and J. Oberlander. 1993. Temporal Coherence and Defeasible Knowledge. *Theoretical Linguistics*, 19(1): 1–35.

[26] McKelvie, D., C. Brew, and H. S. Thomson. 1998. Using SGML as a Basis for Data-Intensive NLP. *Computers and the Humanities* 5(31): 367–388.

[27] McKelvie, D., A. Isard *et al*. 2000. The MATE Workbench – an Annotation Tool for XML Coded Speech Corpora. *Speech Communication* 33(1–2): 97–112.

[28] Oberlander, J. and A. Lascarides. 2000. Laconic Discourses and Total Eclipses: Abduction in DICE. Harry B. and W, Black (eds.), *Abduction, Beliefs and Context: Studies in Computational Pragmatics*, 391–412. London: John Benjamins.

Part 2
The Proposed Research and its Context

1 Background

An essential part of text interpretation involves determining how events relate to one another. This temporal information can be useful for a variety of natural language processing (NLP) tasks, such as generation [25, 28], speech act modelling [55, 51], appointment scheduling [58] and information extraction [39]. In particular, consider an information extraction system whose goal is to gather temporal information about the events. In (1), taken from the British National Corpus (BNC), we would like such a system to extract information about the succeeding event, the resigning event and the reasons that led to the resigning:

(1) a. Lansbury resigned from the parliamentary leadership.
  b. His support for pacifism received only 102,000 votes.
  c. and he was personally attacked by Ernest Bevin.
  d. He was replaced by Clement Attlee, the party's first public-school educated middle-class leader.

The system should infer not only that *Clement Attlee* succeeded *Lansbury* but also that (1b) and (1c) led to Lansbury's resignation. So the system should recognise that (1d) temporally followed (1a) and that the descriptive order between (1a) and (1b,c) mismatches their temporal order (i.e., the vote and the attack by Ernest Bevin happened before Lansbury resigned). Although the interactions between discourse relations (e.g., *Explanation*, *Result*) and event relations have received much attention in linguistics [42, 35, 57, 19], the automatic interpretation of events and their relations is beyond the capabilities of current NLP systems, especially when such relations are not linguistically marked by discourse cue phrases such as *because* or *after*.

To date, work has focussed primarily on the recognition of rhetorical relations, without explicitly taking event relations into account. Some work has concentrated on resolving the antecedents of temporal expressions in the domain of scheduling dialogues [58] and modelling event coreference for information extraction [39]. Hitzeman et al describe a method for analysing the temporal structure of discourse using a unification-based grammar [10]. The approach is promising in that it analyses temporal structure algorithmically via rhetorical relations, discourse cues, tense, and aspect; however, it suffers from overgeneration of rhetorical relations. Ambiguity is reduced by manually specifying constraints and preferences (i.e., temporal centering) on discourse relations. But it is difficult to assess how well the method performs given that no qualitative evaluation is reported. It also relies on relatively deep parsing and interpretation, and wide-coverage grammars that produce the necessary precision are generally not robust enough for processing unrestricted text (e.g., LinGO, `http://lingo.stanford.edu`).

Recognising rhetorical relations is a hard task given that they are not always made explicit via discourse connectives. For example, although the rhetorical relations between (2a) and (2b), and between (3a) and (3b) (also from the BNC) are indicated via the phrases *as a result of* and *because*, competent language users can infer rhetorical relations when such cues are absent (see (1)), by utilising knowledge about the world, the structure of the lexicon, and the cognitive state of the speaker.

(2) a. Mosley resigned from the Labour government in 1930
  b. as a result of the Cabinet turning down his reflationary programme contained in the Mosley Memorandum

(3) a. Mrs Yann Piat, said she resigned from the party a year ago
  b. because she realised that the Front was a dead-end movement and a danger to our country.

Most approaches to the automatic identification of rhetorical relations have either relied on hand-crafted rules which specify, for example, world knowledge; or they are restricted to cases where cue phrases are present (e.g., (2) and (3)). Neither approach on their own achieves the precision or robustness required of the next generation of NLP applications.

