Modeling the Interpretation of Discourse Connectives by Bayesian Pragmatics

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

We propose a framework to model human comprehension of discourse connectives. Following the Bayesian pragmatic paradigm, we advocate that discourse connectives are interpreted based on a simulation of the production process by the speaker, who, in turn, considers the ease of interpretation for the listener when choosing connectives. Evaluation against the sense annotation of the Penn Discourse Treebank confirms the superiority of the model over literal comprehension. A further experiment demonstrates that the proposed model also improves automatic discourse parsing.

1 Introduction

A growing body of evidence shows that human interpretation and production of natural language are inter-related (Clark, 1996; Pickering and Garrod, 2007; Zeevat, 2011; Zeevat, 2015). In particular, evidence shows that during interpretation, listeners simulate how the utterance is produced; and during language production, speakers simulate how the utterance will be perceived. One explanation is that the human brain reasons by *Bayesian inference* (Doya, 2007; Kilner et al., 2007), which is, at the same time, a popular formulation used in language technology.

In this work, we model how humans interpret the sense of a discourse relation based on the Bayesian pragmatic framework. Discourse relations are relations between units of texts that make a document coherent. These relations are either marked by discourse connectives (DCs), such as '*but*', '*as a result*', or implied implicitly, as in the following examples:

1. He came late. *In fact*, he came at noon.
2. It is late. I will go to bed.

The explicit DC '*in fact*' in Example (1) marks a *Specification* relation. On the other hand, a *Result* relation can be inferred between the two sentences in Example (2) although there are not any explicit markers. We say the two sentences (called *arguments*) are connected by an implicit DC.

Discourse relations have a mixture of semantic and pragmatic properties (Van Dijk, 1980; Lewis, 2006). For example, the sense of a discourse relation is encoded in the semantics of a DC (Example (1)), yet the interpretation of polysemic DCs (such as *'since'*, *'as'*) and implicit DCs relies on the pragmatic context (Example (2)).

This work seeks to find out if Bayesian pragmatic approaches are applicable to human comprehension of discourse relations. Our contribution includes: (i) an adaptation of the Bayesian *Rational Speech Acts* model to DC interpretation using a discourse-annotated corpus, the Penn Discourse Treebank; (ii) integration of the proposed model with a state-of-the-art automatic discourse parser to improve discourse sense classification.

2 Related work

There is increasing literature arguing that the human motor control and sensory systems make estimations based on a Bayesian perspective (Doya, 2007; Oaksford and Chater, 2009). For example, it is proposed that the brain's mirror neuron system

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 531–536,
Berlin, Germany, August 7-12, 2016.

recognizes a perceptual input by Bayesian inference (Kilner et al., 2007). Similarly, behavioural, physiological and neurocognitive evidences support that the human brain reasons about the uncertainty in natural languages comprehension by emulating the language production processes (Galantucci et al., 2006; Pickering and Garrod, 2013).

Analogous to this principle of Bayesian language perception, a series of studies have developed the Grice's Maxims (Grice, 1975) based on game-theoretic approaches (Jäger, 2012; Frank and Goodman, 2012; Goodman and Stuhlmüller, 2013; Goodman and Lassiter, 2014; Benz et al., 2016). These proposals argue that the speaker and the listener cooperate in a conversation by recursively inferring the reasoning of each other in a Bayesian manner. The proposed framework successfully explains existing psycholinguistic theories and predict experimental results at various linguistic levels, such as the perception of scalar implicatures (e.g. '*some*' meaning '*not all*' in pragmatic usage) and the production of referring expressions (Lassiter and Goodman, 2013; Bergen et al., 2014; Kao et al., 2014; Potts et al., 2015; Lassiter and Goodman, 2015). Recent efforts also acquire and evaluate the models using corpus data (Orita et al., 2015; Monroe and Potts, 2015).

Production and interpretation of discourse relations is also a kind of cooperative communication between speakers and listeners (or authors and readers). We hypothesize that the game-theoretic account of Bayesian pragmatics also applies to human comprehension of the meaning of a DC, which can be ambiguous or even dropped.

3 Method

This section explains how we model the interpretation of discourse relations by Bayesian pragmatics. The model is based on the formal framework known as *Rational Speech Acts* model (Frank and Goodman, 2012; Lassiter and Goodman, 2015). Section 3.1 explains the key elements of the RSA model, and Section 3.2 illustrates how it is adapted for discourse interpretation.

3.1 The Rational Speech Acts model

The Rational Speech Acts (RSA) model describes the speaker and listener as rational agents who cooperate towards efficient communication. It is composed of a speaker model and a listener model.

