Modeling the Usage of Discourse Connectives as Rational Speech Acts

Frances Yung
Nara Institute of Science and Technology
8916-5 Takayama, Ikoma,
Nara, 630-0101, Japan
pikyufrances-y@is.naist.jp

Kevin Duh
John Hopkins University
810 Wyman Park Drive,
Baltimore, MD 21211-2840, USA
kevinduh@cs.jhu.edu

Taku Komura
University of Edinburgh
10 Crichton Street,
Edinburgh, EH8 9AB, United Kingdom
tkomura@inf.ed.ac.uk

Yuji Matsumoto
Nara Institute of Science and Technology
8916-5 Takayama, Ikoma,
Nara, 630-0101, Japan
matsu@is.naist.jp

Abstract

Discourse relations can either be implicit or explicitly expressed by markers, such as *'therefore'* and *'but'*. How a speaker makes this choice is a question that is not well understood. We propose a psycholinguistic model that predicts whether a speaker will produce an explicit marker given the discourse relation s/he wishes to express. Based on the framework of the Rational Speech Acts model, we quantify the utility of producing a marker based on the information-theoretic measure of surprisal, the cost of production, and a bias to maintain uniform information density throughout the utterance. Experiments based on the Penn Discourse Treebank show that our approach outperforms state-of-the-art approaches, while giving an explanatory account of the speaker's choice.

1 Introduction

Speakers or authors[1] produce informative utterances, such that the listeners or readers can understand his/her message. Grice's *Maxim of Quantity* states that human speakers communicate by being as informative as required, but no more (Grice, 1975). If a speaker always tries to provide as much information as possible, the resulting utterance could become excessively long and tedious. Such utterance is not only effort consuming for the speaker to produce, but also contains redundant information that is not necessary for the listener.

In this work, we model how speakers plan the presentation of discourse structure optimally in terms of informativeness. Specifically, we propose a model that predicts whether the speaker will use or omit a discourse connective, given the sense of discourse relation s/he wants to convey.

Discourse relations are relations between unit of texts (known as *arguments*) that make a document coherent. These relations can be marked in the surface text or inferred by the readers, as shown in the below examples.

1. It was a great movie, **but** I did not like it.
2. It was a great movie, **therefore** I liked it.
3. It was a great movie. I liked it.

The word *'but'* indicates a *Concession* relation in Example (1), and *'therefore'* indicates a *Result* relation in Example (2). We call 'but' and 'therefore' *explicit* discourse connectives (DCs). In Example (3), DCs are absent but a *Result* relation can be inferred. We say the DC is implicit in this case.

Explicit DCs are highly informative cues to identify discourse relations (Pitler et al., 2008) while implicit DCs are more ambiguous. For example, *'I liked it'* can also be read as a *Justification* for the first sentence in Example (3).

Marking a discourse relation or not is subject to ambiguity and redundancy. On one hand, using an explicit DC avoids ambiguity. For example, if the DC 'but' is omitted in Example (1), readers may have problems in inferring the *Concession* sense. On the other hand, if the intended discourse sense is highly predictable, it is verbose or redundant to insert an explicit DC in the utterance, such as the DC *'therefore'* in Example (2).

[1] *'Speakers'* and *'listeners'* are interchangeably used with *'authors'* and *'readers'* in this article

A model that predicts the markedness of discourse relations not only contributes to a better understanding of the human language production mechanism, but is also important in generating natural, humanlike texts and dialogues. In particular, the degree of markedness in discourse relations differs cross-lingually. Yung et al. (2015) analyze the manual alignments of explicit and implicit DCs in a Chinese-English translation corpus and find that 30% of implicit DCs in Chinese are translated to explicit DCs in English. It remains a challenge for machine translation systems to explicitate or implicitate discourse relations in the source texts as human translators do (Becher, 2011; Meyer and Webber, 2013; Zuffery and Cartoni, 2014; Hoek and Zufferey, 2015; Hoek et al., 2015), since the markedness of the translation is subject to the discourse planning of the target text.

In order to explain how human speakers choose the optimal level of markedness in his utterance, we model how speakers rationally balance between ambiguity and redundancy. In particular, we use the Rational Speech Acts (RSA) model (Frank and Goodman, 2012) to predict how speakers reason about the ambiguity of an utterance. In addition, we model how speakers adjust the redundancy of the utterance following the Uniform Information Density (UID) principle (Levy and Jaeger, 2006).

We apply the framework to predict whether an explicit or implicit DC is used in corpus data, given the two arguments of the discourse relations and the discourse sense to be conveyed. Our model not only achieves higher accuracy comparing with previous work (Patterson and Kehler, 2013), but also provides an interpretable account of various cognitive factors behind the predicted decision.

We start by a review of related work in Section 2, followed by the descriptions of our model in Section 3 and experiments in Section 4.

2 Related work

We first provide background information on RSA and UID, which are used in our proposed method. It is followed by introduction of previous work about prediction of DC markedness in corpus data.

2.1 Rational Speech Acts model

The RSA model (Frank and Goodman, 2012) is a variation of the game-theoretic approach in pragmatics (Jäger, 2012). It explains the communicative reasoning of a speaker and a listener in terms of Bayesian probabilities.

A rational listener assumes the utterance s/he hears contains the optimal amount of information. S/he predicts the intended message of a speaker by Bayesian inference (Equation 1).

$$P_{listener}(s|w,C) \propto P_{speaker}(w|s,C)P(s) \quad (1)$$

where w is the *utterance* produced by the speaker; s is the *message* of an utterance; and C is the *context*. $P_{speaker}(w|s,C)$ represents the *listener's predicted* speaker's model, and $P(s)$ represents the *salience* of the message, which is shared knowledge between the speaker and listener.