For example, Hobbs et al use weighted abduction working over linguistic and pragmatic information sources to determine the most plausible rhetorical relation holding between utterances [38]. However, the large number of manually encoded axioms that is required, and the lack of a robust mechanism for constraining inferences over them, make this largely impractical. Cue-based approaches [44, 41, 49] rely on cues provided by discourse connectives, lexical repetition, and punctuation to identify rhetorical relations. Although attractive given their simplicity, such approaches fail to identify a rhetorical relation in the absence of this information. Marcu [44], for example, provides a method for deriving the rhetorical structure of text by identifying cue phrases, breaking sentences into clauses and computing rhetorical structure trees in the style of Rhetorical Structure Theory (RST, [42]). Although this is the first serious attempt to automatically recognise rhetorical relations in real text, it suffers from combinatorial explosion, as the number of RST trees increases exponentially with the number of hypothesised rhetorical relations (even though the constraints on their well-formedeness are in fact too strong; see [48]). One can attempt to remedy this by constraining the space of possible relations; cf. the scoring mechanism in [33]. However, the scores which eliminate candidate relations are determined heuristically, and no method is proposed for their acquisition from real text.

On the one hand, cue-based approaches are simple but not general enough to account for rhetorical relations between text spans where discourse connectives are absent. On the other hand, knowledge-intensive approaches are general enough, but are not sufficiently robust or tractable. Ideally, one would like to take advantage of pragmatic as well as syntactic and lexico-semantic information, while maintaining sufficient precision and robustness.

The aims of this project are twofold. First, we will investigate what types of features are important for the automatic recognition of rhetorical relations and integrate them in a probabilistic model that is able to infer these relations, even in cases where discourse connectives are absent. Second, we will develop algorithms that infer the temporal structure of discourse, borrowing insights from SDRT, [2, 4, 5]), a theory of discourse coherence that is sufficiently formalised to be incorporated in a computationally specifiable interpretation procedure. Furthermore, SDRT claims that weaker assumptions than those made by RST about the structure of texts are sufficient for performing semantic tasks that utilise discourse structure (cf. [48]). Having a simpler discourse structure and a smaller number of rhetorical relations is desirable for a probabilistic framework, as it results in a smaller parameter space. SDRT also emphasises the importance of the lexicon and linguistic cues in recognising rhetorical relations, and therefore depends less on world knowledge and cognitive states (in comparison to [38], for example). Finding the linguistic cues which are the best predictors of rhetorical relations will be one of the main tasks of the proposed project.

2 Programme and Methodology

2.1 Overview

In this project we will develop a hybrid system that combines symbolic constraints on the well-formedeness of discourse structures with a probabilistic model of text coherence and is able to infer the temporal structure of narrative text, even in cases where discourse cues are absent. We will model temporal progression using a probabilistic formulation, giving us a principled approach for combining multiple sources of linguistic knowledge (i.e., features), as well as the ability to derive model parameters automatically from a corpus using data-intensive methods. These methods have been successfully used for a variety of tasks such as lexicon acquisition, speech act recognition, and discourse segmentation (see [43] for an overview). The symbolic component will specify the semantic effects of rhetorical relations on the text, particularly its temporal content; it will also include established constraints on the well-formedeness of rhetorical structure (e.g., temporal centering, right-frontier), which will help to delimit the probabilistic parameter space. Our research programme involves the following tasks:

Task 1. Specification of probabilistic and symbolic components for inferring rhetorical relations and the temporal structure of discourse. We outline the probabilistic model and symbolic reasoning we have in mind in Section 2.2.

Task 2. Creation of a corpus of rhetorical relations. This training corpus will be used to estimate the parameters of our probabilistic model. Instead of creating a corpus which is hand-labelled with rhetorical relations, we intend to exploit information in clauses whose rhetorical relation is made apparent by an explicit discourse connective. A corpus of sentences containing discourse connectives will be created and used for the estimation of various model parameters (see Section 2.3).

Task 3. Exploration of the types of linguistic knowledge (i.e., features) that are useful for inferring rhetorical relations and the temporal order of events (see Section 2.3).

Task 4. Building a system that can identify in an unseen text the temporal order of its events. This system will put forward a hybrid approach by combining probabilistic modelling and symbolic constraints on the well formedeness of discourse structure (see Section 2.2).

Task 5. Evaluation of the system's results. We will evaluate the accuracy of the system against data that is manually annotated with rhetorical and temporal relations (see Section 2.4).

