Predicting the Presence of Discourse Connectives

Gary Patterson and **Andrew Kehler**
Department of Linguistics
UC San Diego
9500 Gilman Drive #0108
La Jolla, CA 92093
{gpatterson,akehler}@ucsd.edu

Abstract

We present a classification model that predicts the presence or omission of a lexical connective between two clauses, based upon linguistic features of the clauses and the type of discourse relation holding between them. The model is trained on a set of high frequency relations extracted from the Penn Discourse Treebank and achieves an accuracy of 86.6%. Analysis of the results reveals that the most informative features relate to the discourse dependencies between sequences of coherence relations in the text. We also present results of an experiment that provides insight into the nature and difficulty of the task.

1 Introduction

A central goal of natural language generation and summarization systems is to produce interpretable, coherent text that rivals material a human would produce. Doing so requires that systems not only have the ability to generate clauses that are grammatical and easy for people to process, but also the ability to employ the appropriate discourse structuring devices needed to yield fluid transitions between these clauses. This is a tricky issue in that it requires that a balance be achieved between the opposing goals of communicative expressiveness and economy. On the one hand, insufficient cueing of inter-clausal relationships can lead to a discourse that is at best difficult to process, and at worst misunderstood. On the other hand, too much explicit marking can result in a clunky and even redundant sounding discourse.

Here we consider the question of when to explicitly mark the COHERENCE RELATIONS in discourse, that is, the inter-clausal relationships that the language producer intends the interpreter to infer between the meanings of clauses (Hobbs, 1979; Mann and Thompson, 1988; Kehler, 2002; Asher and Lascarides, 2003). Consider, for example, the EXPLANATION coherence relation that holds in (1), in which the second clause provides a cause or reason for the eventuality described in the first:

(1) a. Max will visit Australia this summer **because** his father is turning 65.
    b. Max will visit Australia this summer. His father is turning 65.

As example (1) shows, coherence relations can be marked explicitly—by using lexical connectives such as coordinating or subordinating conjunctions (e.g., *because* in 1a) or certain types of prepositional or adverbial phrases—or left implicit as in (1b). Either way, establishing the relation itself requires the reader to go through a complex inferential process necessitating that a variety of assumptions be made, typically supported by context and/or world knowledge, that are not explicitly asserted by the actual linguistic material. In (1), for instance, such inferences would include that Max intends to see his father when he travels to Australia, that his father resides in that country, and that the birthday will taking place during the time of the visit. Importantly, the role of the connective in (1a) is therefore not to *establish* that an EXPLANATION relation holds. Instead, connectives serve the function of directing the addressee's inference processes toward a smaller set of coherence relations than might otherwise be available, among other possible roles.

The fact that both (1a) and (1b) are felicitous may lead us to believe that the choice to insert a connective between clauses is simply optional. This is not

Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 914–923,
Seattle, Washington, USA, 18-21 October 2013.

always the case, however. Sometimes the use of a connective is required, since omitting it would likely result in incorrect inferences being drawn by the addressee. For example, the use of *when* in (2a) implies a backward temporal ordering of events, which is reversed if the connective is left out, as in (2b).

(2) a. Maggie fell over in shock **when** Saul offered to help her.
    b. Maggie fell over in shock. Saul offered to help her.

On the other hand, a connective can seem unnecessary if the relation between the two clauses is sufficiently implied by other cues in the text. For instance, since the act of throwing a vase against a concrete wall would normally be expected to cause the vase to break, the adverbial phrase *as a result* in (3a), while felicitous, seems overly verbose and perhaps even redundant.

(3) a. Susan threw the fragile vase against the concrete wall. **As a result**, it broke.
    b. Susan threw the fragile vase against the concrete wall. It broke.

The foregoing examples suggest that the appropriateness of including an explicit connective is inherently gradient, and is in fact correlated with ease of inference: the more difficult recovering the correct relation would be without a connective, the more necessary it is to include one. This characterization in turn suggests that predicting whether a connective should be included might be a difficult problem for an NLP system to address, since current-day systems lack the requisite world knowledge and capacity for inference that would be necessary to evaluate the ease with which coherence relations can be established on arbitrary examples. However, it is also possible that the decision to include a connective depends in part on stylistic and other types of factors as well, such that there might be predictive information in the kinds of shallow linguistic and textual features that systems *do* have access to. This is the question taken up in this work: Given two adjacent clauses in a text, the type of coherence relation holding between them and a candidate connective that could be used to signal the relation, we ask how well a system can predict whether or not that connective was used by the author of the text. This capability would be useful to generation systems as a post-cursor to discourse-level message planning and sentence realization processes, as well as summarization systems that take existing sentences and have to reconsider connective placement upon reassembling them.