A rational speaker chooses an utterance by softmax optimizing the expected *utility* $(U(w;s,C))$ of the utterance (Equation 2).

$$P_{speaker}(w|s,C) \propto e^{\alpha \cdot U(w;s,C)} \quad (2)$$

α is the decision noise parameter, which is set to 1 to represent a rational speaker [2]. S/He emulates the listener's interpretation and chooses an utterance s/he believes to be informative. Also, an utterance that is easy to produce is preferred.

Utility is thus defined as the *informativeness* $(I(s;w,C))$ of the utterance, deducted by the cost $(D(w))$ to produce it (Equation 3).

$$U(w;s,C) = I(s;w,C) - D(w) \quad (3)$$

Since utterances that are unconventional and surprising are less useful, *Informativeness* is quantified as the *negative surprisal* of the utterance with respect to the message to be conveyed (Equation 4).

$$I(s;w,C) = \ln P(s|w,C) \quad (4)$$

The RSA model has successfully simulated results of psycholinguistic experiments concerning different aspects of human communication, such as scalar implicature and referential expressions (Frank and Goodman, 2012; Goodman and Stuhlmüller, 2013; Bergen et al., 2014; Kao et al., 2014; Potts et al., 2015). Besides experimental data, Orita et al.(2015) applies RSA model to predict the choice of referring expressions in corpus data and Monroe and Potts (2015) optimizes a

[2] $\alpha = 0$ means the decision is totally unrelated to pragmatic reasoning. $\alpha = 1$ represents the Luce's choice axiom (Frank and Goodman, 2012), i.e. a rational decision without bias. $\alpha > 1$ suggests biased choices.

classifier based on RSA by inducing the semantic lexicon from a training corpus. These works focus on the pragmatic use of language, where the informativeness and lexicon of an utterance largely depends on the context (e.g. *'Red'* is not *valid* to be used to refer to a *blue* ball).

In this work, we apply RSA to predict the usage of DCs, which is more universal across different contexts (i.e. A DC can be used or dropped given various discourse senses and contexts). Our model is built upon the *speaker's model* of RSA to predict speaker's choice of explicit or implicit DCs.

2.2 Uniform Information Density

The UID principle views language communication as a form of information transmission through a noisy channel and a constant rate of information flow is optimal according to Shannon's Information Theory (Levy and Jaeger, 2006; Genzel and Charniak, 2002; Shannon, 1948). It states that speakers structure utterances by optimizing *information density*, which is the quantity of information (measured by *surprisal*[3]) transmitted per unit of utterance, such as word.

Information density rises when the utterance is 'surprising' and drops when an utterance is highly predictable. To smooth the peaks and troughs, speakers adjust the ambiguity of an utterance by including or reducing linguistic markers.

Following the UID principle, linguistic choices made by speakers are predicted more accurately by incorporating an *information density predictor* on top of other constraints. The predictor measures how easily a candidate utterance can be predicted and the speaker adjusts information density based on the expected predictability.

UID is applied to explain a variety of speaker's options, such as phonetic (Aylett and Turk, 2004), morphological (Frank and Jaeger, 2008) and syntactic (Jaeger, 2010) reductions, and also referring expressions (Tily and Piantadosi, 2009).

2.3 Explicit vs. Implicit DCs

The choice of discourse marking strategies has been studied in earlier works as a subtask for natural language generation (Scott and de Souza, 1990; Moser and Moore, 1995; Grote and Stede, 1998; Soria and Ferrari, 1998; Allbritton and Moore, 1999). In the absence of large-scale resources, investigations are based on manually derived rules and lexicons or psycholinguistic experiments.

More recently, Asr and Demberg (2012) presents an analysis of the PDTB, showing that 'causal' and 'continuous' senses are more often implicit, or marked by less specific DCs. Indeed these senses are presupposed by listeners according to linguistics theories (Segal et al., 1991; Murray, 1997; Levinson, 2000; Sanders, 2005; Kuperberg et al., 2011). On the other hand, Asr and Demberg (2015) finds that DCs are more often dropped for the discourse relation *Chosen Alternative* (the relation typically signalled by the DC 'instead'), if the context contains negation words, which are identified cues for this relation. Similarly, contextual difference in explicit and implicit discourse relations are reported in attempts to train implicit DC classifiers based on explicit DC instances (Sporleder and Lascarides, 2008; Webber, 2009).

Asr and Demberg (2012; 2015) attribute the corpus statistics to the UID hypothesis, which explains that expected, predictable relations are more likely to be conveyed implicitly, and thus more ambiguously, to maintain steady information flow. However, there are explicit 'causal' and 'continuous' relations and some *Chosen Alternative* are marked even *argument 1* is negated. Although markedness measures are proposed to rate the implicitness of a relation sense (Asr and Demberg, 2013; Jin and de Marneffe, 2015), these measures only quantify the general markedness of the sense in the data, but not the speaker's choice for each particular instance. In contrast, this work specifically measures the predictability of a given relation; generalizes the approach to all discourse senses instead of particular senses or cues; and combines the markedness preference with other language production factors, in order to model each instance of relation.