In the speaker model, the *utility* function U defines the effectiveness for the speaker to use utterance d to express the meaning s in context C.

$$U(d; s, C) = \ln P_L(s|d, C) - cost(u) \quad (1)$$

$P_L(s|d, C)$ is the probability that the *listener* can interpret speaker's intended meaning s. The speaker selects an utterance which, s/he thinks, is informative to the listener. The utility of d is thus defined by its informativeness towards the intended interpretation, which is quantified by *negative surprisal* ($\ln P_L(s|d, C)$), according to Information Theory (Shannon, 1948). The utility is modified by production cost ($cost(d)$), which is related to articulation and retrieval difficulties, etc.

$P_S(d|s, C)$ is the probability for the *speaker* to use utterance d for meaning s. It is proportional to the soft-max of the *utility* of d.

$$P_S(d|s, C) \propto \exp(\alpha \cdot U(d; s, C)) \quad (2)$$

where α, the decision noise parameter, is set to 1.

On the other hand, the probability for the listener to infer meaning s from utterance d is defined by Bayes' rule.

$$P_L(s|d, C) \propto P_S(d|s, C)P_L(s) \quad (3)$$

The listener infers the speaker's intended meaning by considering how likely, s/he thinks, the speaker uses that utterance ($P_S(d|s, C)$). The inference is also related to the *salience* of the meaning ($P_L(s)$), a private preference of the listener.

To summarize, the speaker and listener emulate the language processing of each other. However, instead of unlimitted iterations (i.e. the speaker thinks the listener thinks the speaker thinks..), the inference is grounded on literal interpretation of the utterance. Figure 1 illustrates the direction of pragmatic inference between the speaker and listener *in their minds*.

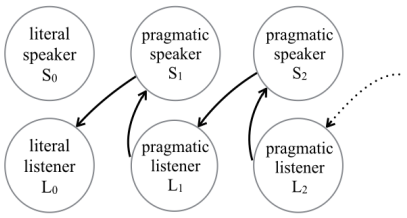

Figure 1: Pragmatic listeners/speakers reason for 1 or more levels, but not the literal listener/speaker.

Our experiment compares the predictions of the literal listener (L_0), the pragmatic listener who

reasons for one level (L_1), and the pragmatic listener who reasons for two levels (L_2). Previous works demonstrate that one level of reasoning is robust in modeling human's interpretation of scalar implicatures (Lassiter and Goodman, 2013; Goodman and Stuhlmüller, 2013).

3.2 Applying the RSA model on discourse relation interpretation

We use the listener model of RSA to model how listeners interpret the sense a DC. Given the DC d and context C in a text, the listener's interpreted relation sense s_i is the sense that maximizes $P_L(s|d, C)$. s_i is specifically defined as

$$s_i = \arg\max_{s \in S} P_L(s|d, C) \tag{4}$$

where S is the set of defined relation senses.

The literal listener, L_0, interprets a DC directly by its most likely sense in the context. The probability is estimated by counting the co-occurrences in corpus data, the Penn Discourse Treebank, in which explicit and implicit DCs are labelled with discourse relation senses.

$$P_{L_0}(s|d, C) = \frac{count(s, d, C)}{count(d, C)} \tag{5}$$

More details about the annotation of PDTB will be explained in Section 4.1.

As shown in Figure 1, the pragmatic speaker S_1 estimates the utility of a DC by emulating the comprehension of the literal listener L_0 (Eq. 1, 2). The probability for the pragmatic speaker S_n to use DC d to express meaning s is estimated as:

$$P_{S_n}(d|s, C) = \frac{\exp(\ln P_{L_{n-1}}(s|d, C) - cost(d))}{\sum_{d' \in D} \exp(\ln P_{L_{n-1}}(s|d', C) - cost(d'))} \tag{6}$$

where $n \geq 1$. D is the set of annotated DCs, including '*null*', which stands for an implicit DC.

The cost function in Equation 6, $cost(d)$, measures the production effort of the DC. As DCs are mostly short words, we simply define the cost of producing *any explicit DC* by a constant positive value, which is tuned manually in the experiments. On the other hand, the production cost for an implicit DC is 0, since no word is produced .

In turn, the pragmatic listener L_1 emulates the DC production of the pragmatic speaker S_1 (Eq. 3). The probability for the pragmatic listener L_n to assign meaning s to DC d is estimated as:

$$P_{L_n}(s|d, C) = \frac{P_{S_n}(d|s, C)P_L(s)}{\sum_{s' \in S} P_{S_n}(d|s', C)P_L(s')} \tag{7}$$

where $n \geq 1$ and S is the set of defined sense. The salience of a relation sense in Equation 7, $P_L(s)$, is defined by the frequency of the sense in the corpus.

$$P_L(s) = \frac{count(s)}{\sum_{s' \in S} count(s')} \tag{8}$$

Lastly, we propose to define the context variable C by the the immediately previous discourse relation to resemble incremental processing. We hypothesize that certain patterns of relation transitions are more expected and predictable. Discourse context in terms of relation sense, relation form (explicit DC or not), and the sense-form pair are compared in the experiments.