To our knowledge, there is no work in the literature that addresses this issue directly. There is a growing body of research (Sporleder and Lascarides, 2008; Pitler et al., 2009; Lin et al., 2009; Zhou et al., 2010) that focuses on building supervised models for classifying implicit relations using a variety of contextual features, such as the polarity of clauses, the semantic class and tense/aspect of verbs, and information from syntactic parses. With respect to explicit relations, Elhadad and McKeown (1990) sketch a procedure to select an appropriate connective to link two propositions as part of a larger text generation system, using linguistic features derived from the sentences. The procedure selects the best connective from a given set of candidates, but does not allow for the option of leaving the relation implicit. More recently, Asr and Demberg (2012a) look at both explicit and implicit relations, and make the observation that certain relation types are more likely to be realized explicitly than others. Relatedly, Asr and Demberg (2012b) discuss which connectives are the strongest predictors of which relation types. However, there is no work of which we are aware that specifically predicts whether connectives should be used or omitted.

2 Classification Model

Our model is a binary classifier trained on data extracted from the Penn Discourse Treebank (PDTB; Prasad et al. (2008)), a large-scale corpus of annotated discourse coherence relations covering the one-million-word Wall Street Journal corpus of the Penn Treebank (Marcus et al., 1993).

2.1 Data

For every relation in the PDTB, the following components are annotated: (i) the connective used to signal the relation; (ii) the textual spans of the two clausal arguments that constitute the relation; (iii) the semantic sense of the relation, according to a hierarchical tagset of senses; and (iv) the attribution of the assertions and beliefs expressed in the text to the

relevant individuals. Crucially for our purposes, for the implicit relations the corpus indicates the most suitable connective if the relation were instead signaled explicitly. For example, the annotators decided that the best connective to signal the REASON relation in (4) would be *because*, rather than other plausible candidates, such as *as* or *since*.

(4) It's a shame their meeting never took place. [IMPLICIT=*because*] Mr. Katzenstein certainly would have learned something. (WSJ0037)

In total, there are 18,459 explicit and 16,053 implicit relations annotated in the PDTB. We excluded a subset of these cases in training our model based on two criteria. First, whereas explicit relations in the PDTB can hold between spans of text that either are or are not adjacent, we excluded the non-adjacent cases. This was done to ensure consistency in discourse structure between the relations considered in the model, since only the implicit relations between adjacent clauses were annotated. Second, we excluded relations that have lower frequency semantic senses or use low frequency connectives. As a result, the model considers only the eight most common semantic senses of relations, which in total account for just less than 90% of the relations in the corpus.[1] Further, for each relation, we only consider the connectives that account for more than 5% of the instances of that relation. After applying these filters, the resulting corpus comprised 10,039 explicit and 11,690 implicit relations.

Table 1 shows the eight relations that were modeled. The majority of these relations exist at the middle layer of the three-level hierarchy of semantic senses annotated in the PDTB.[2] Relations at the highest level – representing the four major semantic categories COMPARISON, CONTINGENCY, EXPANSION and TEMPORAL – were deemed too broad to be of practical use in a generation system, whereas the lowest-level senses were considered either unnecessarily fine-grained or have too few tokens in the corpus to allow for meaningful statistical modeling. Two exceptions were made for REASON and RESULT relations, which do appear at the lowest level in the PDTB hierarchy (beneath the CAUSE category). These were included because they are both attested frequently in the corpus and are undeniably contrastive: with REASON, the second clause provides an explanation for the proposition expressed in the first clause, whereas with RESULT, the second clause describes a consequence of the first. It is reasonable to want these relations to be modeled separately.

Sections 2-22 of the corpus were used as the training set, and sections 23 and 24 were used as the test set. Sections 0 and 1 of the corpus were set aside as a development set for feature design and parameter optimization. The training set comprised 18,218 tokens, distributed as shown in Table 1.