Patterson and Kehler (2013) is the only study we are aware of that predicts the choice of explicit or implicit DCs of each instance of relation. They argue that while the decision is related to the ease to infer the relation, it may also depend on other stylistic or textual factors. A classifier is trained to predict whether a *candidate DC* (i.e. the DC that actually occurs in the text as an explicit DC, or annotated as an implicit DC) is actually present, given the sense of the discourse relation and the arguments. Relatively shallow linguistic features are used, such as whether the relations are em-

[3] This is opposite to *'informativeness'* in RSA, which is defined by *negative surprisal* (Equation 4).

bedded or shared, the previous discourse relation, argument lengths, and content word ratios. The classifier is trained and tested on a subset of relations from the PDTB, after screening away infrequent senses and DCs. An overall high classification accuracy is achieved. Relation-level and discourse-level features are found to be more useful than argument-level features.

However, this work does not target at explaining why an utterance is preferred by the speaker. The focus is a data-driven approach that replicates the occurrence of DCs in the corpus data. Our work differs in that we model the option of markedness from the viewpoint of human language production, explaining the factors behind the speaker's choice. For example, we do not make use of the *candidate DC* as a feature, since it is the result of the speaker's choice, if an explicit DC is preferred. Nonetheless, our model achieves higher accuracy when evaluated on the same test set.

3 The markedness model

Our model is based on the speaker's model of RSA. We first explain how we adapt the RSA model to discourse presentation, followed by the details of each component.

3.1 RSA for discourse relation presentation

According to Equation (2), the probability for a speaker to use utterance w to convey his intended message s in context C is:

$$P(w|s,C) = \frac{e^{U(w;s,C)}}{\sum_{w' \in W} e^{U(w';s,C)}} \tag{5}$$

In the case of discourse connectives, the utterance w comes from the set W = {(exp)licit, (imp)licit}, if both explicit and implicit DCs are grammatically valid to convey s, the sense of discourse relation. Our model thus predicts speaker's choice of DCs based on the following two probabilities:

$$\begin{aligned} P(exp|s,C) &= \frac{e^{U(exp;s,C)}}{e^{U(exp;s,C)} + e^{U(imp;s,C)}} \\ P(imp|s,C) &= \frac{e^{U(imp;s,C)}}{e^{U(exp;s,C)} + e^{U(imp;s,C)}} \end{aligned} \tag{6}$$

According to Equation (3), the *utility* U of an explicit DC equals to its *informativeness* I deducted by production cost D.

$$U(exp;s,C) = I(s;exp,C) - D(exp) \tag{7}$$

$I(s;exp,C)$ is the informativeness of using an explicit DC to present the sense s in discourse-level context C. Each discourse sense has its salience within the discourse context. It means C is also informative, but we want to quantify the informativeness of the DC only. Therefore, we define $I(s;exp,C)$ by the difference between the informativess of 'the explicit DC in context C' and the informativeness of 'context C', which are quantified by negative *surprisal*.

$$I(s;exp,C) = \ln P(s|exp,C) - \ln P(s|C) \tag{8}$$

High $I(s;exp,C)$ means it is informative and not surprising to use an explicit DC for this sense. $P(s|exp,C)$ and $P(s|C)$ are extracted from corpus data. Details are explained in Subsection 3.2.

The principle of UID is incorporated into the RSA model as a bias on the utility of the DCs. A discourse relation is presented not only by the DCs but also the arguments, and the amount of discourse information of the whole utterance (DC + arguments) is fixed. According to UID, information should be transmitted uniformly across the utterance. If the arguments has much information about the sense, the sense is predictable from the arguments and thus the surprisal is small. The information density drops and has to be smoothed by using a more ambiguous, less predictable utterance, which can be achieved by reduction of a DC (Asr and Demberg, 2015).

Therefore, according to UID, an implicit DC is preferred if the arguments are informative. We thus raise the utility of an implicit DC by defining the probability for a speaker to choose an implicit DC to be proportional to the *sum of the the utilities* of a *null* DC and the arguments ($args$)[4].

$$e^{U(imp;s,C)} = e^{U(null;s,C)} + e^{U(args;s,C)} \tag{9}$$

$$U(null;s,C) = I(s;null,C) - D(null) \tag{10}$$

$$U(args;s,C) = I(s;arg,C) - D(args) \tag{11}$$

The amount of information that the null DC provides for the discourse relation is defined similarly as in Equation (8):

$$I(s;null,C) = \ln P(s|null,C) - \ln P(s|C) \tag{12}$$

[4] In turn, an explicit DC is preferred if the arguments are not informative. We could also penalize the utility of an explicit DC by the argument utility, but the result will be the same since the decision is based on Equation 13.

On the other hand, the informativeness of arguments, $I(s; arg, C)$ is quantified by *negative surprisal* in RSA. However, arguments are clauses and sentences. It is not applicable to extract $P(s|args, C)$ from the corpus. We thus approximate $I(s; arg, C)$ by *the confidence of a discourse parser in predicting discourse senses from the arguments*. Details will be explained in Section 3.3.

Lastly, various psycholinguistically motivated measures are explored to approximate the prodcution cost $D(exp)$ in Subsection 3.4. In contrast, no effort is required to produce a *null* DC. Also, we assume that the arguments have been produced to convey other information irrespective of their discourse informativeness, so no extra effort is needed. Therefore, $D(null)$ and $D(args)$ both equal 0.

To summarize, the model predicts that the speaker will use an explicit DC if:

$$e^{U(exp;s,C)} > e^{U(null;s,C)} + e^{U(args;s,C)} \quad (13)$$

and that s/he will use an implicit DC otherwise.

