NAIST-IS-DD1461023

Doctoral Dissertation

Perspectives on the Marking of Discourse Relations: Cognitive Models and Machine Translation

Frances Pikyu Yung

March , 2017

Department of Information Science
Graduate School of Information Science
Nara Institute of Science and Technology

A Doctoral Dissertation
submitted to Graduate School of Information Science,
Nara Institute of Science and Technology
in partial fulfillment of the requirements for the degree of
Doctor of SCIENCE

Frances Pikyu Yung

Thesis Committee:

Professor Yuji Matsumoto	(Supervisor)
Professor Satoshi Nakamura	(Co-supervisor)
Associate Professor Masashi Shimbo	(Co-supervisor)
Associate Professor Kevin Duh (John Hopkins University)	(Co-supervisor)
Assistant Professor Hiroyuki Shindo	(Co-supervisor)
Assistant Professor Hiroshi Noji	(Co-supervisor)

Perspectives on the Marking of Discourse Relations: Cognitive Models and Machine Translation[*]

Frances Pikyu Yung

Abstract

Discourse relations, also known as coherence relations, are the semantic and pragmatic relations between sentences and clauses that make a discourse coherent. On one hand, understanding these relations is the key to comprehend the meaning of a text and the intention of the speaker/writer. On the other hand, producing cohesive discourse with naturally presented discourse relations facilitates communication, for humans and machines alike. Critically, discourse relations can either be explicitly marked by discourse connectives (DCs), such as *therefore* and *but*, or implicitly conveyed in natural language contexts. It is not well understood how speakers choose between the two options, and how this choice impacts applications in natural language processing.

This dissertation explores the marking of discourse relations from two different perspectives. The first part of the study investigates discourse relations from the perspective of human language processing, in the monolingual dimension. A computational psycholinguistic model is proposed to predict whether a discourse relation is marked or not, and how human comprehends explicit and implicit discourse relations. Results are evaluated against corpus annotation as well as behavioral experiments by means of crowd-sourcing.

The second part of the dissertation investigates the marking of discourse relations from an applicational perspective, in the cross-lingual dimension. A bilingual resource of manually annotated discourse relations is constructed, and machine translation experiments are conducted to compare human and machine

[*] Doctoral Dissertation, Department of Information Science, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-DD1461023, March , 2017.

translation of explicit and implicit discourse relations.

This dissertation contributes to the field of computational linguistics by improving the state-of-the-art in the task of predicting speakers' choice of discourse marking, proposing an explanatory cognitive model for the marking of discourse relations based on information-theoretic approaches, building an open-sourced Chinese-English parallel corpus with aligned discourse relations, and advancing our understanding of explicit translation of implicit discourse relations by humans and machine translation systems.

Keywords:

discourse relations, psycholinguistics, cognitive modeling, rational speech acts model, uniform information density, crowdsourcing, linguistic annotation, parallel corpora, statistical machine translation

Acknowledgments

I would like to thank all members of my thesis committee: thank Matsumoto-sensei for all the support and pardon; thank Nakamura-sensei for the always kind comments; thank Shimbo-sensei for reading and correcting my thesis; thank Kevin for giving me advise on research and many other things; thank Shindo-sensei for singing at the karaoke; thank Noji-sensei for advise on my work as well as the whiskey. I'd also like to thank MEXT for the scholarship and Ikoma city for all the facilities and service.

It has been almost six years since I moved to Japan - each year being more dramatic than the previous one. Thank you everybody who had a part in it - folks in the CL lab, the international community in NAIST, neighbours in dorm 8, and other dudes. Even your name is not here, you must have given me help, or lessons, so thank you.

Lastly, I thank my parents for flying all the way from Hong Kong to help me take care of the kids whenever I needed them. Of course, all credits go to Yoshi and Jiro. You made everything happen, from the beginning to the end. Thank you very much and I love you always.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This dissertation investigates the marking of discourse relations in spontaneous and translated texts. In particular, it examines when discourse relations are explicitly presented by discourse markers and when they are implicitly implied, and how this distinction affects the quality of Chinese-English machine translation (MT).

