Prosodic Correlates of Rhetorical Relations

Gabriel Murray
Centre for Speech Technology Research
University of Edinburgh
Edinburgh EH8 9LW
gabriel.murray@ed.ac.uk

Maite Taboada
Dept. of Linguistics
Simon Fraser University
Vancouver V5A 1S6
mtaboada@sfu.ca

Steve Renals
Centre for Speech Technology Research
University of Edinburgh
Edinburgh EH8 9LW
s.renals@ed.ac.uk

Abstract

This paper investigates the usefulness of prosodic features in classifying rhetorical relations between utterances in meeting recordings. Five rhetorical relations of *contrast*, *elaboration*, *summary*, *question* and *cause* are explored. Three training methods - supervised, unsupervised, and combined - are compared, and classification is carried out using support vector machines. The results of this pilot study are encouraging but mixed, with pairwise classification achieving an average of 68% accuracy in discerning between relation pairs using only prosodic features, but multi-class classification performing only slightly better than chance.

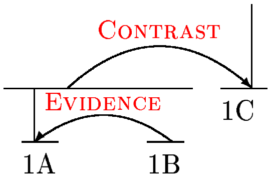
1 Introduction

Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) attempts to describe a given text in terms of its coherence, i.e. how it is that the parts of the text are related to one another and how each part plays a role. Two adjacent text spans will often exhibit a nucleus-satellite relationship, where the satellite plays a role that is relative to the nucleus. For example, one sentence might make a claim and the following sentence give evidence for the claim, with the second sentence being a satellite and the *evidence* relation existing between the two spans. In a text containing many sentences, these nucleus-satellite pairs can be built up to produce a document-wide rhetorical tree. Figure 1 gives an example of a rhetorical tree for a three-sentence text[1].

Theories such as RST have been popular for some time as a way of describing the multi-levelled rhetorical relations that exist in text, with relevant applications such as automatic summarization (Marcu, 1997) and natural language generation (Knott and Dale, 1996). However, implementing automatic rhetorical parsers has been a problematic area of research. Techniques that rely heavily on explicit signals, such as discourse markers, are of limited use both because only a small percentage of rhetorical relations are signalled explicitly and because explicit markers can be ambiguous. RST trees are binary branching trees distinguishing between nuclei and satellites, and automatically determining nuclearity is also far from trivial. Furthermore, there are some documents which are simply not amenable to being described by a document-wide rhetorical tree (Mann and Thompson, 1988). Finally, sometimes more than one relation can hold between two given units (Moore and Pollack, 1992). Given the problems of automatically parsing text for rhetorical relations, it seems prohibitively difficult to attempt rhetorical parsing of speech documents - data which are marked by disfluencies, low information density, and sometimes little cohesion. For that reason, this pilot study sets out a comparatively modest task: to determine whether one of five relations holds between two adjacent dialogue acts in meeting speech. All relations are of the form nucleus-satellite, and the five relation types are *contrast*,

[1] *Contrast* is in fact often realized with a multi-nuclear structure

Proceedings of the Analyzing Conversations in Text and Speech (ACTS) Workshop at HLT-NAACL 2006, pages 1–7,
New York City, New York, June 2006. ©2006 Association for Computational Linguistics



[I love coffee.]^{1A} [I drink it every morning.]^{1B}

[But my brother has never even tried it.]^{1C}

Figure 1: *Sample RST tree*

elaboration, *cause*, *question* and *summary*. This work solely investigates the usefulness of prosodic features in classifying these five relations, rather than relying on discourse or lexical cues. A central motivation for this study is the hope that rhetorical parsing using prosodic features might aid an automatic summarization system.