3.2 Informativeness of DCs

This section explains how we estimate the informativeness in Equations (8) and (12). In discourse production, the utterance lexicon, $W = \{exp, imp\}$ in Equation (5), and the set of speaker's intended messages (all possible discourse relation senses) are always *valid*[5]. Thus $P(s|C)$, $P(s|exp, C)$, and $P(s|null, C)$ are universal distributions and can be extracted from corpus data based on the co-occurrences of senses, DCs, and contexts. We extract these empirical distributions from the training portion of the corpus.

We define context C as the surrounding discourse relations. Specifically, the discourse contexts (and their abbreviation in Table 2) are: the full discourse sense annotated in PDTB (S), the 4-way top level sense (TS), the form of discourse presentation (F) such as 'explicit' or 'implicit'[6], and the pair of sense and form (SF or TSF). The contexts are taken from window sizes of 1 to 2: previous one (10) , next one (01), previous two (20), next two (02), previous one paired with next one (11). We hypothesize that the speaker also thinks ahead the coming discourse structures when planning the current ones. Various discourse contexts are compared in the experiment.

3.3 Informativeness of arguments

$I(s; arg, C)$ in Equation (11) refers to the amount of information in the arguments that contributes to the interpretation of the discourse sense. According to UID, information density drops when the discourse sense is predictable from the arguments alone, and an implicit DC is preferred.

Presence of features in the arguments that signal a particular sense makes the sense more predictable, and thus promote the reduction of a DC. For example, the DC *'instead'* is less used to present the *Chosen Alternative* sense if the first argument is negated (Asr and Demberg, 2015).

Generalizing this idea to capture various cues in the arguments for various senses, we approximate $I(s; arg, C)$ by the confidence of an automatic discourse parser in predicting the discourse sense. An implicit relation parser uses various features in the arguments to identify the implicit relation sense (Pitler et al., 2009; Lin et al., 2009; Park and Cardi, 2012; Rutherford and Xue, 2014). If the arguments contain much informative features, the parser will predict the sense more confidently.

We propose two methods, for comparison, to measure the confidence of the parser prediction. A confident prediction means the parser will assign a high probability to the one output sense. Therefore, we use the *negative surprisal* of the estimated probability P_p of the parser output sense s_{output} (Equation 14) to approximate $I(s; arg, C)$.

$$I(s; arg, C) \approx w_a \cdot \ln P_p(s_{output}) \quad (14)$$

At the same time, the probability distribution of all senses is less uniform if one sense is assigned a high probability. We thus alternatively approximate $I(s; arg, C)$ by the *negative entropy* of the probability distribution estimated by the parser (Equation 15)[7].

$$I(s; arg, C) \approx w_a \sum_{s_p \in O} P_p(s_p) \log P_p(s_p) \quad (15)$$

where O is the set of senses defined in the parser and w_a is a positive weight tuned on the dev set.

[5] In case of referring expressions, for example, the lists of referents and grammatically correct pronouns differ case by case, e.g. *'she'* is not a *valid* pronoun for a male.

[6] We use the 5 forms of discourse presentation defined in the PDTB: explicit DC, implicit DC, alternative lexicalization, entity relation and 'no relation'.

[7] Note that we use information-theoretic measures to approximate $I(s; arg, C)$, but these approximations are not related to the formulation of RSA nor UID.

We measure the *general informativeness* of the arguments to imply *any* discourse senses, so s_{output} does not necessarily equal s.

We employ the implicit sense classifier from the winning parser of shared task 2015 (Wang and Lan, 2015), which is designed to identify a subset of 14 implicit senses plus the *entity relation*. The two arguments of a relation instance, which can actually be explicit or implicit, are passed to the implicit DC classifier and $I(s; arg, C)$ is approximated based on the output probabilities [8]. Although the performance of this state-of-the-art implicit DC classifier is still unsatisfactory (34.45% on PDTB Section 23[9]) , our method only makes use of the probability estimation of the prediction[10].

Our motivation of using the implicit DC classifier is based on the hypothesis that the classifier can better predict the sense of relations that are actually implicit, than those that are actually explicit, since more features in the arguments are identifiable. In fact, it is the case. The classification accuracy of the originally explicit relations is significantly lower. This supports our motivation to use the parser estimation as an information density predictor.

3.4 Cost function

The cost function $D(exp)$ models speaker's effort required to produce an explicit DC for the intended discourse sense. We propose 5 versions of the cost function that are inspired by existing psycholinguistic findings.

Mean DC length: Production cost intuitively increases with word length. We define the mean DC length of a discourse relation as the mean word length of all valid DCs for that sense, normalized by the average word length of all DCs. A lexicon of possible DC per each discourse sense is derived from the whole corpus. For multi-word DCs, a white space is simply counted as one character. We do not use the length of the *candidate DC* (refer to Section 2.3), because we view that speakers first decide to use an explicit DC or not, then decide which DC best expresses the relation.

DC/arg2 ratio: Similarly, we use the mean word count normalized by the word count of *argument 2* as another version of cost function.

Prime frequency: Structural priming refers to the tendency for human to process a linguistic construction (the target) more easily if the construction is used before. In terms of language production, a speaker tends to repeat a previous construction (the prime) since it consumes less effort than to generate an alternative construction. We use the reciprocal of the count of primes (any explicit DC occurring before the current position) as the production cost, since the strength of priming effect is known to be increasing with the frequency of the primes (Levelt and Kelter, 1982; Bock, 1986; Smith and Wheeldon, 2001).