This chapter delineates the basics of this study. First of all, the motivation (Section 1.1) and research questions (Section 1.4) of the study are explained. This is followed by the definition of *discourse relation marking* (Section 1.2) and the methodology used in this study (Section 1.3). Lastly, Section 1.5 outlines the organization of the dissertation.

1.1 Motivation

Natural language does not occur as a random sequence of sentences and clauses. Instead, sentences and clauses relate to each other to form a meaningful discourse. To understand the overall meaning of an article, word-level and sentence-level semantics are not enough. It is essential to understand the semantic and pragmatic relations between clauses or sentences at document level, and these relations are discourse relations.

As a key for language processing, discourse relations draw much attention in psycholinguistic research as well as natural language processing (NLP). There are extensive studies on human comprehension of discourse relations, such as signals of discourse relations, long and short-term processing difficulties, and interaction with other levels of language processing [88, 15, 82, 81, 42, 121, 120, 60, 93]. NLP of discourse relations focuses on automatic analysis of the hidden discourse structure given a surface text. The task is known as discourse parsing or shallow discourse parsing depending on the annotation formalism of the data used. In turn, automatic discourse analysis is used in applications such as automatic summerization, sentiment analysis and text coherence assessment [80, 135, 52, 147].

Research on both human and machine processing of discourse relations centers on the interpretation of discourse relations. On the other hand, human production and machine generation of discourse relations are not extensively explored. To have a complete picture in human language processing, understanding the mechanism behind speakers/writers' production choice is as important as modeling listeners/readers' [1] comprehension. While most literature agree that explicitly marked discourse relations are easier to interpret and/or memorize [121, 120], we still lack a consolidated explanation on how speakers choose to mark a discourse relation or not.

Example 1 illustrates the distinction between marked (explicit) and unmarked (implicit) discourse relations and shows that the choice is not desultory.

(1a) It was a great movie, **but** I did not like it.

(1b) It was a great movie. **Therefore**, I liked it.

(1c) It was a great movie. I liked it.

(1d)* It was a great movie. I did not like it.

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

Explicit and implicit relations differ in their level of ambiguity. Explicit relations can be signaled by a variety of lexical, syntactic and semantic features, of which DCs are the most informative cues to identify discourse relations [110]. In contrast to explicit relations, implicit relations are more ambiguous. For example, "I liked it" can also be read as a JUSTIFICATION for the first sentence in Example 1c.

[1] In this dissertation, *"speakers"* and *"listeners"* are interchangeably used with *"authors"* and *"readers"*, respectively.

Marking a discourse relation or not affects the readability of a text, which is subject to ambiguity and redundancy. On one hand, using an explicit DC avoids ambiguity. For example, if the DC 'but' is omitted as in Example 1d, readers may have problems in inferring the *Concession* sense, and misinterpret that 'I' generally do not like good movies. On the other hand, if the intended discourse sense is highly predictable, it could be verbose or redundant to insert an explicit DC in the utterance, such as the DC *'therefore'* in Example 1b.

The marking of discourse relations is not simply optional and does not depend on the relation sense alone. According to corpus statistics, explicit and implicit discourse relations are equally frequent, yet the corresponding sense distributions are largely different. However, there are not any relation sense that are always explicit or always implicit [113].

The primary motivation of this dissertation is thus to seek a theoretically sound and empirically assessed account on speakers' choice of discourse marking, in order to contribute to a balanced understanding of both the comprehension and production aspects of the discourse phenomenon.

In terms of NLP applications, a model that predicts the natural presentation of discourse relations is also important for automatically generating coherent, human-like texts and dialogues.

In particular, the degree of marking in discourse relations is cross-linguistically different [92, 148]. It remains a challenge for MT systems to explicitate (traslate an implicit DC to an explicit DC) or implicitate (translate an explicit DC to an implicit DC) discourse relations in source texts, as human translators do [8, 47, 48, 74, 92, 148, 157], as it is not yet clear how DC explicitation and implicitation are subject to the convention of discourse marking in the source and target texts. In addition, it is not clear to what extend such cross-lingual distinction is actually incorporated in human translation, which is treated as the gold standard in MT.