These tasks are interleaved and will therefore be carried out in parallel throughout the project. It is not possible to specify the probabilistic model without specifying its features and an inventory of rhetorical relations, for example.

2.2 Modelling Temporal Order

Modelling involves three interdependent tasks: (a) inferring rhetorical relations (Task 1), (b) inferring the temporal structure of the discourse (Tasks 1 and 4), and (c) determining which linguistic features are important for probabilistic classification (Task 3). To give a concrete example assume the discourse (1a)–(1c). In order to work out the temporal order between (1a) and (1b) we need to infer their rhetorical relationship. Assuming a small set of rhetorical relations (e.g., *Narration*, *Elaboration*, *Explanation*, *Background*, *Result*, see [42] and [40] for classifications of discourse relations) our task is to predict the right rhetorical relation between (1a) and (1b) using a probabilistic model which enables us to combine multiple knowledge sources.

Formally, $P(S_a, r_j, S_b)$ represents the probability that a rhetorical relation r_j holds between structures S_a and S_b (these can be sentences or clauses, see (1a)–(1c)). By varying r_j we are trying to find which rhetorical relation maximises $P(S_a, r_j, S_b)$ (see (4)).

$$
\begin{aligned}
(4)\quad *r &= \operatorname*{argmax}_{r_j \in R} P(S_a, r_j, S_b) \\
&= \operatorname*{argmax}_{r_j \in R} P(S_a)P(r_j|S_a)P(S_b|S_a, r_j) \\
&\approx \operatorname*{argmax}_{r_j \in R} P(r_j|S_a)P(S_b|r_j) \\
&\approx \operatorname*{argmax}_{r_j \in R} P(r_j)P(S_a|r_j)P(S_b|r_j)
\end{aligned}
$$

Here, we ignore the term $P(S_a)$ as it is a constant and use Bayes' Rule to derive $P(S_a|r_j)$ from $P(r_j|S_a)$. S_a and S_b are vectors of features $s_{a_1}, s_{a_2}, \cdots s_{a_n}$ and $s_{b_1}, s_{b_2}, \cdots s_{b_n}$ characteristic of the relation r_j (e.g., tense, punctuation, polarity). By making the simplifying assumption that these features are conditionally independent given the rhetorical relation, the probability of observing the conjunctions $s_{a_1}, s_{a_2}, \cdots s_{a_n}$ and $s_{b_1}, s_{b_2}, \cdots s_{b_n}$ is the following:

(5) $$r^* = \underset{r_j \in R}{\operatorname{argmax}} P(r_j) \prod_i P(s_{a_i}|r_j) \prod_i P(s_{b_i}|r_j)$$

A Bayesian approach (see (5)) is particularly attractive since it combines knowledge about the prior probabilities of alternative relations (for instance we might expect *Elaborations* to be more likely than *Explanations*) with knowledge of the probabilities of observing various data given a particular rhetorical relation. Furthermore, note that treating discourse relations as a classification task allows us to explore the contribution of classification techniques other than naive Bayes (e.g., decision tree learning, maximum entropy). The combination of different classifiers has been shown to increase system performance (at least for NP-chunking and part-of-speech tagging, see [56] for an overview) and we intend to explore its contribution to the identification of rhetorical relations.

Note that (5) is a model of *local coherence*: it infers the most likely relation between S_a and S_b only, *without* having a notion of *global coherence*, i.e., of how a group of clauses or sentences relate to each other. We will create an incremental model of global coherence (i.e., discourse structure) by combining the probabilistic model in (5) with a symbolic component that incorporates well-formedeness constraints (e.g., right frontier) on discourse structure. Discourse structure is typically represented hierarchically, i.e., by *trees* [35, 57, 18], because some rhetorical relations are classified as subordinating (e.g., *Explanation*, *Elaboration*), and others as coordinating (e.g., *Narration*, *Background*) [29].

Consider the tree in Figure 1, which represents the discourse structure of (1a)–(1c). By the symbolic model of well-formed discourse structures given in, for example, [29], only two attachment sites are available for connecting the next clause (1d): (1a) and (1c). (1b) isn't available, because it's not on the right frontier, and so the probability of (1d) attaching to it is 0. But we must choose, probabilistically, which of the two attachment sites and which rhetorical relation to use. We make these decisions by choosing the site which yields the rhetorical relation $r_j \in R$ with the highest probability; i.e., we choose (1a) as the attachment site of (1d) if there's a relation $r_j \in R$ which maximises $P(S_a, r_j, S_d)$, and $P(S_a, r_j, S_d)$ is larger than $P(S_c, r_i, S_d)$ for *any* $r_i \in R$. Generalising from this illustrative example, we intend to not only develop novel algorithms which combine probabilistic and symbolic reasoning, but also experimentally evaluate their individual contribution using a broadly heuristic (and therefore computationally inexpensive) approach.