2 Previous Research

Early work on automatic RST analysis relied heavily on discourse cues to identify relations (Corston-Oliver, 1998; Knott and Sanders, 1998; Marcu, 1997; Marcu, 1999; Marcu, 2000) (e.g., "however" signaling an *antithesis* or *contrast* relation. As mentioned above, this approach is limited by the fact that rhetorical relations are often not explicitly signalled, and discourse markers can nevertheless be ambiguous. A novel approach was described in (Marcu and Echihabi, 2002), which used an unsupervised training technique, extracting relations that were explicitly and unamibiguously signalled and automatically labelling those examples as the training set. This unsupervised technique allowed the authors to label a very large amount of data and pairs of words found in the nucleus and satellite as the features of interest. The authors reported very encouraging pairwise classification results using these word-pair features, though subsequent work using the same bootstrapping technique has fared less well (Sporleder and Lascarides, to appear 2006).

There is little precedent for applying RST to speech dialogues, though (Taboada, 2004) describes rhetorical analyses of Spanish and English spoken dialogues, with in-depth corpus analyses of discourse markers and their corresponding relations. The work in (Noordman et al., 1999) uses short read texts to explore the relationship between prosody and the level of hierarchy in an RST tree. The authors report that higher levels in the hierarchy are associated with longer pause durations and higher pitch. Similar results are reported in (den Ouden, 2004), who additionally found significant prosodic differences between causal and non-causal relations and between semantic and pragmatic relations.

Litman and Hirschberg (1990) investigated whether prosodic features could be used to disambiguate *sentential* versus *discourse* instances of certain discourse markers such as "incidentally." Passonneau and Litman (1997) explored the discourse structure of spontaneous narrative monologues, with a particular interest in both manual and automatic segmentation of narratives into coherent discourse units, using both lexical and prosodic features. Grosz and Hirschberg (1992) found that read AP news stories annotated for discourse structure in the Grosz and Sidner (1986) framework showed strong correlations between prosodic features and both global and local structure. Also in the Grosz and Sidner framework, Hirschberg and Nakatani (1996) found that utterances from direction-giving monologues significantly differed in prosody depending on whether they appeared as segment-intial, segment-medial or segment-final.

3 Defining the Relations

Following Marcu and Echihabi's work, we included *contrast*, *elaboration* and *cause* relations in our research. We chose to exclude *condition* because it is always explicitly signalled and therefore trivial for classification purposes. We also include a *summary* relation, which is of particular interest here because it is hoped that classification of rhetorical relations will aid an automatic speech summarization system. As in Segmented Discourse Representation Theory (SDRT) (Asher and Lascarides, 2004), an alternative framework for representing text structure, we included *question/answer* to our relations list. All training and testing pairs consist of a nucleus followed by a satellite, and the relations are defined as follows:

- **Contrast:** The information in the satellite contradicts or is an exception to the information in the nucleus. Example:

  - Speaker 1: *You use it as a tool*
    Speaker 1: *Not an end user*

- **Elaboration:** The information from the nucleus is discussed in greater detail in the satellite. Example:

  - Speaker 1: *The last time I looked at it was a while ago*
    Speaker 1: *Probably a year ago*

- **Cause:** The situation described in the satellite results from the situation described in the nucleus. Example:

  - Speaker 1: *So the GPS has crashed as well*
    Speaker 1: *So the first person has to ask you where you are*

- **Summary:** The information in the satellite is semantically equivalent to the information in the nucleus. It is not necessarily more succinct. Example:

  - Speaker 1: *The whole point is that the text and lattice are isomorphic*
    Speaker 1: *They represent each other completely*

- **Question/Answer:** The satellite fulfills an information need explicitly stated in the nucleus. Example:

  - Speaker 1: *What does the P stand for anyway?*
    Speaker 2: *I have no idea*

We also took the simplifying step of concentrating only on dialogue acts which did not internally contain such relations as defined above, which could confound the analysis. For example, a dialogue act might serve as a *contrast* to the preceding dialogue act while also containing a *cause* relation within its own text span.