The second motivation of this dissertation is thus to assess the impact of discourse marking on the NLP application of MT. The specific language pair under investigation is Chinese-English translation, because of the distinct contrast between the discourse marking strategies in Chinese and English texts.

Example 2 shows two versions of English translation of a Chinese sentence as output by *Google Translate*.

Example 2a – Source text

(如果-if) 交納款有困難的，(便-then) 可暫緩積欠，(但是-but) 新不欠，(而且-furthermore) 掛免罰，(並-and) 逐年清。

Example 2b – Original MT

Difficult to pay taxes, may suspend arrears, the new tex is not owed, penalties linked tax free, paid annually.

Example 2c – MT with inserted source DCs

If you have difficulty to pay taxes, you can suspend the arrears, **but** the new tax is not owed **and** taxes linked to impunity **and** paid annually.

Example 2d – Reference translation

Those having difficulty paying taxes can temporarily postponing old debt **but** not owing on new taxes, **and** suspending taxes **and** waiving fines, **and** paying off year by year.

(adapted from the Translated Chinese Tree Bank Article 89)

In the original Chinese sentence shown in Example 2a, all explicit DCs are omitted. This results in a harmonic surface pattern (same number of syllables per clause), which is appreciated in Chinese writing style. When this sentence is input into the MT system as is, the output (Example 2b) results in a sequence of broken clauses. On the other hand, if implicit DCs that signal the underlying discourse relations are inserted into the source text, as represented by the glossed words in brackets in Example 2a, the clauses are joined by the translated DCs to a complete sentence (Example 2c). In addition, the dropped pronoun 'you' is properly generated, potentially due to improvement in syntactical parsing of the source sentence.

The potential to improve discourse-level machine translation motivates investigation on the cross-lingual choice of discourse marking – whether the marking of the source text should be maintained or reversed in the translation.

1.2 Definition of Marking of Discourse Relations

This section gives a precise definition on what *marking* of discourse relations refer to throughout the dissertation. It is necessary since there is no general consensus on the distinction between marked (or explicit) and unmarked (or implicit) discourse relations.

First of all, the list of DCs differs in different formalism, resources, and languages. In addition, discourse relations can also be alternatively 'marked' by other expressions. For example, the discourse relations in Examples 1a and 1b can also be expressed respectively as follows.

(1e) It was a great movie. Surprisingly, I did not like it.

(1f) The reason I liked the movie was that it was great.

Some studies consider relations in Examples 1e and 1f as marked/explicit relations as well [24], while others only consider DCs as discourse markers and only Examples 1a and 1b as marked/explicit relations In the extreme case, even Example 1c can be categorized as a marked relation, due to the lexical relation between 'great' and 'like' [24].

In this dissertation, *marking of discourse relations* is defined as the *expression of discourse relations by explicit DCs*. Only relations signaled by explicit DCs are considered marked/explicit relations, and relations neither signaled by DCs nor other alternative expressions are considered unmarked/implicit relations. The purpose of this definition is to precisely investigate the production of discourse relations independent of other semantic or syntactic constraints. Expressions other than DCs often carry extra information on top of the discourse relation, as in Example 1e, or embedded in the syntactical structure of the rest of the discourse. Therefore, relations marked by alternative expressions are excluded in this study.

Specifically, the list of explicit DCs depends on the resource used in this study. The first part of the dissertation (Chapters 3, 5) investigates the production of discourse relations in English speakers/writers, based on the Penn Discourse Treebank (PDTB), in which 100 distinct types of DCs are defined. The second part of the dissertation (Chapters 6, 7) investigates the cross-lingual production of discourse relations in Chinese-to-English translation, based on a parallel corpus annotated with discourse relations, in which 156 DC types are identified on the English side and 227 DC types are identified on the Chinese side.

1.3 Methodology

This dissertation examines the marking of discourse relations from two different perspectives. From a psycholinguistic perspective, a *Discourse Marking Model* is proposed to predict speakers' choice of relation marking by quantifying the ambiguity and redundancy of a particular choice. From an applicational perspective, the cross-lingual production of discourse relations is analyzed in human translations and applied to machine translation. This section outlines the methodology used in the two perspectives respectively.