Relation Type	**Explicit**		**Implicit**	
Asynchronous	1,120	(70%)	469	(30%)
Conjunction	2,940	(61%)	1,906	(39%)
Contrast	2,044	(66%)	1,054	(34%)
Instantiation	203	(16%)	1,093	(84%)
Reason	771	(28%)	1,938	(72%)
Restatement	76	(4%)	2,081	(96%)
Result	354	(21%)	1,295	(79%)
Synchronous	748	(86%)	126	(14%)
Total	**8,256**	(45%)	**9,962**	(55%)

Table 1: Distribution of Training Set

As Table 1 shows, the preference for an overt connective varies significantly according to the type of relation. The ASYNCHRONOUS, CONJUNCTION, CONTRAST and SYNCHRONOUS relations are realized explicitly the majority of the time, whereas INSTANTIATION, REASON, RESTATEMENT and RESULT relations are more often left implicit. We can also see that some relation types (such as RESTATEMENT, INSTANTIATION and SYNCHRONOUS) exhibit a strong preference to be realized in a particular form, whereas other types show more variability in whether they are realized explicitly or implicitly.

The distribution of tokens in Table 1 can be used to determine a baseline accuracy against which the performance of our model is evaluated. A naive model that uses the semantic type of the coherence relation as the sole predictive feature makes a binary classification based simply on the majority category for that relation type. A baseline model using this methodology results in classification accuracy of 77.0% over the held-out test set.

[1] The next most common relation type, CONDITION, was excluded because it is always marked explicitly in the corpus.

[2] For more details of the PDTB sense hierarchy, see Prasad et al. (2008).

2.2 Model

We built a composite model containing binary logistic regression classifiers for each coherence relation, trained on a set of linguistic features extracted from each token in the training set. Logistic regression was chosen because it produces a model with high performance and results that are easily interpretable. The features included in the model fall into the following three broad classes: relation-level, argument-level, and discourse-level.

Relation-level features

In addition to the semantic type of the relation, we include as a feature the connective used to signal the relation in the text (or, for the implicit relations, the connective indicated by the annotators as most appropriate). This feature (*Connect*) is included based upon the observation that connectives vary as to their rates of being realized explicitly—even for connectives that signal relations with the same semantic sense. Consequently, given a relation of a particular semantic type, an indication of the best fitting connective may be a consistent predictor of whether or not this relation is realized explicitly.

We also include a feature reflecting the attribution of the relation. As mentioned above, the PDTB is annotated to describe the attribution of the propositions expressed within a relation to individuals or entities in the text. For example, in the relation shown in (5), the first clause contains a direct quotation, clearly attributing the proposition expressed to the individual *Rep. Stark*. However, the second clause contains no such indication of attribution to an entity in the text, and so the proposition is instead assumed to be asserted by the writer of the article.

(5) "No magic bullet will be discovered next year, an election year," says Rep. Stark. **But** 1991 could be a window for action. (WSJ0314)

Inspection of the corpus data suggests that when one argument of a relation contains a proposition that is attributable to an individual in the text (either by direct or indirect quotation) but the other is assumed by default to be attributed to the author, this relation is more likely to be realized explicitly. This may well have an explanation based on sentence processing: the intervening attribution phrase *'says Rep. Stark'* may serve as a distraction, with the result that the intended coherence relation is harder to infer without a connective. Consequently, we include a factor (*AttMismatch*) indicating if the two arguments are not attributed to the perspective of the same individual.

Finally, in any particular genre there may be formulaic prose whose systematic features can be exploited by a system tasked with generating text within that same genre. In this case, the genre represented by the corpus data comprises copy-edited articles from the Wall Street Journal, many of which refer to company earnings reports or other financial events, and are written in a highly prescribed style. Accordingly, we may suspect that there is a greater prevalence of implicit relations in these cases, since the reader is assumed to be habituated to the way in which the information in this type of article is presented. Consequently, for the domain at hand we include a binary feature (*Financial*) indicating whether the relation pertains to financial information. This feature takes the value 1 if the textual spans of both arguments in the relation contain percentage amounts or dollar figures.