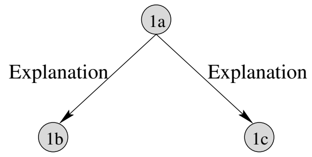

Figure 1: Discourse structure for (1a)–(1c)

Once we have established the rhetorical relations between text spans we can determine the temporal ordering between events by assuming a (symbolic) mapping between rhetorical relations and temporal relations. For example, if *Background* relates two events, these events overlap; if the rhetorical relation is *Result*, then one event precedes the other (see [10] for details).

2.3 Parameter Estimation

An important aspect of our proposal is to determine what type of information is needed for the probabilistic classification of rhetorical relations (see Tasks 3 and 2). So far, cue-based approaches to discourse relations have concentrated on a limited number of surface features such as punctuation and lexical repetition. In addition to those two features, we will systematically explore the contribution of a wide range of features including: verb and noun semantics, tense, aspect, modality, polarity, and temporal expressions (e.g., *tomorrow*, *the previous day*). SDRT claims that verb semantics is particularly important for recognising discourse relations [19], and a manual inspection of texts from the BNC supports this claim, suggesting that (1) is typical (e.g., the relationships between resigning and being attacked, and resigning and being replaced). Indeed, an inspection of sentences from the BNC which feature *because* suggests that the verb semantics supports the discourse relation even if the cue phrase were absent, and this will be important for our proposed method of estimating parameters, which we'll describe shortly. Most of the above linguistic features can be easily retrieved from the corpus. Verb and noun relations can be retrieved from WordNet [47], although methods have been proposed for the acquisition of their relations from corpora (see [43] for an overview).

A standard approach to estimating these parameters and consequently training the assumed probabilistic model is to create a corpus that is hand-labelled with rhetorical relations. Although this approach has been successfully applied to the recognition of speech-acts [55], it is highly labour-intensive. Instead of relying on manual annotation, we will estimate our model parameters from a corpus of text segments whose rhetorical relations are indicated by discourse connectives. Knott proposed a set of discourse relations signalled by discourse cues present in real text [40]. Drawing on this body of work, we will assume a mapping between discourse connectives and rhetorical relations (e.g., *because* indicates *Explanation* and *as a result of* indicates

Result, see (2) and (3)). Some discourse connectives may be ambiguous (e.g., *and*, *also*). However, automatic methods exist that have been successfully used for their disambiguation [37, 53].

More specifically, we will estimate rhetorical relations from the British National Corpus (BNC, [30]), a large, synchronic corpus of British English, consisting of 90 million words of text and 10 million words of speech (we will ignore the speech and dialogue data, however, both in training and testing). The BNC is a balanced corpus, i.e., it was compiled so as to represent a wide range of present day British English. Its texts are classified with information about their source (e.g., *fiction*, *educational*), and we can exploit this information to isolate narrative texts for training (e.g., we train on ones marked *fiction*, but ignore those marked *educational*). The BNC is representative across domains, and this ensures that training will take place on a representative sample of narrative texts. Finally, note that once an initial model is trained, bootstrapping [36] can be used to create larger amounts of training data which is not limited to corpus tokens containing discourse connectives.

We will use a part-of speech tagged, lemmatised, and parsed version of the BNC. Recognising the syntactic role of constituents relatively accurately is important for the correct estimation of the model parameters. For example, distinguishing between main and subordinate clauses is important for deciding discourse structure. We will use a parsed version of the BNC which was obtained using a robust statistical parser [31]. The parser utilises a probabilistic context-free grammar (PCFG) for English. In a PCFG, context-free rules are annotated with probabilities, and the probability of a parse is computed as the product of the probabilities of the relevant rules. The grammar proposed by [31] is an extension of the standard PCFG model in that it incorporates information about the lexical heads of constituents. Such a head-lexicalised PCFG provides a grammar model that combines lexicalised rules and lexical coherence relations between constituents.