Prime distance: We also use the prime-target distance, normalized by the length of the article, as another version of the production cost. Psycholinguistic findings suggest that the priming effect is more subtly affected by the prime-target distance (Gries, 2005; Bock et al., 2007; Jaeger and Snider, 2008).

Distance from start: We use the relative position of the relation within the article as the production cost. We hypothesize that more effort is needed as the production proceeds.

The range of values of the cost function depends on the cost definition. We thus adjust the values with a constant weight w_c that is tuned on the dev set in the experiments:

$$D(exp) = w_c \cdot cost(exp) \qquad (16)$$

4 Experiment

We apply the model to simulate speaker's choice of explicit or implicit DC for discourse relations in the PDTB corpus. The aim of the experiment is to answer two questions: (1) Does the model explain the factors affecting speaker's choice of DC markedness? If the hypotheses of the model is appropriate, each component in the model should

[8] The implicit DC classifier is trained by Naïve Bayes based on features including syntactic features, polarity, immediately preceding DC, and Brown cluster pairs. Syntactic features are based on automatic parsing using Stanford CoreNLP (Manning et al., 2014). The parser is trained on the same sections of the PDTB as the training set used in our experiment.

[9] http://www.cs.brandeis.edu/~clp/conll15st/results.html

[10] We use the parser's probability estimates as is; conceivably it may be improved by an additional probabilistic calibration step (Nguyen and O'Connor, 2015).

contribute to the prediction accuracy. (2) How does the prediction performance compare with the state-of-the-art, i.e. Patterson and Kelher (2013)?

We first describe the details of the data we use in the experiments.

4.1 Data: The Penn Discourse Treebank

The Penn Discourse Treebank (PDTB) is the largest available discourse-annotated corpus in English (Prasad et al., 2008). The text are news articles collected from the Wall Street Journals. Below are 3 examples of the annotation.

1. The OTC market has only a handful of takeover-related stocks. **But** (Explicit;*Comparison-Contrast*) they fell sharply. *(WSJ2379)*

2. Japan's Finance Ministry had set up mechanisms ... to give market operators the authority to suspend trading in futures at any time. (Implicit: **but**; *Comparison*) Maybe it wasn't enough. *(WSJ0097)*

3. **Before** (Explicit; *Temporal-Asynchronous-Precedence*) becoming a consultant in 1974, Mr. Achenbaum was a senior executive at J. Walter Thompson Co..*(WSJ0295)*

Explicit DCs are labelled with relation senses (Example 1). If an explicit DC is absent *between two sentences* within the same paragraph and an implicit relation can be inferred, a candidate DC and the relation sense are annotated (Example 2).

Our model is based on the assumption that $W = \{explicit, implicit\}$ for all relations, yet it is notable that *intra-sentential implicit* DCs are **not** annotated in the PDTB (Prasad et al., 2014). We thus exclude intra-sentential samples, such that $W = \{explicit, implicit\}$ is always true and free of grammatical constraints. Also, as a result of the annotation procedure, implicit DCs always occur *in between 2 arguments* in their original order, i.e. Arg1-DC-Arg2. To preserve the original order of the discourse arguments, which is also part of the communicative structure intended by the speaker but out of the scope of this model, we only use samples in the Arg1-DC-Arg2 order. For example, Example (3) is excluded from our training data. Finally, annotations of other forms of discourse relations, such as entity relations and attributions, are also excluded.

The screened data set contains 5,201 explicit and 16,049 implicit relations[11]. Sections 2-22 are used as the training set, from which probability distributions are extracted. For easier comparison with previous work, we select the dev set (sections 0-1) and test set (sections 23-24) in the same way as in Patterson and Kehler (2013), where only relations of infrequent DCs and senses are removed. The resulting dev and test sets contain 1720 and 1878 relations respectively. Samples not included in our screened dataset are classified as *explicit* by default.

	sense	exp	imp
1	Expansion.Conjunction	1,380	3,314
2	Comparison.Contrast	1,283	1,200
3	Expansion.Restatement.Specification	75	2,406
4	Contingency.Cause.Reason	28	2,295
5	Contingency.Cause.Result	269	1,649
6	Expansion.Instantiation	119	1,383
7	Comparison.Contrast.Juxtaposition	507	672
8	Comparison.Concession.Contra-expectation	475	179
9	Temporal.Asynchronous.Precedence	117	479
10	Expansion.List	84	374
...	...	...	...
17	Expansion.Conjunction.–Temporal.Synchrony	74	114
...	...	...	...
50	Contingency.Pragmatic cause.Justification #Expansion.Instantiation	0	6
...	...	...	...
122	Contingency	0	1
Total		**5,201**	**16,049**

Table 1: Sense distribution of explicit and implicit DCs in screened data set.