Argument-level features

For each relation, the model includes features capturing the size or complexity of each of its two arguments. The arguments were identified by the annotators according to a principle of minimality, whereby the annotations indicate the shortest text spans necessary for the appropriate coherence relation to be interpreted. However, the annotators also indicated other text that is in some way relevant to the interpretation of the arguments. This supplementary material can include unrestricted relative clauses, appositives, or other parenthetical information. Our observation of the data indicates that relations which have supplementary material annotated alongside one or both of their arguments are more often than not realized explicitly with connectives. As a result, we include binary features (*Supp1*, *Supp2*) indicating whether the first and second arguments of the relation include such supplementary information. We also include features (*Length1*, *Length2*) reflecting a simple measure of the length of each argument, calculated as the log transformed count of the number of words in the arguments' minimal text spans.

One measure of the complexity of an argument

is the number of clauses it contains. It might be thought that the greater the syntactic complexity of an argument, the more likely it is that the relation containing it is marked explicitly, so as to give the reader more help in drawing the correct intended inference between the arguments. As a proxy for the number of clauses in each argument, we include features (*NPSbj1*, *NPSbj2*) equal to the total number of main, subordinate, or complement clause subjects included within the textual spans of the respective arguments, determined using the syntactic parses available in the corpus data.

We also consider whether the underlying richness of the informational content expressed by the argument may influence the presence or omission of a connective. Considering the way in which readers process text in real time, it would intuitively be more difficult to infer the intended relation between two clauses without the aid of a connective if the arguments themselves had greater processing demands owing to increased lexical retrieval, reference and anaphor resolution requirements, and so forth. Given this intuition, we may expect that arguments with higher density of information are correlated with the increased use of connectives as a means of facilitating the inference of the relation type and thereby easing the overall processing burden. Consequently, our model includes features (*ContDensity1*, *ContDensity2*) calculated as the ratio of the count of words in each argument that are content words (i.e. ignoring articles, prepositions and pronouns), divided by the total number of words, as well as features (*PronDensity1*, *PronDensity2*) calculated as the ratio of pronouns in each argument to the number of noun phrases.

Finally, the accessibility of the subject of the second argument in a relation may play a role in determining whether the relation is explicitly marked. Specifically, informal observation of the data suggests that there is a tendency for the second argument of an implicit relation to begin with a longer, contentful noun phrase, rather than a pronoun. Consequently, our model includes a binary feature (*FirstA2Pron*) indicating whether the first word in the second argument is a pronoun.

Discourse-level features

The final class of features takes account of the way in which a relation fits into the broader discourse structure in the text. In their work on implicit relation classification, Pitler et al. (2008) identified various dependencies between bigram sequences of explicitly- and implicitly-realized relations of different semantic types. These results suggest that the semantic type and the presence of a connective in one relation may be predictive of whether or not the following relation in the text is marked with a connective. Consequently, we include features indicating the semantic type of the relation occurring immediately prior in the text (*PrevSemType*), and whether this relation was marked implicitly or explicitly (*PrevForm*).

The other discourse-level features take account of the dependencies between the relation in question and its neighboring relations in the text. As part of a supervised learning model developed to classify the semantic class of implicit relations in the PDTB, Lin et al. (2009) found features based on the two main types of discourse dependency pattern in the corpus ('shared' and 'fully embedded' arguments) to be highly predictive. We speculatively include similar features in our model to see if they are helpful in predicting the presence of connectives.

The first type of dependency between adjacent relations is one where the second argument of one relation is also the first argument of the following relation, as in Figure 1. Accordingly, we include two binary features indicating whether an argument is shared with the preceding relation (*Arg1isPrevArg2*) or the following relation (*Arg2isNextArg1*) in the corpus.

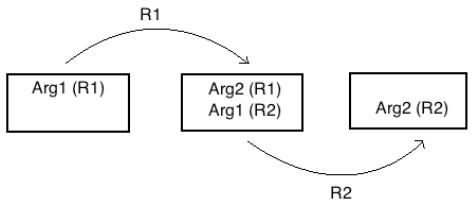

Figure 1: Shared argument

The other main type of discourse dependency, a 'fully embedded' dependency, is one where an entire relation (including both of its arguments) is completely embedded within one argument of an adjacent relation in the text, as in Figure 2. To capture

this type of dependency structure, we include two binary features (*EmbedNext*, *EmbedPrev*) indicating whether the current relation is embedded within either one of its adjacent relations. We also include two binary features (*Arg1Embed*, *Arg2Embed*) to indicate whether either argument of the current relation completely contains an embedded relation.