1.3.1 Psycholinguistic approach for discourse marking

This work advocates that speakers' choice of discourse marking is not merely an optional preference characterized by external factors, such as personal writing style. Instead, speakers choose the optimal level of marking in their utterances to facilitate communication. The proposed approach is to combine two information-theoretic frameworks, namely the Rational Speech Acts (RSA) model and the Uniform Information Density (UID) principle, to model how speakers rationally balancing ambiguity and redundancy.

On one hand, the RSA model [31] formalizes the inter-relation of language comprehension and production in terms of a listener model and speaker model, which are interwoven. Recent findings in human language processing suggest that listeners simulate how an utterance is produced to guide comprehension and speakers consider the ease of comprehension when planning production [21, 63, 106]. Based on these findings, the RSA model quantifies the *informativeness* of the choice of discourse marking by the likelihood for the listeners to disambiguate the discourse relation.

On the other hand, the Uniform Information Density (UID) principle [71] is applied to model how redundant utterances are avoided. The UID principle views language communication as a form of information transmission through a noisy channel, through which a constant rate of information flow is optimal according to Shannon's Information Theory [35, 71, 125]. Speakers thus structure utterances by optimizing *information density*, which is the quantity of information (measured by *surprisal*) transmitted per *unit of utterance*, typically *a word*. In particular, a highly predictable utterance triggers a drop in *information density*, which has to be smoothed by choosing a more ambiguous utterance, such as by leaving out linguistic markers.

In short, the proposed model implements Grice's *Maxim of Quantity* [38] by computing how speakers try to be informative (using the RSA model), but not too informative (based on the UID principle). This model is applied to predict whether an explicit or implicit DC is used to express a discourse relation, given the context of the discourse relation and the discourse sense to be conveyed. The model is evaluated using the actual presence or absence of DCs in the PDTB as the gold standard.

As an extension, the RSA model is also applied to illustrate that listeners interpret explicit and implicit discourse relations by simulating the production process.

1.3.2 Applicational approach for discourse marking

A model that predicts speakers' marking of discourse relations not only advances our understanding on human language processing, but is also important for a number of NLP applications, such as natural language generation. In particular, this study investigates the impact of discourse marking on Chinese-English MT. The investigation is divided into two steps: cross-lingual annotation of discourse relations and MT experiments based on oracle pre-explicitation of implicit relations.

Discourse relations are annotated on the raw text of the English Chinese Translation Treebank [11]. Grounded on the linguistic characteristics of Chinese, an annotation scheme is designed to annotate Chinese discourse structures as sequences. An end-to-end discourse parser is built to prove that the annotation formalism is machine-learnable. Then, using the translation spotting technique [17, 91], each connective on the Chinese source text, either explicit or implicit, is aligned with their translation on the English target side. The annotation statistics are analyzed to see how the level of discourse marking is transferred from the source text to the target text in human translation.

Next, based on the observation that implicit Chinese relations are often translated explicitly in English, explicit DCs are inserted into the source text to represent the implicit discourse relations. In other words, implicit source relations are artificially explicitated. This pre-processing is based on the manual annotation instead of automatic discourse analysis of the source text, in order to isolate the effect of the discourse parser performance from the assessment of the pre-processing effect. MT results of the pre-processed and original source texts are compared and the transfer of discourse marking is analyzed. Results of the experiments provides directions on how to translation implicit discourse relations in MT.

1.4 Research questions

This dissertation contributes by answering below research questions:

1. From a cognitive perspective, can speakers' choice of discourse marking be explained by RSA and UID?
2. How does the proposed information-theoretic model compared with a machine-learning approach, such as the state-of-the-art [105] of the task of discourse marking prediction?
3. From an applicational perspective, how is discourse marking *reproduced* cross-lingually in human Chinese-English translation ?
4. Can MT be improved when source implicit discourse relations are *pre-explicitated*?

Answers to these questions will be summarized in Chapter 8.