4 Experiment

This section describes experiments that evaluate the model against discourse-annotated corpus. We seek to answer the following questions: (1) Can the proposed model explain the sense interpretation (annotation) of the DCs in the corpus? (2) Is the DC interpretation refined by the context in terms of previous discourse structure? (3) Does the proposed model help automatic discourse parsing? We first briefly introduce the corpus resource we use, the Penn Discourse Treebank.

4.1 Penn Discourse Treebank

The Penn Discourse Treebank (PDTB) (Prasad et al., 2008) is the largest available discourse-annotated resource in English. The raw text are collected from news articles of the Wall Street Journals. On the PDTB, all explicit DCs are annotated with discourse senses, while implicit discourse senses are annotated between two adjacent sentences. Other forms of discourse relations, such as 'entity relations', are also labeled. In total, there are 5 form labels and 42 distinct sense labels, some of which only occur very sparsely.

We thus use a simplified version of the annotation, which has 2 form labels (*Explicit* and *Non-explicit* DC) and 15 sense labels (first column of Table 3), following the mapping convention of the CONLL shallow discourse parsing shared task (Xue et al., 2015). Sections 2-22 are used as the

training set and the rest of the corpus, Sections 0, 1, 23 and 24, are combined as the test set. Sizes of the data sets are summarized in Table 1.

	Train	Test	Total
Explicit	15,402	3,057	18,459
Non-Exp	18,569	3,318	21,887
Total	33,971	6,375	40,346

Table 1: Sample count per data set

4.2 Does RSA explain DC interpretation?

The RSA model argues that a rational listener does not just stick to the literal meaning of an utterance. S/he should reason about how likely the speaker will use that utterance, in the current context, based on the informativeness and production effort of the utterance. If the RSA model explains DC interpretation as well, discourse sense predictions made by the pragmatic listeners should outperform predictions by the literal listener.

In this experiment, we compare the DC interpretation by the literal listener L_0, and pragmatic listeners L_1 and L_2. Given a DC d and the discourse context C for each test instance, the relation sense is deduced by maximizing the probability estimate $P_L(s|d, C)$. $P_{L_0}(s|d, C)$ is simply based on co-occurrences in the training data (Eq. 5). $P_{L_1}(s|d, C)$ and $P_{L_2}(s|d, C)$ are calculated by Eq. 6 and 7, in which the salience of each sense is also extracted from the training data (Eq. 8).

	context C	Explicit	Non-Explicit
L_0	**constant (BL)**	.8767	.2616
	prev. form	.8754	.2616
	prev. sense	.8727	.2507
	form-sense	.8684	.2692
L_1	constant	**.8853***	.2616
	prev. form	**.8830**	.2616
	prev. sense	.8671	**.2698***
	form-sense	.8621	**.2671**
L_2	constant	**.8853***	.2616
	prev. form	**.8830**	.2616
	prev. sense	.8671	.2616
	form-sense	.8621	.2616

Table 2: Accuracy of prediction by L_0, L_1 and L_2. Improvements above the baseline are bolded. * means significant at $p < 0.02$ by McNemar Test.

Table 2 shows the accuracy of discourse sense prediction by listeners L_0, L_1 and L_2, when provided with various discourse contexts. Predictions by L_1, when they are differ from the predictions by L_0 under 'constant' context, are more accurate than expected by chance. This provides support that the RSA framework models DC interpretation. Overall, predictions of non-implicit senses hardly differ among different models, since an implicit DC is much less informative than an explicit DC. Moreover, previous relation senses or forms do not improve the accuracy, suggesting that a more generalized formulation of contextual information is required to refine discourse understanding. It is also observed that predictions by L_2 are mostly the same as L_1. This implies that the listener is unlikely to emulate speaker's production iteratively at deeper levels.

4.3 Insights on automatic discourse parsing

Next, we investigate if the proposed method helps automatic discourse sense classification. A full discourse parser typically consists of a pipeline of classifiers: explicit and implicit DCs are first classified and then processed separately by 2 classifiers (Xue et al., 2015). On the contrary, the pragmatic listener of the RSA model considers if the speaker would prefer a particular DC, explicit or implicit, when expressing the intended sense.

In this experiment, we integrate the output of an automatic discourse parser with the probability prediction by the pragmatic listener L_1. We employ the winning parser of the CONLL shared task (Wang and Lan, 2015). The parser is also trained on Sections 2-22 of PDTB, and thus does not overlap with our test set. The sense classification of the parser is based on a pool of lexicosyntactic features drawn from gold standard arguments, DCs and automatic parsed trees produced by CoreNLP (Manning et al., 2014).