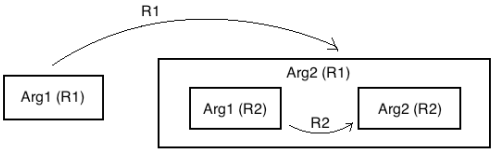

Figure 2: Fully embedded argument

The two relations in (6) exemplify a typical instantiation of this embedded dependency structure.

(6) It is an overwhelming job. [IMPLICIT= *because*] There are so many possible proportions **when** you consider how many things are made out of eggs and butter and milk. (WSJ0261)

In this example, there is an implicit REASON relation holding between the two complete sentences, and an explicit SYNCHRONOUS relation signaled by the connective *when* holding between the two clauses of the second sentence. Since the REASON relation fully embeds another relation within its second argument, the feature *Arg2Embed* for this relation takes the value 1. For the SYNCHRONOUS relation, the feature *EmbedPrev* takes the value 1 since the entire relation is fully contained within the second argument of the preceding relation in the text.

3 Results and Evaluation

3.1 Classification accuracy

The model was evaluated by assessing the accuracy of its predictions against the unseen test set. The model achieved an overall accuracy of 86.6%, an improvement of 9.6% above baseline.[3] Table 2 shows the model accuracy for each relation type, together with the baseline performance based on the majority category for that type.

[3]During the preparation of the final version of this paper, a model was trained with an SVM using the same set of features, which resulted in a modest improvement in performance (87.3%). The ensuing discussion of results, however, will continue to pertain to the regression model.

Relation Type	Accuracy	Baseline
Asynchronous	91.7%	79.7%
Conjunction	84.5%	78.2%
Contrast	81.1%	65.0%
Instantiation	83.3%	82.5%
Reason	88.2%	68.3%
Restatement	95.2%	95.2%
Result	84.4%	76.9%
Synchronous	96.5%	92.9%
Total	**86.6%**	**77.0%**

Table 2: Classification Accuracy by Relation Type

The model achieved an improvement in accuracy across all relation types but one: RESTATEMENT relations, for which the baseline accuracy was already close to 100%. The greatest improvement in accuracy was seen for REASON relations, for which the model accuracy was 19.9% above baseline. We now discuss which of the factors in each of the feature classes turned out to be the most predictive.

3.2 Significant predictors

We trained the model on subsets of the features to investigate the predictive power of the different feature classes. The accuracy assessed against the test set is shown in Table 3.

Feature Class	# Features	Accuracy
Relation Level	4	80.4%
Argument Level	11	77.2%
Discourse Level	8	80.9%
Rel + Arg Levels	15	82.8%
Rel + Disc Levels	12	85.1%
Arg + Disc Levels	19	82.4%
All Features	23	86.6%

Table 3: Classification Accuracy by Feature Class

The classes of Relation-level and Discourse-level features each separately yielded significantly better performance over baseline (one-sided tests of proportion, z=2.50 and z=2.92, respectively; $p<0.01$ for both), whereas the Argument-level features alone performed only marginally better than baseline. However, all three classes of features are needed to attain the highest model performance.

Across all relation types, we found that the features relating to the discourse dependencies between a relation and its neighbors were the strongest and

most consistent predictors of whether that relation is explicit or implicit. A relation that is fully embedded within a single argument of an adjacent relation in the text (indicated by the features *EmbedPrev* and *EmbedNext*) has a much higher likelihood of being signaled explicitly. Conversely, a relation that fully contains another relation within one of its arguments (indicated by *Arg1Embed* and *Arg2Embed*) has a significantly higher likelihood of being implicit. The result is consistent with the embedded discourse dependency shown in (6), in which the implicit REASON relation fully contains an explicit SYNCHRONOUS relation within its second argument.

The model also found that the features which indicate whether a relation has shared arguments with either the preceding or following relations in the text (*Arg1isPrevArg2*, *Arg2isNextArg1*) are both predictors of an implicit outcome. In other words, if a clause in the text serves as the argument for two adjacent relations, then both of these relations are more likely to be realized implicitly.