1.5 Organization of the dissertation

The rest of this dissertation is organized as follows:

Chapter 2 provides an introduction to discourse relations, including various formalism and resources, human discourse processing and NLP of discourse relations. Related works on RSA and UID are also introduced.

Chapters 3 and 5 investigate the marking of discourse relations from the psycholinguistic perspective. Chapter 3 describes the proposed Discourse Marking Model that predicts speakers' choice of explicit or implicit DCs when producing a particular discourse relation in a particular context. Chapter 4 presents a psycholinguistic behavioral study that evaluate humans' performance of the discourse marking task in comparison with the model prediction. Chapter 5 explores the marking of discourse relations in the reverse direction, i.e. how listeners' interpret explicit and implicit discourse relation based on the ambiguity of the utterance.

Chapters 6 and 7 investigate the marking of discourse relations from the applicational perspective. Chapter 6 introduces the annotation of discourse relations on a translation corpus, starting from the sequential annotation of Chinese discourse structures and followed by the cross-lingual alignment of DCs. Finding of the annotation statistics is also discussed. Chapter 7 describes the MT experiments based on several variations of *pre-explicitation* of implicit relations in the source text. The results are compared base on automatic metrics and manual analysis of random samples.

Chapter 8 summarizes the contributions of the dissertation and discusses directions for future work.

Chapter 2

Background

This chapter introduces three lines of existing work that are related to this study. Section 2.1 gives an overview on various theories of discourse relations. In particular, the PDTB-styled discourse annotation is explained with examples (Section 2.1.1), since the proposed cognitive model is based on PDTB. A summary of theories and experimental findings in human discourse processing is also given in Section 2.1.2.

Section 2.2 outlines literature on the automatic analysis of discourse relations (Section 2.2.1), in particular of English and Chinese languages. Efforts to incorporate discourse structures to machine translation are also introduced (Section 2.2.2).

Lastly, background on information-theoretic approaches for cognitive modeling of language processing is introduced in Section 2.3, including Bayesian approach for pragmatic language processing (Section 2.3.1) and the principle of constant entropy rate (Section 2.3.2). These are the theories behind the key components of the proposed Discourse Marking Model.

2.1 Discourse relations

There are numerous theories that explain the discourse phenomenon, e.g. [46, 40]. The major difference is the taxonomy of relation senses defined. For example, the Rhetorical Structure Theory [83] represents discourse relations in a tree structure, where a *satellite* text span is related to a *nucleus* text span, and a taxonomy of 23 senses are defined. This work is based on the lexical grounded theory used in the annotation of the PDTB, which represents a discourse relation by an explicit/implicit DC binding exactly two discourse units.

2.1.1 The Penn Discourse Treebank

This work applies a compuational model to predict the actual marking of discourse relations in corpus data given a particular discourse relation. To achieve this, a corpus annotated with discourse relations and marking is essential. There are various corpora annotated with discourse relations, such as the RST Discourse Treebank [16] and Discourse GraphBank [142], but discourse markers are annotated and associated with discourse relations only in two resources: the PDTB [113] and RST Signalling Corpus [24]. The proposed model in this work is trained and evaluated against the annotation of the PDTB. This section gives a breif introduction to the annotation scheme of this corpus.
PDTB is the largest available discourse-annotated corpus in English. The PDTB consists of news articles collected from the *Wall Street Journal*. Discourse relations are identified between two discourse units, usually clauses or sentences, and the sense of the relation is annotated from a defined list of sense labels. 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 limit how far futures prices could fall in a single session and ... to give market operators the authority to suspend trading in futures at any time. (Implicit: **but**; COMPARISON) Maybe it wasn't enough. *(WSJ0097)*

(3) This cannot be solved by provoking a further downturn; reducing the supply of goods does not solve inflation.
(Implicit 1: **instead** EXPANSION.ALTERNATIVE.CHOSEN ALTERNATIVE),
(Implicit 2: **so**; CONTINGENCY.CAUSE.RESULT and EXPANSION.ALTERNATIVE)
Our advice is this: Immediately return the government surpluses to the economy through incentive-maximizing tax cuts, and find some monetary policy target that balances both supply and demand for money. *(WSJ0553)*

The PDTB follows a lexically-grounded approach in the annotation of discourse relations [141]. Each discourse relation is represented in a predicate-argument-like structure, where discourse connectives (DCs) relates two text spans (*Arg1* and *Arg2*), as shown in Figures 2.2 and **??**.

