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.

- **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

Example: thical5olus. (it)Tj 8n.8582 0 Td (the)Tj 1necessarid (sema0Td he)Tj 1more36 0 Td

Speakat *3thTj 8th3.1781 nti(from)Tj 26.2363 0 Td (the)Tj 1t1.3164 0ellitesatf -2u-*

6ultsed

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

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

	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**most confusable with**

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*.

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

nalled 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 of