The next most predictive feature was the connective used to signal the relation (*Connect*). This feature was a significant predictor for every relation type. Eliminating this feature from the final model reduces the overall accuracy by 2.5%. The other features in the model were less significantly predictive, and generally worked in the expected direction. Longer arguments (*Length1, Length2*) and the indication of a financial genre (*Financial*) were generally associated with predicted implicit outcomes, whereas the presence of supplementary material (*Supp1, Supp2*), a mismatch of attribution (*AttMismatch*), more 'content rich' arguments (*ContDensity1, ContDensity2*), and a pronoun appearing as the first word of the second argument (*FirstA2Pron*) all tended to increase the odds in favor of a predicted explicit outcome.

The features indexing syntactic complexity (*NPSbj1* and *NPSbj2*) were found to be marginally predictive of an explicit outcome for most relation types, but the overall effect in the model was relatively small—resulting in only a 0.2% improvement—meaning that the level of performance reported on this task depends very little on the model having access to full syntactic parses. Somewhat unexpectedly, the factors indicating the semantic type of the previous relation in the text (*PrevSemType*) and whether or not this relation was explicitly signaled by a connective (*PrevForm*) were found not to be significant predictors. Our analysis of the training data confirmed the findings of Pitler et al. (2008) in that certain bigrams of coherence relation types are significantly more prevalent than others. However, the differences in the frequencies were evidently not sufficiently correlated with the explicit/implicit distinction as to make the type or form of the previous relation a significant feature in the model.

3.3 Error analysis

We analyzed a sample of cases incorrectly predicted by the model to see if there were any consistent traits. We focus our attention here on the CONTRAST relations, which is the type with the lowest model accuracy. The majority of these errors were cases where the model predicted that the relation would be explicit—the most likely outcome for a CONTRAST relation—whereas in the corpus the intended relation was signaled by linguistic cues other than an overt connective. For instance, the strong syntactic parallelism of the two arguments in (7), and the opposite polarity of the lexical items *delight* and *detriment*, combine to induce a contrastive relationship without the need for a connective.

(7) To the delight of some doctors, the bill dropped a plan passed by the Finance Committee. [IMPLICIT=*but*] To the detriment of many low-income people, efforts to boost Medicaid funding were also stricken. (WSJ2372)

Other ways that the contrast relation is signaled implicitly include contrasting temporal modifiers (*It wasn't so long ago X. Now, Y*), repetition of the predicate in the argument (*. . . it could only happen once. . . . it's happening again*), or even by the use of punctuation such as a semicolon. Previous work (Sporleder and Lascarides, 2008; Lin et al., 2009) has sought to make use of such cues to identify and classify implicit relations in the text. The results of this brief error analysis suggest that such indirect cues could also be useful factors in determining whether to choose to use a connective for a given relation type when generating text.

4 Judgment Study

The system described in the last section outperformed a baseline majority-category classifier on the task of deciding whether a relation should be made explicit or left implicit. This result might be considered surprising, for two reasons that we have previously discussed. First, the system was able to make this improvement using relatively shallow features extracted from the text, without access to the richer types of contextual information and world knowledge required for establishing coherence relations during actual discourse comprehension. Second, the data suggest that the appropriateness of including a connective is not as cut-and-dried as a binary classification task may suggest, but is instead gradient, with many cases for which the inclusion of a connective appears to be optional. Obviously, the PDTB does not avail us of the opportunity to evaluate this gradience directly (or even use a 3-way required/optional/redundant distinction), since the producer of the actual text samples in the corpus had to ultimately decide whether or not to use a connective. The apparent optionality of many examples thus puts limits on how well we can expect a system to perform, since there is no way to reliably predict cases in which the decision is made arbitrarily.

This observation leads us to ask how well humans perform on this same task. Do they make highly accurate predictions, or does optionality limit their performance? In order to shed light on this question, we carried out an experiment to see how consistently humans choose to use lexical connectives to signal intended coherence relations between clauses.