4 Experimental Setup

4.1 Corpus Description

All data was taken from the ICSI Meetings corpus (Janin et al., 2003), a corpus of 75 unrestricted domain meetings averaging about an hour in length each. Both native and non-native English speakers participate in the meetings. The following experiments used manual meeting transcripts and relied on manual dialogue act segmentation (Shriberg et al., 2004). A given meeting can contain between 1000 and 1600 dialogue acts. All rhetorical relation examples in the training and test sets are pairs of adjacent dialogue acts.

4.2 Features

Seventy-five prosodic features were extracted in all, relating to pitch (or *F0*) contour, pitch variance, energy, rate-of-speech, pause and duration. To approximate the pitch contour of a dialogue act, we measure the pitch slope at multiple points within the dialogue act, e.g., the overall slope, the slope of the first 100 and 200 ms, last 100 and 200 ms, first half and second half of the dialogue act, and each quarter of the dialogue act. The pitch standard deviation is measured at the same dialogue act subsections. For each of the four quarters of the dialogue act, the energy level is measured and compared to the overall dialogue act energy level, and the number of silent frames are totalled for each quarter of the dialogue act as well. The maximum F0 for each dialogue act is included, as are the length of the dialogue act both in seconds and in number of words. A very rough rate-of-speech feature is employed, consisting of the number of words divided by the length of the dialogue act in seconds. We also include a feature of pause length between the nucleus and the satellite, as well as a feature indicating whether or not the speakers of the nucleus and the satellite are the same. Finally, the cosine similarity of the nucleus feature vector and satellite feature vector is included, which constitutes a measurement of the general prosodic similarity between the two dialogue acts. The motivation for this last feature is that some relations such as *question* would be expected to have very different prosodic characteristics in the satellite versus the nucleus, whereas other relations such as *summary* might have a nucleus and satellite with very similar prosody to each other.

While there are certainly informative lexical cues to be exploited based on previous research, this pilot study is expressly interested in how efficient prosody alone is in automatically classifying such rhetorical relations. For that reason, the feature set is limited solely to the prosodic characteristics described above.

4.3 Training Data

Using the PyML machine learning tool[2], support vector machines with polynomial kernels were trained on multiple training sets described below, using the default libsvm solver[3], a sequential minimal optimization (SMO) method. Feature normalization and feature subset selection using recursive feature elimination were carried out on the data. The following subsections describe the various training approaches we experimented with.

4.3.1 Manually Annotated Data

For the first experiment, a very small set of manually labelled relations was constructed. Forty examples of each relation were annotated, for a total training set of 200 examples. Each relation has training examples that are explicitly and non-explicitly signalled, since we want to discover prosodic cues for each relation that are not dependent on how lexically explicit the relation tends to be. The percentage of either unsignalled or amibiguously signalled relations across all of the manually-labelled datasets is about 57%, though this varies very much depending on the relation. For example, only just over 20% of *questions* are unsignalled or ambiguously signalled whereas nearly 70% of *elaborations* are unsignalled.

4.3.2 Unsupervised

Following Marcu and Echihabi, we employ a bootstrapping technique wherein we extract cases which are explicitly signalled lexically and use those as our automatically labelled training set. Because those lexical cues are sometimes ambiguous or misleading, the data will necessarily be noisy, but this approach allows us to create a large training set without the time and cost of manual annotation. Whereas Marcu and Echihabi used these templates to extract relation examples and learn further lexical information about the relation pairs, we are using similar templates based on discourse markers but subsequently exploring the extracted relation pairs in terms of prosodic features. Three hundred examples of each relation were extracted and automatically labelled, for a training set of 1500 examples, more than ten times the size of the manually labelled training set. Examples of the explicit lexical cues used to construct the training set are provided in Table 1:

Relation	**Nucleus**	**Satellite**
Contrast	...	*However...*
	...	*But...*
	...	*Except...*
	...	*Although...*
Cause	...	*Therefore...*
	...	*As a result...*
	...	*And so...*
	...	*Subsequently...*
Elaboration	...	*Which...*
	...	*For Example...*
	...	*Specifically...*
Summary	...	*Basically...*
	...	*In other words...*
	...	*I mean...*
	...	*In short...*
Q/A	*Why/What/Where/When*	...
	Who/Did/Is/Are	...