Senses in the PDTB are defined in a hierarchy of 2 to 3 levels. Some relations have multiple senses. Up to 2 DCs can be annotated to an implicit relation and in turn each (implicit or explicit) DC can be labelled with up to 2 senses. Most existing works split a multi-sense sample into separated

[11] 4 cases of intra-sentential implicit relations, due to sentence splitting errors of the PTB (single sentences wrongly splitted into two), are removed.

	discourse	arg. info.	cost function	Dev: Sections 0-1			Test: Sections 23-24		
	context C	$e^{U(args;s,C)}$	$D(exp)$	accuracy	$F1_{exp}$	$F1_{imp}$	accuracy	$F1_{exp}$	$F1_{imp}$
BL	constant	0	0	.849	.872	.817	.854	.875	.823
SOA (Patterson and Kehler, 2013)				–	–	–	.866	–	–
(a)	F10	0	0	.855	.876	.826	.855	.876	.826
	SF10	0	0	.859	.877	.835	.855	.874	.829
	F20	0	0	.854	.875	.825	.854	.875	.825
	F11	0	0	.851	.872	.822	.854	.875	.825
	TS10	0	0	.852	.872	.822	.854	.875	.824
(b)	constant	surprisal	0	$.895^{++}$	.901	.887	.870	.881	.857
	constant	entropy	0	$.895^{++}$	.902	.888	.870	.881	.856
(c)	constant	0	mean DC length	$.894^{++}$	.897	.890	$.876^{+}$	.886	.865
	constant	0	DC/arg2 ratio	$.895^{++}$	.900	.889	.873	.882	.863
	constant	0	prime frequency	$.886^{+}$	.888	.885	.873	.882	.862
	constant	0	prime distance	$.892^{++}$	.902	.881	.875	.886	.862
	constant	0	distance from start	$.893^{++}$	.894	.892	$.877^{+}$	.879	.875
(d)	F10	entropy	DC/arg2 ratio	$\mathbf{.902}^{++}$	**.903**	**.901**	$.882^{+}$	.883	.881
	TSF01	surprisal	prime frequency	$.895^{++}$	.898	.892	$.889^{++*}$	**.893**	.885
	TS01	entropy	prime distance	$.895^{++}$	.900	.889	$\mathbf{.890}^{++*}$	.892	**.888**

Table 2: Accuracies and F1 scores of predicted DC markedness. The best values are bolded. $^{+}/^{++}$:significant improvement over baseline (**BL**) accuracy at $p < 0.05$ and $p < 0.001$ respectively; *:significant improvement over state-of-the-art (**SOA**) accuracy at $p < 0.03$ (by Pearson's χ^2 test) (refer to Section 3.2 for abbreviations of discourse context C.)

samples, each labelled with one of the senses. However, it is notable that the individual senses of a multi-sense relation are not disjoint[12] and *having multiple senses* is *part of the sense* (Asr and Demberg, 2013; Prasad et al., 2014). Multi-sense is an important factor of our DC production model: a speaker could have chosen an explicit DC for each sense, but if s/he has to express two senses at the same time, an implicit DC could be more usable. Therefore, we treat all combination of senses as *individual senses*, each containing 1 to 3 joint sense labels[13] This results in a total of 122 senses.

Table 1 is a summary of the distribution in descending order of frequency. In fact, joint multi-senses are not rare: the most frequent multi-sense is the 17th most frequent sense.

4.2 Results

We apply the *markedness* model to predict the speaker's choice of DC markedness on the dev and test sets. Table 2 shows the results under various settings, evaluated by accuracy and the harmonic mean of precision and recall for explicit and implicit relations respectively.

Row **BL** shows the results of the *markedness* model without the cost function and argument informativeness component, and with constant context C. We consider this setting as the baseline, in which the prediction is solely based on the distributions of $P(s|exp)$ and $P(s|imp)$. Considerably high accuracy is achieved, suggesting that the speaker's choice of markedness is strongly related to the intended discourse sense.

Row (a) shows the prediction results based on the distributions of $P(s|exp, C)$ and $P(s|imp, C)$, where C is the discourse context. The 5 best combinations of contexts and window sizes are shown. Refining the utility of DCs by these contextual constraints, in particular previous contexts, improves the classification accuracy, but the improvement is not significant. This suggests that speaker's choice of markedness not only depends on surrounding discourse relations but also other contextual factors.

Row (b) shows the contribution of the *argument informativeness* component, under constant dis-

[12] Similarly, certain level 2 senses, as in Example (2), are backoffed from level 3 senses due to annotator disagreement. This is also a kind of multi-sense.

[13] There is only 1 sample of 3 joint labels in our screened dataset.

course context and production cost. Classification accuracy increases (significantly for the dev set) when the usability of explicit DC is deducted by the estimated informativeness of the arguments, supporting the UID principle. Predictions based on the surprisal of the parser output sense and the entropy of the parser output distribution are similar. We also experiment by adjusting with the estimated *argument informativeness* only if the parser output sense is correct (matching at the top level sense). Similar improvement is observed.

Row (c) shows the contribution of the cost function, when discourse context is set as constant and argument informativeness is not considered. Adjusting the utility of explicit DCs by their production cost increases the classification accuracy most significantly. Among the various features to model production cost, 'DC length' and 'distance from start' features give the best results.

Row (d) shows the performance of predictions based on the 3 best combinations of components. The highest accuracies and F_1 scores are achieved for both explicit and implicit relations.

These results answer the first question of the experiment purpose: the proposed model explains the speaker's choice of DC markedness in terms of DC and argument informativeness, and production cost, while contextual discourse structure is a moderate constraint to the choice.

The answer to the second question is also positive. Significant improvement above the state-of-the-art (Row **SOA**) is achieved by the 2 best combinations (89.0%, 88.9% vs. 86.6%).

Lastly, we compare the results with a linear classifier trained on the *features* specified in the model, i.e. the discrete values of the intended sense and various discourse context definitions, and real values of various cost functions and argument informativeness estimates. Note that in the proposed model, the training data is used to derive the $P(s|exp, C)$ and $P(s|null, C)$ distributions only, while the linear classifier learns from the features and DC markedness of the training set[14]. The classifier achieves accuracy of 88.3% on the test set, which does not significantly outperform previous work. This suggests the advantage of the information-theoretic configuration of our model.