In our previous work we have shown that the parses obtained using a head-lexicalised PCFG can be used to reliably estimate subcategorization frequencies [15]. These frequencies yield a high correlation with manually annotated data. Obtaining subcategorization frames reliably is important for the proposed work. Recent work on lexicon acquisition [50, 46, 11] uses subcategorization information to derive the meanings of verbs or verb classes which in turn can be used as indicators of the type of rhetorical relation between text spans. The parsed version of the corpus will not only be used for the acquisition of the relations between verbs and nouns, but also for the identification of discourse cues, clauses, and the acquisition of morphological and syntactic information indicating aspect and aktionsart (see [52] for an approach that acquires aspectual classes from corpora).

In general, our approach avoids the pitfalls of Marcu [44] and Corston-Oliver [33] discussed earlier: (a) we avoid the combinatorial explosion of rhetorical relations by using statistical inference techniques; (b) our approach can be applied to text segments even if discourse connectives are absent; and (c) we propose to derive the model parameters automatically from a corpus using data-intensive techniques, rather than from a manually annotated training corpus.

2.4 Evaluation

We will assess accuracy of the hybrid model for recognising discourse structure (see Task 5). We will also evaluate the system's accuracy at guessing temporal ordering and establish a baseline for the task.

Evaluation will require corpus data annotated with discourse relations. The evaluation data will consist of narrative texts with different domains, styles and registers. This is important for assessing whether the approach generalises across corpora. Texts from the Brown corpus [34], the Wall Street Journal [45] and the monologues in [32] are an example of an evaluation set. Rhetorical relations and discourse structure will be coded by more than one annotator, and inter-annotator agreement will be computed to assess reliability.

Evaluation will focus on three interdependent tasks: (a) evaluation of the discourse relations automatically assigned, (b) evaluation of derived discourse structure, and (c) evaluation of temporal ordering. Given that discourse structure is in fact a tree structure (at least, in the simplified model assumed here), we can make use of well-established measures that have been proposed for evaluating parse trees. For example, we will measure precision, i.e., number of discourse segments for which correct attachments and discourse relations were derived, by comparing the system's output to the manual data (tasks (a) and (b)). Similarly, precision will be used to evaluate the accuracy of the temporal relations (see task (c)). We will also adopt a measure similar to what is known as cross-brackets in the parsing literature. Cross-brackets are the number of constituents for which the manual gold-standard has a bracketing such as ((A B) C) but the candidate parse has a bracketing such as (A (B C)). As far as the proposed research is concerned A, B, and C are not syntactic constituents but discourse segments (see task (b)).

We will further devise evaluation scores which will measure overall correctness by taking into account both discourse structure and the labelling of discourse relations (see tasks (a)–(c)). For example, a tree for which the system guessed the right structure but no correct relation should be considered worse than a tree for which the system guessed most rhetorical relations and some attachment sites right.

2.5 Research Plan

Corpora. Create training and testing corpora. Establish the inventory of rhetorical and temporal relations required. Measure inter-annotator agreement at coding discourse structure. Establish baseline for the task. *Run-*

time: months 10.

Modelling. Specify the probabilistic model and its parameters. Explore the parameter space. Create an inventory of features. *Runtime:* months 18.

Feature Exploration. Experiment with features and their interactions. Establish most predictive features for each rhetorical and temporal relation. *Runtime:* months 8.

Discourse Structure. Interface the probabilistic with the symbolic component. Create a system that infers the temporal ordering of narrative texts. *Runtime:* months 18.

Evaluation. Evaluate the research results by testing the system's performance against manually annotated data. Develop evaluation scores. Compare the system's performance on different text domains. *Runtime:* months 12.

3 Appointment of Effort

Collaborations on SDRT will continue with Nicholas Asher (Texas). This project will be done within LTG and ICCS, Division of Informatics, University of Edinburgh. As explained earlier, this environment provides a wide variety of expertise in statistical parsing (e.g., Osborne, Mikheev), semantics (e.g., Moore, Steedman, Webber) and corpus annotation (e.g., McKelvie).