4.1 Methodology

We selected a balanced sample of 100 clause-pair tokens from the test set, reflecting the distribution of the different major relation types (six relations were represented in the sample). This sample comprised 44 explicit and 56 implicit tokens, consistent with the distribution in the overall corpus. The experimental stimulus for each item consisted of two versions of the same clause pair, one including a connective between the clauses, and the other without. For relations that were realized explicitly in the corpus, as in (8a), the alternative implicit stimulus omitted the connective and showed the second argument as a separate sentence, as in (8b).

(8) a. Mr. Nesbit also said the FDA has asked Bolar Pharmaceutical Co. to recall at the retail level its urinary tract antibiotic, **but** so far the company hasn't complied with that request.

b. Mr. Nesbit also said the FDA has asked Bolar Pharmaceutical Co. to recall at the retail level its urinary tract antibiotic. So far the company hasn't complied with that request.

For the implicit relations, the alternative explicit stimulus for the experiment used the connective annotated in the PDTB as the one being most appropriate. For each item, a short passage was created including the preceding and following sentences in the text to serve as context. The relative ordering of the presentation of the explicit and implicit forms was randomized, without regard to the actual corpus outcome for that stimulus.

Using Amazon's Mechanical Turk, judges were presented with the two passages for each item. They were told to assume that the passages had the same intended meaning, and were asked to judge which of the two sounded more natural. We collected 30 responses for each item.[4]

4.2 Results

We classified each experimental item as either explicit or implicit, based on the majority response of the judges. Using this classification, the judges' responses matched the actual outcomes in 68 of the 100 cases.[5] The distribution of correctly-judged items across relation types is shown in Table 4. The judgments for REASON relations most closely matched the corpus outcomes, with 9 out of the 12 explicit tokens and all 6 implicit tokens in the cor-

[4]The data from a small number of judges were discarded due to an unreasonably fast response time or because their judgments showed a unanimous preference across every experimental item. This left a total of 2,925 judgments over the 100 experimental items, from 113 different judges.

[5]Using the majority response of judges for each item to measure classification accuracy is consistent with the statistical model, whereby probabilities are rounded up or down to arrive at a binary classification. If accuracy is instead calculated in terms of average correctness over the individual responses, performance drops to 60.4%.

pus correctly identified by the judges. The lowest scoring relation type was CONTRAST, for which 9 of the 10 explicit tokens were judged correctly but only 4 out of the 11 implicit tokens were correctly identified.

Relation Type	Items	Correct	Accuracy
Conjunction	22	15	68.2%
Contrast	21	13	61.9%
Instantiation	9	6	66.7%
Reason	18	15	83.3%
Restatement	16	9	62.5%
Result	14	10	64.2%
Total	100	68	68.0%

Table 4: Results of Mechanical Turk study

There were hence 32 experimental items for which the majority response by the judges did not match the actual corpus outcome. In two-thirds (21) of these cases, the judges indicated a preference for a connective when the relation in the corpus was implicit. These mismatches occurred across the range of relation types. This suggests that the judges tended to err on the side on inserting a connective, even when it may not have been strictly necessary. While the reason for this is not clear, one possibility is that the texts reflected the genre and the highly-prescribed editing guidelines for the newspaper articles that comprise the corpus, under which unnecessary or redundant words are excised. Without such pressures to edit the copy down to a minimal form, the judges may have preferred to see the relations signaled explicitly in cases in which either decision would result in a felicitous passage.

In the remaining 11 cases, for which the relations in the corpus were explicitly signaled with a connective, the judges on average indicated a preference to leave the relation implicit. Interestingly, all of these cases were either CONJUNCTION or CONTRAST relations, semantic types which are usually signaled explicitly with a connective. We inspected these cases to ascertain why judges may have preferred an outcome opposite to that actually seen in the text. We found that all 7 of the CONTRAST mismatches were instances where the second argument of the relation in the corpus was a sentence beginning with the coordinating conjunction *but*, as in (9).

(9) At those levels stocks are set up to be hammered by index arbitragers. **But** nobody knows at what level the futures and stocks will open today. (WSJ2300)

Similarly, three of the mismatched CONJUNCTION relations had a sentential second argument beginning with the conjunction *and*. The responses of the judges to these cases may simply reflect a dispreference for sentence-initial conjunctions, a practice which is frowned upon in prescriptive grammar books, but apparently allowed by the Wall Street Journal style sheet.

