A Maximum Entropy Approach to Recognizing Discourse Relations in Spoken Lanugage

Jeanette Pettibone and Heather Pon-Barry
{pettibon|ponbarry}@stanford.edu

June 6, 2003

Abstract

Our project examines the automatic classification of discourse relations in spoken language data. We use a maximum entropy model and train our classifier to recognize general discourse relations (e.g. contrast, elaboration) as well as discourse relations unique to conversation (e.g. acknowledge, question-confirmation). We then compared these results to those of a simple naive bayes classifier. The overall performance of our maximum entropy classifier was better than the naive bayes performance, especially for the classification of conversational discourse relations, with accuracies as high as 85.7%.

1 Introduction

When viewed individually, each sentence in a discourse can be seen as a well-formed grammatical unit. However, when we look at the discourse as a whole, the relationship between the sentences becomes extremely important. For example, look at the following pairs of sentences (taken from (Kehler 2002:2)):

(1) a. John took a train from Paris to Istanbul. He has family there.

b. John took a train from Paris to Istanbul. He likes spinach.

The pair of sentences in example (1a) have a relationship. The second sentence explains the cause of the first. In contrast, the pair in (1b) have no relationship to each other. This makes the pair together sound odd when together.[1]

Our goal here is to build an automatic discourse relation recognizer for a corpus of spoken language. We are building on the work of Marcu and Echihabi (2002) who use a Naive Bayes classifier to recognize discourse relations in a corpus of written text. Our task is different from

[1] Of course, we can create a context in which this pair is not strange. For example, if it was an established fact that Istanbul had fabulous spinach dishes and that people go there specifically to eat spinach, this pair would not seem infelicitous.

theirs in that spoken data has much more noise than written. When people are speaking, they tend to make many false starts and repairs. There are also numerous pauses filled with 'uh' and 'um'. Another difference is our approach. We will be using a maximum entropy model, which eliminates the independence assumption of Naive Bayes and has been proven to out-perform Naive Bayes in a number of NLP tasks.

1.1 Possible applications

Marcu and Echihabi (2002) point out a few possible applications of a discourse relation recognizer. A recognizer could aid in the discourse parsing which would then help to improve text summarization. Also, it would help in information retrieval. For example, a question such as *Why did Sally put her dog in the garage?* could be more easily answered if there is a CAUSE relation labeled.

Because our project examines relations specifically in spoken language, it also has possible applications involving dialogue systems. A classifier that robustly handles a variety of conversational discourse relations would be useful for interpreting speech acts and understanding dialog structure, two tasks that are essential for any dialogue manager.

2 Design and Methods

2.1 A Theory of Discourse Relations

As Knott et. al. (2001) point out, the first two steps in designing a theory of discourse relations are to select the relations that you want to use and to specify the span over which those relations hold.

(1) **the discourse relations**

We started with the five discourse relations of Marcu and Echihabi (2002). Their relations are a parsed down selection of the 23 relations originally proposed by Mann and Thompson (1988). These relations are CONTRAST, CAUSE-EXPLANATION-EVIDENCE, CONDITION, ELABORATION, NONE-OF-THE-ABOVE. We used these relations but, as we tried to specify the needs of spoken language, we added additional relations. Our final list is:

- CONTRAST
  "And we have those creatures down here that uh, like to go hunting for a midnight snack"
  "But they won't necessarily attack a human or bite you or anything"
  "They will if you attack them"
- EVIDENCE[2]
  "We don't really have any problem with finding day care for him because I'm in the service, and we just take him to the day care on base"

[2]Note that by reducing the name of the relation, we do not mean to be reducing the information that it contains. It is still the same as Marcu and Echihabi's CAUSE-EXPLANATION-EVIDENCE

- CONDITION
  "I would really be disturbed if I found out that the place I took it to was going to just put it in the landfill"
- ELABORATION
  "I don't think the teachers get enough support, quite honestly"
  "Yeah"
  "Uh, for example, my wife, she has kids that are on drugs, and of course like everyone, she has kids that get pregnant even, even in the seventh grade"
- ACKNOWLEDGE
  "I want to preserve the value of the house by making sure the yard is taken care of"
  "Yes, okay"
- DISAGREE
  "I thought well, golly gee. She's got to be lying. I'd say I slept with him for a hundred thousand bucks, you know"
  "Yeah. But I don't know, I mean, it's just the press is always going to be like that I think"
- QUESTION-CLARIFICATION
  "One is three and one's one"
  "Wow. You got two little ones then, huh?"
  "Yeah"
- OTHER