Lascarides will devote 10% of her time to the project. She will act as primary supervisor to the PhD student, and will develop the symbolic model of discourse structure and its semantic effects. The named RA (Lapata) will be responsible for specifying the probabilistic model, and for interfacing the probabilistic and symbolic components. Many of the software packages for processing the BNC that she developed for her PhD thesis will be re-used on this project. The students' work will focus on the data-intensive portion of the project, including exploring the parameter space, establishing the most predictive features, and evaluating the results (PhD training and writing up the thesis is additional to the above research plan). Finally, three annotators are required for assessing task difficulty, and for creating a gold standard against which the system will be evaluated.

4 Relevance to Beneficiaries

The practicality and range of many NLP applications would be significantly enhanced if one could automatically compute the implicit content of unrestricted text. This project addresses this, by computing discourse relations and their temporal effects in a reliable and robust manner. As a consequence, the potential applications are wide and varied. We mention just two here.

First, researchers have long hypothesised that discourse structure can aid summarisation [42, 54]. The output of our system can be used as the input to a summarisation system equipped with a discourse module. Furthermore, our system's output can be used as a discourse tagging tool which guesses the temporal relations and attachment sites of discourse segments. This tool can be used for the semi-automatic creation of corpora annotated with discourse information.

Second, the results of this project would be beneficial to information extraction. Knowing the temporal order of events is important for identifying whether these events are coreferent or not. Knowledge of the interrelations among events can also produce more accurate fact extraction.

5 Research Dissemination

The results of the research will be disseminated in journal and conference papers (e.g., ACL) that undergo peer review at an international level. A version of the documented software will be made available (over the web) to other users for research purposes.

6 Justification of Resources

Staff. A salary is requested to employ Lapata on the AR1 scale. The proposed work requires a postdoctoral researcher with specialist skills in data-intensive methodologies and linguistics, especially semantics. Although apparently 'expensive' in RA terms, the experience she brings to the project will make her more cost-effective than a junior RA. Computing support is included to cover the skilled manpower required for the continuing provision of a smoothly-running distributed computing infrastructure, including printing, mail, backup and archiving facilities, and the routine management of project specific hardware and software. The administrative support will allow us to keep interested companies informed of progress; investigate new contacts; organise meetings and travel; book rooms and prepare, copy and distribute documents.

Travel. Joint research with Nicholas Asher will require one site visit a year, since collaborations cannot be carried out via email alone. Attendance at relevant conferences will enable the dissemination of results and allow the project to keep abreast of new developments. Attending conferences was of great benefit to the development of SDRT and DICE, and also to Lapata's data-intensive NLP work, both of which we will use as starting points for this project. It will be necessary to attend conferences in linguistics and computational linguistics. The following are essential: COLING 2002, 2004; EACL 2003; ACL 2002, 2003, 2004; ANLP 2003, EMNLP 2003.

Equipment and Computing Consumables. The proposed work will be data-intensive, and in our experience a dedicated workstation with large-capacity storage media, directly attached, is best suited for this type of work. Sun workstations provide the best price/performance ratio for this. Infrastructure access costs have been included in the budget. We need three annotators to manually mark text, so as to evaluate the difficulty of the task, and to attain a gold standard against which system performance will be evaluated.

Exceptional Items. There are, as outlined above, a number of tasks that together constitute the kind of project that forms a successful PhD. The most cost-effective way of achieving these sub-tasks is therefore via a PhD studentship, rather than a postdoc RA.

References

[29] Asher, N. 1993. *Reference to Abstract Objects in Discourse*. London: Kluwer Academic Publishers.

[30] Burnard, L. 1995. *Users Guide for the British National Corpus*. British National Corpus Consortium, Oxford University Computing Service.

[31] Carroll, G. and M. Rooth. 1998. Valence Induction with a Head-lexicalized PCFG. In Nancy I. and A. Voutilainen (eds.), *Proceedings of the 3rd Conference on Empirical Methods in Natural Language Processing*, 36–45. Granada, Spain.

[32] Chafe, W. L. 1980. *The Pear Stories: Cognitive, Cultural and Linguistic Aspects of Narrative Production*. Norwood, NJ: Ablex Publishing Corporation.

[33] Corston-Oliver, S. H. 1998. Beyond String Matching and Cue Phrases: Improving Efficiency and Coverage in Discourse Analysis. In *Proceedings of the COLING/ACL Workshop on Discourse Relations and Discourse Markers*, 9–15. Montréal, Canada.