5 Conclusion

We present a language production model that predicts a speaker's choice of using an explicit DC or not given the discourse relation s/he wants to express. Our model gives an cognitive account of the speaker's choice and also outperforms previous work on the same task.

Our study shows that a speaker organizes the discourse structure by balancing the pro (informativeness) and con (production cost and redundancy) of using an explicit marker, although the option is a subtle preference in the absence of other grammatical constraints. Using an information-theoretic approach, our model tackles the option as a rational preference by the speaker, who wants to contribute to an informative speech act. Furthermore, we take a logical step forward to formalize the idea of the UID theory, that redundant explicit markers are avoided if the discourse relation is clear enough from the context.

As future work, we plan to improve the *markedness* model by making fuller use of the training data, such as learning a more expressive formulation of the context governing the choice of explicit or implicit DCs. We also plan to evaluate the effectiveness of the model in applications, such as natural language generation or machine translation tasks. On the other hand, as discourse presentation differs across genres (Webber, 2009) and mediums (Tonelli et al., 2010), the model can be applied to predict the explicitation of discourse relations from, for example, news articles to spoken dialogues. Another direction is to apply the RSA framework in the opposite direction - to build a *listener's model* that simulates a listener's recognition of a discourse sense given an utterance, as proposed in Yung et al.(2016).

Acknowledgments

We thank the anonymous reviewers for their valuable feedback on the previous versions of this paper.

References

David Allbritton and Johanna Moore. 1999. Discourse cues in narrative text: Using production to predict comprehension. In *AAAI Fall Symposium on Psychological Models of Communication in Collaborative Systems*.

[14] When extracting the argument informativeness features from the training set, using the automatic discourse parser, we penalize the parser estimates of the *implicit* samples by a constant ratio, since the discourse parser is also trained on these samples. We use LIBLINEAR (Fan et al., 2008) to build the classifiers.

Fatemeh Torabi Asr and Vera Demberg. 2012. Implicitness of discourse relations. In *COLING*, pages 2669–2684. Citeseer.

Fatemeh Torabi Asr and Vera Demberg. 2013. On the information conveyed by discourse markers. In *Proceedings of the Fourth Annual Workshop on Cognitive Modeling and Computational Linguistics*, pages 84–93.

Fatemeh Torabi Asr and Vera Demberg. 2015. Uniform information density at the level of discourse relations: Negation markers and discourse connective omission. *Proceedings of the International Conference on Computation Semantics*, pages 118–128.

Matthew Aylett and Alice Turk. 2004. The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, 47(1):31–56.

Viktor Becher. 2011. When and why do translators add connectives? a corpus-based study. *Target*, 23(1).

Leon Bergen, Roger Levy, and Noah D. Goodman. 2014. Pragmatic reasoning through semantic inference.

Kathryn Bock, Gary S Dell, Franklin Chang, and Kristine H Onishi. 2007. Persistent structural priming from language comprehension to language production. *Cognition*, 104(3):437–458.

J Kathryn Bock. 1986. Syntactic persistence in language production. *Cognitive psychology*, 18(3):355–387.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874.

Michael C. Frank and Noah D. Goodman. 2012. Predicting pragmatic reasoning in lanugage games. *Science*, 336(6084):998.

Austin Frank and T Florian Jaeger. 2008. Speaking rationally: Uniform information density as an optimal strategy for language production. *Proceedings of the Annual Meeting of the Cognitive Science Society*, pages 933–938.

Dmitriy Genzel and Eugene Charniak. 2002. Entropy rage constancy in text. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 199–206.

Noah D Goodman and Andreas Stuhlmüller. 2013. Knowledge and implicature: modeling language understanding as social cognition. *Topics in cognitive science*, 5(1):173–184.

H Paul Grice. 1975. Logic and conversation. *Syntax and Semantics*, 3:41–58.

Stefan Th Gries. 2005. Syntactic priming: A corpus-based approach. *Journal of psycholinguistic research*, 34(4):365–399.

Brigitte Grote and Manfred Stede. 1998. Discourse marker choice in sentence planning. In *Proceedings of the Ninth International Workshop on Natural Language Generation*, pages 128–137.

Jet Hoek and Sandrine Zufferey. 2015. Factors influencing the implicitation of discourse relations across languages. In *Proceedings 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (isa-11)*, pages 39–45. TiCC, Tilburg center for Cognition and Communication.

Jet Hoek, Jacqueline Evers-Vermeul, and Ted JM Sanders. 2015. The role of expectedness in the implicitation and explicitation of discourse relations. In *Proceedings of the Second Workshop on Discourse in Machine Translation (DiscoMT)*, pages 41–46. Association for Computational Linguistics.

T Florian Jaeger and Neal Snider. 2008. Implicit learning and syntactic persistence: Surprisal and cumulativity. In *The 30th Annual Meeting of the Cognitive Science Society (CogSci08)*, page 827.

T Florian Jaeger. 2010. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive psychology*, 61(1):23–62.

Gerhard Jäger. 2012. Game theory in semantics and pragmatics. In Claudia Maienborn, Klaus von Heusinger, and Paul Portner, editors, *Semantics: An International Handbook of Natural Language Meaning*, volume 3, pages 2487–2425. Mouton de Gruyter.

Lifeng Jin and Marie-Catherine de Marneffe. 2015. The overall markedness of discourse relations. Proceedings of the Conference on Empirical Methods on Natural Language Processing.