Table 1: Templates for Unsupervised Method

4.3.3 Combined

Finally, the two training sets discussed above were combined to create a set of 1700 training examples.

4.4 Development and Testing Data

For the development set, 35 examples of each relation were annotated, for a total set size of 175 examples. We repeatedly tested on the development set as we increased the prosodic database and experimented with various classifier types. The smaller final test set consists of 15 examples of each relation, for a total set size of 75 examples. Both the test set and development set consist of explicitly and non-explicitly signalled relations. As mentioned above, the percentage of either unsignalled or amibiguously signalled relations across all of the manually-labelled datasets is about 57%

Both pairwise and multi-class classification were

[2] http://pyml.sourceforge.net

[3] http://www.csie.ntu.edu.tw/ cjlin/libsvm/

Relation Pair	Super.	Unsuper.	Combo
Contrast/Cause	0.60	0.67	0.64
Contrast/Summary	0.63	0.57	0.60
Contrast/Question	0.74	0.73	0.80
Contrast/Elaboration	0.61	0.53	0.56
Cause/Summary	0.59	0.60	0.69
Cause/Question	0.84	0.77	0.81
Cause/Elaboration	0.59	0.54	0.56
Summary/Question	0.59	0.60	0.63
Summary/Elaboration	0.70	0.63	0.70
Elaboration/Question	0.90	0.73	0.84
AVERAGE:	**68%**	**64%**	**68%**

Table 2: Pairwise Results on Development Set

carried out. The former set of experiments simply aimed to determine which relation pairs were most confusable with each other; however, it is the latter multi-class experiments that are most indicative of the real-world usefulness of rhetorical classication using prosodic features. Since our goal is to label meeting transcripts with rhetorical relations as a preprocessing step for automatic summarization, multi-class classification must be quite good to be at all useful.

5 Results

The following subsections give results on a development set of 175 relation pairs and on a test set of 75 relation pairs.

5.1 Development Set Results

5.1.1 Pairwise

The pairwise classification results on the development set are quite encouraging, showing that prosodic cues alone can yield an average of 68% classification success. Because equal class sizes were used in all data sets, the baseline classification would be 50%. The manually-labelled training data resulted in the highest accuracy, with the unsupervised technique performing slightly worse and the combination approach showing no added benefit to using manually-labelled data alone. Relation pairs involving the *question* relation generally perform the best, with the single highest pairwise classification being between *elaboration* and *question*. *Elaboration* is also generally discernible from *contrast* and *summary*.

	Cause	Contr.	Elab.	Q/A	Summ.
Cause	**15**	7	11	1	9
Contrast	8	**16**	9	6	5
Elaboration	6	4	**6**	2	4
Question	2	8	4	**17**	**10**
Summary	4	0	5	9	7
SUCCESS:	**34.8%**				

Table 3: Confusion Matrix for Development Set

Relation Pair	Super.	Unsuper.	Combo
Contrast/Cause	0.67	0.47	0.57
Contrast/Summary	0.60	0.43	0.50
Contrast/Question	0.70	0.73	0.77
Contrast/Elaboration	0.67	0.37	0.77
Cause/Summary	0.67	0.63	0.70
Cause/Question	0.87	0.77	0.80
Cause/Elaboration	0.47	0.57	0.50
Summary/Question	0.43	0.60	0.57
Summary/Elaboration	0.77	0.57	0.57
Elaboration/Question	0.80	0.60	0.57
AVERAGE:	**67%**	**58%**	**61%**

Table 4: Pairwise Results on Test Set

5.1.2 Multi-Class

The multi-class classification on the development set attained an accuracy of 0.35 using a one-against-the-rest classification approach, with chance level classification being 0.20. The confusion matrix in Table 3 illustrates the difficulty of multi-class classification; while *cause*, *contrast* and *question* relations are classified with considerable success, the *elaboration* relation pairs are often misclassified as *cause* and the *summary* pairs misclassifed as *question*.