For each test sample, the parser outputs a probability estimate for each sense. We use these estimates to replace the *salience* measure ($P_L(s)$) (in Eq. 8) and deduce $P'_{L_1}(s|d, C)$, where C is the previous relation form.

$$P'_{L_1}(s|d, C) = \frac{P_{S_1}(d|s, C)P_{parser}(s)}{\sum_{s' \in S} P_{S_1}(d|s', C)P_{parser}(s')} \quad (9)$$

Table 3 compares the performance of the original parser output and the prediction based on P'_{L_1}.

[1]This does not match with Table 1 as samples labeled with 2 senses are double counted. Multi-sense training samples are splitted into multiple samples, each labelled with one of the senses. In testing, a prediction is considered correct if it matches with one of the multiple senses.

discourse relation sense tags	parser output	P'_{L_1} output	test counts
Conjunction	.7022	**.7079**	1479
Contrast	**.7382**	.7152	1152
Entity	.5174	**.5249**	862
Reason	.4844	**.5105**	661
Restatement	.2773	**.2871**	567
Result	.4019	**.4150**	405
Instantiation	.4346	**.4357**	282
Synchrony	.6553	**.7007**	264
Condition	.9087	**.9302**	238
Succession	.7022	**.7210**	204
Precedence	.7523	**.7762**	200
Concession	.3048	**.4382**	146
Chosen alternative	.5000	**.5200**	36
Alternative	.8421	**.8929**	28
Exception	1.00	1.00	1
Accuracy / Total	.5833	**.5916**	6525[1]

Table 3: F1 scores of original parser output vs parser output modified with P'_{L_1}. Higher scores are bolded. The improvement in accuracy is significant at $p < 0.05$ by McNemar Test.

Significant improvement in classification accuracy is achieved and the F1 scores for most senses are improved. This confirms the applicational potential of our model on automatic discourse parsing.

5 Conclusion

We propose a new framework to model the interpretation of discourse relations based on Bayesian pragmatics. Experimental results support the applicability of the model on human DC comprehension and automatic discourse parsing. As future work, we plan to deduce a more general abstraction of the context governing DC interpretation. A larger picture is to design a full, incremental discourse parsing algorithm that is motivated by the psycholinguistic reality of human discourse processing.

References

Anton Benz, Gerhard Jäger, Robert Van Rooij, and Robert Van Rooij. 2016. *Game theory and pragmatics*. Springer.

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

Herbert H Clark. 1996. *Using language*. Cambridge university press.

Kenji Doya. 2007. *Bayesian brain: Probabilistic approaches to neural coding*. MIT press.

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

Bruno Galantucci, Carol A Fowler, and Michael T Turvey. 2006. The motor theory of speech perception reviewed. *Psychonomic bulletin & review*, 13(3):361–377.

Noah D Goodman and Daniel Lassiter. 2014. Probabilistic semantics and pragmatics: Uncertainty in language and thought. *Handbook of Contemporary Semantic Theory. Wiley-Blackwell*.

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

HP Grice. 1975. Logic and conversation in p. cole and j. morgan (eds.) syntax and semantics volume 3: Speech acts.

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.

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.

James M Kilner, Karl J Friston, and Chris D Frith. 2007. Predictive coding: an account of the mirror neuron system. *Cognitive processing*, 8(3):159–166.

Daniel Lassiter and Noah D Goodman. 2013. Context, scale structure, and statistics in the interpretation of positive-form adjectives. In *Semantics and Linguistic Theory*, volume 23, pages 587–610.

Daniel Lassiter and Noah D Goodman. 2015. Adjectival vagueness in a bayesian model of interpretation. *Synthese*, pages 1–36.

Diana Lewis. 2006. Discourse markers in english: a discourse-pragmatic view. *Approaches to discourse particles*, pages 43–59.

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.

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

Mike Oaksford and Nick Chater. 2009. Prcis of bayesian rationality: The probabilistic approach to human reasoning. *Behavioral and Brain Sciences*, 32:69–84, 2.

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

Martin J Pickering and Simon Garrod. 2007. Do people use language production to make predictions during comprehension? *Trends in cognitive sciences*, 11(3):105–110.

Martin J Pickering and Simon Garrod. 2013. An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, 36(04):329–347.

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*.

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

Teun A Van Dijk. 1980. The semantics and pragmatics of functional coherence in discourse. *Speech act theory: Ten years later*, pages 49–66.

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

Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Rashmi PrasadO Christopher Bryant, and Attapol T Rutherford. 2015. The conll-2015 shared task on shallow discourse parsing. *CoNLL 2015*, page 1.

Henk Zeevat. 2011. Bayesian interpretation and optimality theory. *Bidirectional Optimality Theory. Palgrave Macmillan, Amsterdam*, pages 191–220.

Henk Zeevat. 2015. Perspectives on bayesian natural language semantics and pragmatics. In *Bayesian Natural Language Semantics and Pragmatics*, pages 1–24. Springer.