(2) **the span over which discourse relations apply**
As we're interested in looking at spoken language, we used the transcripts of spoken language in the Switchboard corpus. In looking through the conversations, there seemed to be a trend that the current speaker in a dialogue continue speaking and so there was more discourse coherence from one one turn of the speaker to the next. We decided on a span of three turns: the previous turn of participant A, the turn of participant B, and the current turn of A. For example:

(2) A: Well, we got the bloodhound, and the rabbits and the bloodhound didn't mix well, so –
B: Uh-huh.
A: – they were easier to get rid of than the big dog, and –

By having quite a large span, we hope to capture the discourse relations from one turn to the next. However, we do risk not being able to identify exactly which are the pieces within the span that share the relationship.

2.2 Maximum Entropy

Maximum Entropy modeling has been used in a wide variety of classification problems within NLP. [3] In the field of computational discourse in particular, Kehler (1997) uses maximum entropy instead of a discourse algorithm like Centering Theory for determing coreference.

Maximum entropy is a great model for linguists. It allows them to place a large number of constraints on the model they build and therefore allows them to incorporate any sort of prior expectations that they have concerning the information that is important for the classification process. Each of these constraints, or *features*, are then used to estimate the expectation of each feature given the training data. This gives us the *parameters* of our maximum entropy model which will be used in the classification task.

For this project, we used simple binary features. Each feature denoted the occurence of a particular word in a span with a particular discourse relation. For example, one feature was the occurence of the word *but* in a span with the relation CONTRAST. In addition, we also had features about the word's position which denoted the occurence of a particular word before the cue word given a particular class and the occurence of a particular word after the cue word given a particular class. By using the positional features, we imporved the performance of our classifier.

We used Rob Malouf's (2002) maximum entropy software to evaluate the features and give us the parameters of the model.

2.3 Our Discourse Relation Recognizer

We wrote a script that found the features in the training data, evaluated them with the maximum entropy software, and then used the parameters of our maximum entropy model to classify spans with the appropriate discourse relation.

Step 1 (**pre-processing**) We cleaned up the Switchboard corpus so that there was one turn per line. Then we divided it into 3 parts: 80set, and the last 10percentages as we were careful to maintain the integrity of each conversation within the corpus. The Switchboard contains about 3 million words in about 2400 conversations. For our purposes, this means that We had around 140,000 spans in the training data.

Step 2 Our script first went through the training data to gather the features and their frequencies. Because we did not have an annotated corpus, we used "cue words" as approximate labels for the discourse relations. For example:

[3] Manning and Schütze (2000:607-8) point out some examples, including "tagging(Ratnaparkhi 1997), prepositional phrase attachment (Ratnaparkhi 1998), sentence boundary detection (Mikheev 1998), named entity recognition (Borthwick et. al. 1998) and partial parsing (Skut and Brants 1998)" More recently, Johnson and Goldwater have used maximum entropy instead of the Gradual Learning Algorithm for Optimality Theory.[get a reference here]

- CONTRAST: but, although
- EVIDENCE: because
- CONDITION: if, unless
- ELABORATION: which, that, for example, so that
- ACKNOWLEDGE: yeah, yes, right, true, uh-huh
- DISAGREE: no, i don't think
- QUESTION: huh, right at EOS

Once we gathered all the features within each span, we outputed them to an event file that was in the correct format to be the input for the maximum entropy software.

Step 3 We then used the file of events that we created as input to Rob Malouf's maximum entropy software. The software then analyzed the features and returned the parameters in a new file. Our script then went through the new file and collected all the parameters and stored them in an array.

Step 4 Using our newly created parameters, we ran the test data and for each instance, selected the discourse relation with the highest overall probability.

Step 5 (**post-processing**) We previously modified our Naive Bayes classifier to also be able to chose the best relation given the test data. We then compared the results of our maximum entropy classifier with those of the Naive Bayes.