Justine T Kao, Jean Y Wu, Leon Bergen, and Noah D Goodman. 2014. Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences*, 111(33):12002–12007.

Gina R Kuperberg, Martin Paczynski, and Tali Ditman. 2011. Establishing causal coherence across sentences: An erp study. *Journal of Cognitive Neuroscience*, 23(5):1230–1246.

Willem JM Levelt and Stephanie Kelter. 1982. Surface form and memory in question answering. *Cognitive psychology*, 14(1):78–106.

Stephen C Levinson. 2000. *Presumptive meanings: The theory of generalized conversational implicature*. MIT Press.

Roger Levy and T. Florian Jaeger. 2006. Speakers optimize information density through syntactic reduction. *Advances in neural information processing systems*, (849-856).

Ziheng Lin, Minyen Kan, and Hwee Tou Ng. 2009. Recognizing implicit discourse relations in the penn discourse treebank. *Proceedings of the Conference on Empirical Methods on Natural Language Processing*.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkey, Steven J. Bethard, and David McClosky. 2014. The standord corenlp natural language processing toolkit. *Proceedings of the Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.

Thomas Meyer and Bonnie Webber. 2013. Implicitation of discourse connectives in (machine) translation. In *Proceedings of the 1st DiscoMT Workshop at ACL 2013 (51st Annual Meeting of the Association for Computational Linguistics)*, number EPFL-CONF-192528.

Will Monroe and Christopher Potts. 2015. Learning in the rational speech acts model. *arXiv preprint arXiv:1510.06807*.

Megan Moser and Johanna D Moore. 1995. Investigating cue selection and placement in tutorial discourse. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, pages 130–135. Association for Computational Linguistics.

John D Murray. 1997. Connectives and narrative text: The role of continuity. *Memory & Cognition*, 25(2):227–236.

Khanh Nguyen and Brendan O'Connor. 2015. Posterior calibration and exploratory analysis for natural language processing models. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*.

Naho Orita, Eliana Vornov, Naomi H. Feldman, and Hal Daumé III. 2015. Why discourse affects speakers' choice of referring expressions. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Joonsuk Park and Claire Cardi. 2012. Improving implicit discourse relation recognition through feature set optimization. *Proceedings of Annual Meeting on Discourse and Dialogue*.

Gary Patterson and Andrew Kehler. 2013. Predicting the presence of discourse connectives. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, pages 914–923.

Emily Pitler, Mridhula Raghupathy, Hena Mehta, Ani Nenkova, Alan Lee, and Aravind Joshi. 2008. Easily identifiable discourse relations. Technical report, University of Pennsylvania.

Emily Pitler, Annie Louis, and Ani Nenkova. 2009. Automatic sense prediction for implicit discourse relations in text. *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*.

Christopher Potts, Daniel Lassiter, Roger Levy, and Michael C. Frank. 2015. Embedded implicatures as pragmatic inferences under compositional lexical uncertainty. Manuscript.

Rashmi Prasad, Nikhit Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The penn discourse treebank 2.0. *Proceedings of the Language Resource and Evaluation Conference*.

Rashmi Prasad, Bonnie Webber, and Aravind Joshi. 2014. Reflections on the penn discourse treebank, comparable corpora, and complementary annotation. *Computational Linguistics*.

Attapol Rutherford and Nianwen Xue. 2014. Discovering implicit discourse relations through brown cluster pair representation and coreference patterns. *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*.

Ted Sanders. 2005. Coherence, causality and cognitive complexity in discourse. In *Proceedings of the Symposium on the Exploration and Modelling of Meaning*.

Donia Scott and Clarisse Sieckenius de Souza. 1990. Getting the message across in rst-based text generation. *Current research in natural language generation*, 4:47–73.

Erwin M Segal, Judith F Duchan, and Paula J Scott. 1991. The role of interclausal connectives in narrative structuring: Evidence from adults' interpretations of simple stories. *Discourse processes*, 14(1):27–54.

C.E. Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27(379-423; 623-656).

Mark Smith and Linda Wheeldon. 2001. Syntactic priming in spoken sentence production–an online study. *Cognition*, 78(2):123–164.

Claudia Soria and Giacomo Ferrari. 1998. Lexical marking of discourse relations-some experimental findings. In *Proceedings of the ACL-98 Workshop on Discourse Relations and Discourse Markers*.

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

Harry Tily and Steven Piantadosi. 2009. Refer efficiently: Use less informative expressions for more predictable meanings. *Proceedings of the workshop on the production of referring expressions*.

Sara Tonelli, Giuseppe Riccardi, Rashmi Prasad, and Aravind K Joshi. 2010. Annotation of discourse relations for conversational spoken dialogs. In *Proceedings of the Language Resource and Evaluation Conference*.

Jianxiang Wang and Man Lan. 2015. A refined end-to-end discourse parser. *CoNLL 2015*, page 17.

Bonnie Webber. 2009. Genre distinctions for discourse in the penn treebank. 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-Volume 2*, pages 674–682. Association for Computational Linguistics.

Frances Yung, Kevin Duh, and Yuji Matsumoto. 2015. Crosslingual annotation and analysis of implicit discourse connectives for machine translation. In *Workshop of Discourse in machine translation, EMNLP*, page 142.

Frances Yung, Taku Komura, Kevin Duh, and Yuji Matsumoto. 2016. Modeling the interpretation of discourse connectives by bayesian pragmatics. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Sandrine Zuffery and Bruno Cartoni. 2014. A multifactorial analysis of explicitation in translation. *Target*, 26(3).