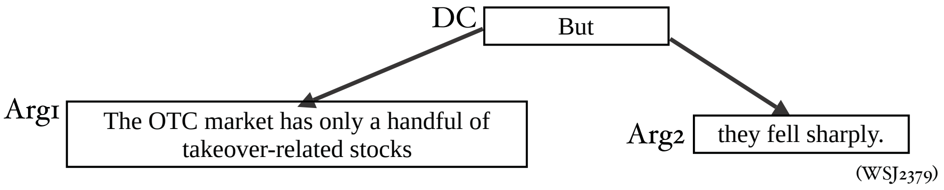

DC
But
Arg1
The OTC market has only a handful of takeover-related stocks
Arg2
they fell sharply.
(WSJ2379)

Figure 2.1: Example of a discourse connected by an explicit DC

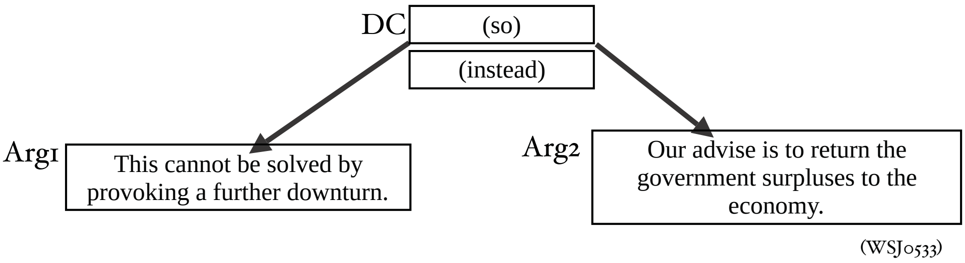

DC
(so)
(instead)
Arg1
This cannot be solved by provoking a further downturn.
Arg2
Our advise is to return the government surpluses to the economy.
(WSJ0533)

Figure 2.2: Example of a discourse connected by an implicit DC

In the annotation process, *explicit* DCs are first identified, based on a list of DCs that are accumulated in the course of annotation, and labelled with relation senses (Example (1)). Other expressions that signal discourse relations, such as “the reason is”, are identified as *alternative lexicalization (AltLex)* and labelled with relation senses as well. If explicit markers are absent *between two sentences* within the same paragraph, there are three options for annotation: i) if a discourse relation can be inferred and expressed by a DC, the relation is labelled as *implicit* and the candidate DC and relation sense are annotated (Example (2)); ii) if a discourse relation cannot be inferred but the two sentences are about the same entity, the relation is labelled *EntRel*; and iii) if the two sentences are unrelated, they are tagged as *NoRel*.

Senses in the PDTB are defined in a hierarchy of two to three levels, as shown in Figure 2.3. Some relations have multiple senses. Up to two DCs can be annotated to an implicit relation and, in turn, each (implicit or explicit) DC can be labelled with up to two senses (Example (3)). Similarly, certain level 2 senses, as in Example (2), are resulting from the back-off strategy in annotation, i.e. when the annotators disagree on the level 3 senses. This is also a kind of multipe sense. It is arguable that multi-sense discourse relations are non-compositional, thus each combination of multiple senses is considered as an individual sense. This will be discuessed in the marking model in Chapter 3.