5.2 Test Set Results

5.2.1 Pairwise

The pairwise results on the test set are similar to those of the development set, with the manually-labelled training set yielding superior results to the other two approaches, and relation pairs involving *question* and *elaboration* relations being particularly discernible. The average accuracy of the supervised approach applied to the test set is 67%, which closely mirrors the results on the development set. The most confusable pairs are *summary/question* and *cause/elaboration*; the former is quite surprising in that the *question* nucleus would be expected to have a prosody quite distinct from the others.

5.2.2 Multi-Class

The multi-class classification on the test set was considerably worse than the development set, with a success rate of only 0.24 (baseline: 0.2).

5.3 Features Analysis

This section details the prosodic characteristics of the *manually labelled* relations in the training, development, and test sets.

The *contrast* relation is typically realized with a low rate-of-speech for the nucleus and high rate-of-speech for the satellite, little or no pause between nucleus and satellite, a relatively flat overall F0 slope for the nucleus, and a satellite that increases in energy from the beginning to the end of the dialogue act. Of the manually labelled data sets, 74% of the examples are within a single speaker's turn.

The *cause* relation typically has a very high duration for the nucleus but a large amount of the nucleus containing silence. The slope of the nucleus is typically flat and the nuclear rate-of-speech is low. The satellite has a low rate-of-speech, a large amount of silence, a high maximum F0 and a high duration. There is typically a long duraton between nucleus and satellite and the speakers of the nucleus and the satellite are the same. Of the manually labelled data sets, nearly 94% of the examples are within a single speaker's turn.

The *elaboration* relation is often realized with a high nuclear duration, a high satellite duration, a long pause in-between and a low rate-of-speech for the satellite. The satellite typically has a high maximum F0 and the speakers of the nucleus and satellite are the same. 95% of the manually labelled examples occur within a single speaker's turn.

With the *summary* relation, the nucleus typically has a steep falling overall F0 while the nucleus has a rising overall F0. There is a short pause and a short duration for both nucleus and satellite. The rate-of-speech for the satellite is typically very high and there is little silence. 48% of the manually labelled examples occur within a single speaker's turn.

Finally, the *question* relation has a number of unique characteristics. The rate-of-speech of the nucleus is very high and there is very little silence. Surprisingly, these examples do not have canonical question intonation, instead having a low maximum F0 for the nucleus and a declining slope at the end of the nucleus. The overall F0 for the satellite steeply declines and there is a high standard deviation. The energy levels for the second and third quarters of the satellite are high compared with the average satellite energy and there is very little silence in the satellite as a whole. There is little or no pause between satellite and nucleus and both nucleus and satellite have relatively short durations. The maximum F0 for the satellite is typically low, and the speaker of the satellite is almost always different than the speaker of the nucleus - 99% of the time.

6 Conclusion

These experiments attempted to classify five rhetorical relations between dialogue acts in meeting speech using prosodic features. We primarily focused on pitch contour using numerous features of pitch slope and variance that intend to approximate the contour. In addition, we incorporated pause, energy, rate-of-speech and duration into our feature set. Using an unsupervised bootstrapping approach, we automatically labelled a large amount of training data and compared this approach to using a very small training set of manually labelled data. Whereas Marcu and Echihabi used such a bootstrapping approach to learn additional lexical information about relation pairs, we used the automatically labelled examples to learn the prosodic correlates of the relations. However, even a small amount of manually-labelled training data outperformed the unsupervised method, which is the same conclusion of Sporleder and Lascarides (Sporleder and Lascarides, to appear 2006), and a combined training method gave no additional benefit. One possible explanation for the poor performance of the bootstrapping approach is that some of the templates were inadvertently ambiguous, e.g., “I mean” can signal an *elaboration* or a *summary* and *which* can signal an *elaboration* or the beginning of a *question* relation. Furthermore, one possible drawback in employing this bootstrapping method is that there may be a complementary distribution between prosodic and lexical features. We are using explicit lexical cues to build an automatically labelled training set, but such explicitly cued relations may not be prosodically distinct. For example, a question that is signalled by “Who” or “What” may not have canonical question intonation since it is lexically signalled. This relates to a finding of Sporleder and Lascarides, who report that the unsupervised method of Marcu and Echihabi only generalizes well to relations that are already explicitly signalled, i.e. which could be found just by using the templates themselves.