[34] Francis, N. H. Kucera and A. Mackie. 1982. *Frequency Analysis of English Usage: Lexicon and Grammar*. Boston: Houghton Mifflin.

[35] Grosz, B. J. and C. Sidner. 1986. Attention, Intentions, and the Structure of Discourse. *Computational Linguistics* 12(3): 175–204.

[36] Hindle, D. and M. Rooth. 1993. Structural Ambiguity and Lexical Relations. *Computational Linguistics*, 19(1): 103–120.

[37] Hirschberg, J. and D. Litman. 1993. Empirical Studies on the Disambiguation of Cue Phrases. *Computational Linguistics* 19(3): 501–530.

[38] Hobbs, J. R., M. Stickel *et al*. 1993. Interpretation as Abduction. *Journal of Artificial Intelligence* 63(1–2): 69–142.

[39] Humphreys, K., R. Gaizauskas, and S. Azzam. 1997. Event Coreference for Information Extraction. In *Proceedings of the ACL Workshop on Operational Factors in Practical Robust Anaphora Resolution for Unrestricted Texts*, 75–81. Madrid, Spain.

[40] Knott, A. 1996. A Data-Driven Methodology for Motivating a Set of Coherence Relations. Ph.D. thesis, University of Edinburgh.

[41] Kurohashi, S. and M. Nagao. 1994. Automatic Detection of Discourse Structure by Checking Surface Information in Sentences. In *Proceedings of the 15th International Conference on Computational Linguistics*, 1123–1127. Kyoto, Japan.

[42] Mann, W. C. and S. A. Thomson. 1988. Rhetorical Structure Theory. *Text* 8(3): 243–281.

[43] Manning, C. D. and H. Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: The MIT Press.

[44] Marcu, D. 1997. The Rhetorical Parsing of Natural Language Texts. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, 96–103. Madrid, Spain.

[45] Marcus, M. P., B. Santorini, and M. A. Marcinkiewicz. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics* 19(2): 313–330.

[46] Merlo, P. and S. Stevenson. 1999. Automatic Verb Classification Using Distributions of Grammatical Features. In *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics*, 45–51. Bergen, Norway.

[47] Miller, G. A., R. Beckwith *et al*. 1990. Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography* 3(4): 235–244.

[48] Moore, J. and M. Pollack. 1992. A Problem from RST: The Need for Multi-Level Discourse Analysis. *Computational Linguistics* 18(4): 537–544.

[49] Ono, K., K. Sumita, and S. Miike. 1994. Abstract Generation Based on Rhetorical Structure Extraction. In *Proceedings of the 15th International Conference on Computational Linguistics*, 344–348. Kyoto, Japan.

[50] Rooth, M., S. Riezler *et al*. 1999. Inducing a Semantically Annotated Lexicon via EM-Based Clustering. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, 104–111, College Park, MD.

[51] Rosé, C., B. Di Eugenio *et al*. 1995. Discourse Processing of Dialogues with Multiple Threads. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, 31–38. Cambridge, MA.

[52] Siegel, E. V. 1999. Corpus-Based Linguistic Indicators for Aspectual Classification. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, 112–119. College Park, MD.

[53] Siegel, E. V. and K. R. McKeown. 1994. Emergent Linguistic Rules from Inducing Decision Trees: Disambiguating Discourse Clue Words. In *Proceedings of the 12th National Conference on Artificial Intelligence*, 820–826. Seattle, WA.

[54] Sparck Jones, K. 1993. What Might be in a Summary. In *Proceedings of the 1st Conference on Information Retrieval*, 9–26. Konstanz, Germany.

[55] Stolcke, A, K. Ries *et al*. 2000. Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech. *Computational Linguistics* 26(3): 339–373.

[56] Tjong Kim Sang, E. F. 2000. Noun Phrase Recognition by System Combination. In *Proceedings of the 1st North American Annual Meeting of the Association for Computational Linguistics*. Seattle, WA.

[57] Webber, B. L. 1991. Structure and Ostension in the Interpretation of Discourse Deixis. *Language and Cognitive Processes* 6(2): 107–135.

[58] Wiebe, J. M., T. P. O'Hara *et al*. 1998. An Empirical Approach to Temporal Reference Resolution. *Journal of Artificial Intelligence Research* 9: 247–293.