PDTB's annotation scheme is adapted by discourse treebanks of other languages, such as the Chinese Discourse Treebank (CDTB) [155]. Combinition of the RST and PDTB formalisms is also proposed. [152] adds the distinction

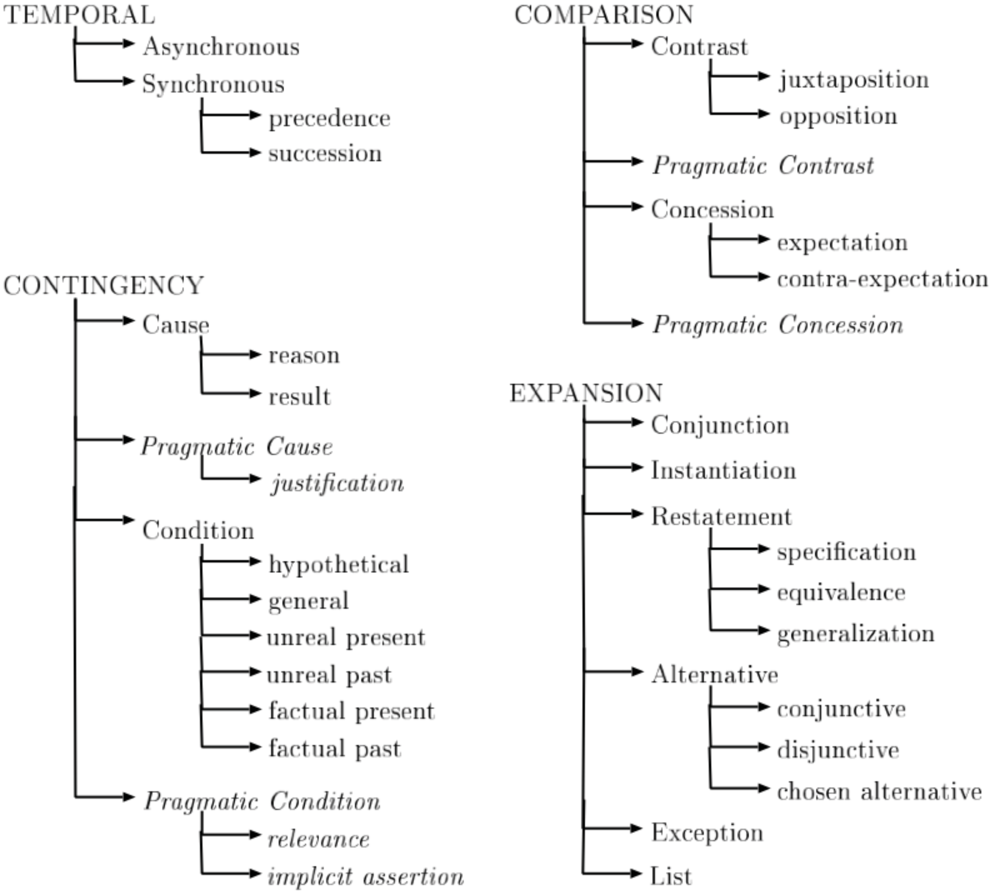


Figure 2.3: Sense hierarchy of PDTB, reprinted from [113]

of *satellite* and *nucleus* to PDTB-style annotation, and [76] labels the connectives in an RST tree. Other efforts to exploit Chinese discourse relations include cross-lingual annotation projection based on machine translation or word-aligned parallel corpus [151, 75].

To investigate the cross-lingual difference in the marking of discourse relations, this work presents the first bilingual annotation effort of both explict and implicit discourse relations on a Chinese-English translation corpus. Discourse relations are first annotated on the Chinese side of the corpus and then aligned to the English side. The annotation scheme for Chinese discourse relations is independent of the exisiting approaches. Details will be explained in Chapter 6.

2.1.2 Human discourse processing

The marking of discourse relations is a well studied topic in the psycholinguistic literature, but the focus is on how the marking affects the interpretation of discourse relations.

Most literature agree that explicitly marked discourse relations are easier to interpret and/or memorize [121, 120]. Some studies also find that the relations signaled by DCs are recalled better in long term (i.e. improved mental representation) [81, 82, 88, 93], while other studies conclude that the effect is not significant, or may depend on medium (written or speech) or relations [87, 115, 121, 120, 119, 129].

This work focuses on the marking of discourse relations in human language production, using the *speaker model* of the RSA model. As an extension, the *listener model* of RSA is also applied to model the pragmatic interpretation of marked and unmarked relations. This model is described in Chapter 5.