3 Linguistic Assumptions

By building a classifier for both general discourse relations and conversational discourse relations, we are mixing two areas (relational semantics, and conversational analysis) that are usually kept separate. We feel this is justified because the classifications we chose are mutually exclusive, with the exception of CONTRAST and DISAGREE (e.g. A: "Is her party this Friday?" B: "I don't think so, I heard it's on Saturday" would fall under both classifications). However, cases like this occur rarely in the Switchboard data. In the 106 instaces we hand-evaluated, this sort of ambiguity never arose.

4 Results and Discussion

4.1 Optimization

Marcu and Echihabi (2002) found that using the "most representative" words in the corpus (i.e. nouns, verbs, cue words), they significally improved the results of their discourse relation recognizer. However, this did not improve our results. One of the problems is that we are working with spoken languge. Words like *yeah, um* are labeled as nouns and so when we pulled out only the nouns, verbs, and cue words, we were left with a corpus that full of mostly filler words and confirmations. If we also extracted these words, we would be left with a very small amount of information. Also, though we did not study this, it is possible that words like *uh* and *um* are more used in certain relations than others. This would make an interesting study for the future.

One of the advantages of using a maximum entropy model is that we can use as many features as we find appropriate. As mentioned before, we used not only features denoting the occurence of a word within a span with a particular relation (which might be most similar to Naïve Bayes), we also added features concerning the position of the word. Although our maximum entropy classifier using just the occurence of the words still worked better than our Naïve Bayes classifier, there was significant improvement when we added the additional features.

4.2 Results

One problem we faced was the task of evaluation. We originally thought that there was a classifier available, but soon learned that it was really a markup tool to aid classification-by-hand. So, in order to evaluate our performance, we annoted by hand relations for 106 randomly selected instances from our test results. Table 1 shows a confusion matrix (rows = ME classification, columns = correct classificaiton) with the results from our maximum entropy classifier. Table 2 shows the precision and recall measures.

Table 1. Confusion matrix for Maximum Entropy classifier

class	contrast	evidence	condition	elaboration	acknowledge	disagree	question	other
contrast	**4**	0	3	7	3	0	0	2
evidence	3	**5**	1	2	3	0	0	1
condition	1	2	**6**	4	4	0	0	1
elaboration	0	0	0	**12**	5	0	0	0
acknowledge	0	0	0	2	**18**	0	0	1
disagree	0	0	0	1	2	**2**	0	0
question	0	0	0	1	1	0	**0**	1
other	0	0	0	0	1	0	1	**4**

							instances	106
							accuracy	0.481132

Table 2. Precision and Recall for Maximum Entropy Classifier

class	precision	recall
contrast	0.2105	0.5000
evidence	0.2941	0.7143
condition	0.3333	0.6000
elaboration	0.7059	0.3871
acknowledge	0.8571	0.4865
disagree	0.4000	1.0000
question	0.0000	0.0000
other	0.6667	0.4000

Tables 3 and 4 show the results from our naive bayes classifier.

Table 3. Confusion matrix for Naïve Bayes classifier

class	contrast	evidence	condition	elaboration	acknowledge	disagree	question	other
contrast	**4**	1	1	5	10	1	2	0
evidence	1	**6**	1	6	8	0	0	1
condition	4	1	**2**	6	4	1	1	3
elaboration	0	1	0	**1**	2	0	0	1
acknowledge	0	0	0	1	**4**	1	0	2
disagree	0	0	0	0	0	**0**	0	0
question	0	0	0	1	0	0	**1**	2
other	2	0	0	1	10	0	0	**7**
							instances	106
							accuracy	0.235849

Table 4. Precision and Recall for Naïve Bayes Classifier

class	precision	recall
contrast	0.1667	0.3636
evidence	0.2609	0.6667
condition	0.2609	0.6667
elaboration	0.2000	0.0476
acknowledge	0.5000	0.1053
disagree	0.0000	0.0000
question	0.2500	0.2500
other	0.3500	0.4375

4.3 Discussion

While the overall accuracy of our maximum entropy classifier was not as high as expected (48.1%), it is significantly better than the overall accuracy of our naive bayes classifier (23.6%). For 7 out of the 8 relations examined, the maximum entropy classifier has higher precision, and for 6 out of the 8 relations it has higher recall.