The pairwise results were quite encouraging, with the supervised training approach yielding average accuracies of 68% on the development and test sets. This illustrates that prosody alone is quite indicative of certain rhetorical relations between dialogue acts. However, the multi-class classification performance was not far above chance levels. If this automatic rhetorical analysis is to aid an automatic summarizaton system, we will need to expand the prosodic database and perhaps couple this approach with a limited lexical/discourse approach in order to improve the multi-class classification accuracy. But most importantly, if even a small amount of training data leads to decent pairwise classification using only prosodic features, then greatly increasing the amount of manual annotation should provide considerable improvement.

7 Acknowledgements

Thanks to Mirella Lapata and Caroline Sporleder for valuable feedback. Thanks to two anonymous reviewers for helpful suggestions. This work was partly supported by the European Union 6th FWP IST Integrated Project AMI (Augmented Multi-party Interaction, FP6-506811, publication 177).

References

N. Asher and A. Lascarides. 2004. *Logics of Conversation*. Cambridge University Press, Cambridge, GB.

S. Corston-Oliver. 1998. *Computing representations of the structure of written discourse*. Ph.D. thesis, UC Santa Barbara.

H. den Ouden. 2004. *The Prosodic Realization of Text Structure*. Ph.D. thesis, University of Utrecht.

A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters. 2003. The ICSI meeting corpus. In *Proceedings of IEEE ICASSP 2003, Hong Kong, China*.

Alistair Knott and Robert Dale. 1996. Choosing a set of coherence relations for text-generation: a data-driven approach. In Giovanni Adorni and Michael Zock, editors, *Trends in natural language generation: an artificial intelligence perspective*, pages 47–67. Springer-Verlag, Berlin.

A. Knott and T. Sanders. 1998. The classification of coherence relations and their linguistic markers: An exploration of two languages. *Journal of Pragmatics*, 30:135–175.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

D. Marcu and A. Echihabi. 2002. An unsupervised approach to recognizing discourse relations. In *The Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA*.

D. Marcu. 1997. From discourse structures to text summaries. In *The Proceedings of the ACL’97/EACL’97 Workshop on Intelligent Scalable Text Summarization, Madrid, Spain*, pages 82–88.

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

D. Marcu. 2000. The rhetorical parsing of unrestricted texts: A surface-based approach. *Computational Linguistics*, 26(3):395–448.

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

L. Noordman, I. Dassen, M. Swerts, and J. Terken. 1999. Prosodic markers of text structure. In K. Van Hoek, A. Kibrik, and L. Noordman, editors, *Discourse Studies in Cognitive Linguistics*, pages 133–149. John Benjamins Publications, Amsterdam, NL.

E. Shriberg, R. Dhillon, S. Bhagat, J. Ang, , and H. Carvey. - 2004. The ICSI meeting recorder dialog act (MRDA) corpus. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue, Cambridge, MA, USA*, pages 97–100.

C. Sporleder and A. Lascarides. to appear, 2006. Using automatically labelled examples to classify rhetorical relations: An assessment. *Natural Language Engineering*.

M. Taboada. 2004. *Building Coherence and Cohesion: Task-Oriented Dialogue in English and Spanish*. John Benjamins Publications, Amsterdam, NL.