For this sample of 100 relations, the model achieves a classification accuracy of 84%. This may seem at first blush to be an odd result, since it appears that the model is surpassing human performance. As we have suggested, however, this could be the result of our experimental judges having different preferences than the writers and editors at the Wall Street Journal for cases in which connective placement is truly optional. We therefore sought to evaluate the effect of optionality on these results.

If inaccurate predictions are associated with optionality of connective use, we might expect that both human judges and the classification model would be less certain about their categorizations of these examples than for the cases that were correctly classified. This was indeed the case. First, there was a significant difference in the variability of judges' responses between items that were incorrectly classified and those that were correct (66% vs. 73%, respectively; two-sample t test: t=2.60, df=73, $p<0.02$). Thus, as a group the judges were less sure of themselves in those cases in which they incorrectly decided to use or omit the connective, suggesting that either option may have been acceptable. Second, we analyzed the levels of confidence our model had for its judgments on correctly and incorrectly categorized cases, measured in terms of the probability of the predicted outcome assigned by the model. The analysis revealed that the average model confidence for the relations that were incorrectly classified was significantly lower than the average model confidence for the correctly-classified items (71% vs. 88%, respectively; t=5.65, df=25, $p<0.001$). Taken together, these results are consistent with the idea that, at least for a significant portion of the data, the incorrect judgments made by

both the judges and the model may have occurred on passages for which either including or omitting the connective would have been acceptable.

5 Conclusion

We have presented a model that predicts whether the coherence relation holding between two clauses is marked explicitly with a lexical connective or left implicit. Whereas there is reason to think that an author's decision to use a connective is in part influenced by properties of the extra-linguistic context that are inaccessible to NLP systems (such as semantics and world knowledge), we find that relatively simple linguistic features derivable from the clauses and from local discourse dependencies can be exploited to reach a level of performance significantly greater than that achieved by a baseline. The variability in the judgments of native speakers when presented with these data suggests that the use of a connective is in many cases simply optional; in such cases the decision may reflect lower-level stylistic choices on the part of the author. This in turn indicates that there may be an inherent upper bound to the performance of computational systems on this task.

Acknowledgments

We thank Roger Levy for useful discussions about this work and three anonymous reviewers for their helpful feedback.

References

Nicholas Asher and Alex Lascarides. 2003. *Logics of conversation*. Cambridge University Press.

Fatemeh Torabi Asr and Vera Demberg. 2012a. Implicitness of discourse relations. In *Proceedings of the 24th International Conference on Computational Linguistics*, pages 2669–2684.

Fatemeh Torabi Asr and Vera Demberg. 2012b. Measuring the strength of linguistic cues for discourse relations. In *Proceedings of the Workshop on Advances in Discourse Analysis and its Computational Aspects (ADACA)*, pages 33–42.

Michael Elhadad and Kathleen R McKeown. 1990. Generating connectives. In *Proceedings of the 13th Conference on Computational Linguistics-Volume 3*, pages 97–101.

Jerry R Hobbs. 1979. Coherence and coreference. *Cognitive science*, 3(1):67–90.

Andrew Kehler. 2002. *Coherence, reference, and the theory of grammar*. CSLI Publications, Stanford.

Ziheng Lin, Min-Yen Kan, and Hwee Tou Ng. 2009. Recognizing implicit discourse relations in the Penn Discourse Treebank. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1*, pages 343–351.

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

Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Emily Pitler, Mridhula Raghupathy, Hena Mehta, Ani Nenkova, Alan Lee, and Aravind K Joshi. 2008. Easily identifiable discourse relations. In *Proceedings of the 22nd International Conference on Computational Linguistics*.

Emily Pitler, Annie Louis, and Ani Nenkova. 2009. Automatic sense prediction for implicit discourse relations in text. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2*, pages 683–691.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse Treebank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*.

Caroline Sporleder and Alex Lascarides. 2008. Using automatically labelled examples to classify rhetorical relations: An assessment. *Natural Language Engineering*, 14(3):369–416.

Zhi-Min Zhou, Yu Xu, Zheng-Yu Niu, Man Lan, Jian Su, and Chew Lim Tan. 2010. Predicting discourse connectives for implicit discourse relation recognition. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 1507–1514.