In particular, the maximum entropy classifier has high precision for acknowledgement (85.7%) and elaboration (70.6%). We feel that some of these results are skewed by sparsity of the relation coupled with the limitations of our evaluation procedure (hand-classification). For example, if we were able to evaluate all 24,000 instances in our test set rather than just 106, we would have more than 1 instance of a question-clarification relation, and a better indication of our system's performance.

The two relations with the best precision, acknowledgement and elaboration, are also the two relations that occurred most frequently in our data. Given the nature of the Switchboard corpus, this is not surprising–most conversations tend to be monologues with interspersed acknowledgements where the person giving the monologue switches back and forth.

Surprisingly, acknowledge and elaboration are among the relations with the lowest recall. We are not sure why this is the case, but it may be related to the way in which we extracted instances and features from the training data. It is possible that the cue words we used for the elaboration relation do not reflect all the ways in which elaboration occurs. Or maybe we shouldn't have used *but* as a cue word for contrast. There were many cases that our classifier labeled an elaboration relation as contrast (see Table 1). In written text, "but" is a good indicator of a contrast relation, however in conversation it is frequently used when elaborating (e.g. "I bought a CD, but I haven't listened to it yet") in addition to when signaling contrast.

5 Conclusions and Future Work

Our results indicate that using a maximum entropy model is a good approach to classifying discourse relations in spoken language–it performs much better than a naive bayes model on the same task. However, our results also indicate that in order to build a decent classifier, there are many issues that need to be investigated further.

While it is promising that our classifier performs well on acknowledge and elaboration relations, in order to be glean anything useful about the structure or content of conversations, it needs to perform well on other relations (e.g. disagree, question-clarification) too. We think a classifier of this nature would be very useful for conversational dialogue systems if it could recognize relations such as confusion, or change-of-topic, but early on we realized that this was beyond the scope of this project.

One area for improvement is strategies for feature selection. The ability to design interesting features is one of the characteristics that makes maximum entropy models so appealing, and we wish we had more time to experiment with this. Perhaps our choice of a three-turn instance was too broad, and our choice of features too unrefined. For example, it would be an interesting experiment to incorporate identity of the speaker (either A or B) into our features, vary the number of turns per instance, and only look at relations that hold between the two speakers. There are many possibilities for future work.

References

[1] Adam L. Berger, Stephan A. Della Pietra, and Vincent J. Della Pietra. 1996. A maximum entropy approach to natural language processing. In *Computational Linguistics*, 22:39-71.

[2] Barbara J. Grosz and Candace L. Sidner. 1986. Attention, Intentions, and the Strusture of Discourse. In *Computational Linguistics*, 12:175-204.

[3] Daniel Jurafsky and James H. Martin. 2000. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall.

[4] Andrew Kehler. 1997. Probabilistic coreference in information extraction. In *EMNLP*, 2:163-173.

[5] Andrew Kehler. 2002. *Coherence, Reference, and the Theory of Grammar*. CSLI Publications.

[6] Alistair Knott, Jon Oberlander, Michael O'Donnell, and Chris Mellish. 2001. Beyond elaboration: the interaction of relations and focus in coherent text. In T. Sanders, J. Schilperoord, and W. Spooren, eds., *Text representation: linguistic and psycholinguistic aspects*, 181-196. John Benjamins Publishing Company.

[7] Robert P. Malouf. 2002. A comparison of algorithms for maximum entropy parameter estimation. In *Proceedins of the Sixth Conference on Natural Language Learning (CoNLL-2002)*, 49-55.

[8] Robert P. Malouf. 2002. Maximum entropy software. Downloadable at `http://www-rohan.sdsu.edu/ malouf/pubs.html`.

[9] William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8:243-281.

[10] Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press.

[11] Daniel Marcu and Abdessamad Echihabi. 2002. An Unsupervised Approach to Recognizing Discourse Relations. In *Proceedings of the 40th Annual Meeting of the Association for Computational Lingusitics*. Philadelphia, PA, July 7-12.