****

To summarize, discourse relations connect individual clauses and sentences in a text such that the text is cohesive and meaningful. Among various theories proposed to formalized discourse relations, the PDTB adopts a lexically grounded approach to associate each discourse relation with a discourse connective, which is either marked or implicit. From the view of human language processing, it is generally agreed that explicitly marked discourse relations are easier for humans to interpret.

2.2 Natural language processing of discourse relations

The first part of the thesis tackles the task of automatic prediction of discourse marking, which is one of the NLP tasks dealing with discourse relations. This section gives a brief survey on NLP of discourse relations, in particular discourse parsing and application of discourse on MT.

2.2.1 Automatic discourse parsing

Following the release of discourse-annotated corpora, such as PDTB and RST-treebank, machine learning-based analyzers of discourse relations are proposed. These discourse parsers generally work in a pipeline of steps [79]. The text is first segmented into discourse units, followed by identification of relations between the units. Earlier approaches depend on feature engineering and various features have been exploited [78, 108, 103, 109, 80]. Implicit discourse relations are much harder to learn than explicit discourse relations [108, 156], although explicit relations can also be ambiguous. For example, classification of the 4 main relation senses (temporal, contingency, comparison, expansion) reaches 94% accuracy for explicit relations [109], but only range from F-scores of 20 for 'temporal' to 76 for 'expansion' relations, due to uneven distribution of the relation senses[108, 156]. Therefore, recent interest of research focuses on inference of implicit discourse relations, particularly based on latent representation of texts (e.g. [57, 126]) and by creating pseudo implicit DCs training instances from explicit relations [86, 118].

Similar approaches have been applied to Chinese discourse parsing, yet less data is available and the performance is inferior comparing with that of English. For example, classification of inter-sentence discourse relations reaches an F-score of 64 [50] and 2-way classification of intra-sentence contingency and comparison relations reaches an F-score of 71 [51], training on a moderately sized (81 articles) corpus and considering explicit and implicit relations collectively.

Overall, the current state-of-the-art of end-to-end discourse parsing, based on the PDTB data, is 31 (F1) for English and 41 F1 for Chinese [144]. Although there is still room for improvement, automatic discourse analysis is incorporated in a number of applications, or jointly modeled in with other tasks, such as sentiment analysis and language modeling [58].

2.2.2 Discourse relations in machine translation

Humans translate from document to document, because the meaning of a particular sentence depends also on discourse structure. Although research of machine translation (MT) had long been limited to sentence-level translation, recent efforts start to explore the possibility to incorporate linguistic information outside the sentence boundary, such as topical structure, coreference chains, and lexical coherence [45].The second half of this thesis examines the transfer of discourse relation marking in machine translation. This section summarizes previous works on the machine translation of discourse relations.

Earlier studies of discourse relations in MT includes [85], which proposed a discourse transfer model to re-construct the target discourse tree from the source discourse tree. However, incorporation to an SMT system was not discussed in the work. Recent works focus on the translation of ambiguous DCs, such as 'since' in the temporal sense vs. 'since' in the reason sense. This is achieved by annotating the DCs in the training data by 'translation spotting', which is to manually align the DCs of the source text to their translation in the target text, either occurring as DCs or other expressions [91, 111, 90, 89, 17]. Experiments of these works have been conducted in translations among English, French, Czech, German and Arabics and only explicit DCs were considered. Explicitation of implicit DCs is observed in translations between European languages [8, 157]. On the other hand, it is also reported that certain English explicit DCs are not translated explicitly in French or German [92]. Comparing with other languages, Chinese sentences is 'discourse-like', consisting of a sequence of discourse units. Therefore, this work hypothesizes that explicitation is more common in Chinese-to-English translation.

Previous works on discourse-awared Chinese-to-English translation include [136], which extracts translation rules from the RST-styled discourse structure output by an automatic parser. An improvement of 1.16 BLEU point is reported, considering only intra-sentential explicit relations. Similarly, [137] presents a tree-to-string translation model in which translation rules and language model are conditioned by the syntactic structure based on complex sentence parsing (CCS) [153]. CCS parses cover certain inter-sentential discourse relations that are either explicit or implicit. Improved BLEU scores are reported, but it is not clear how much the the improvement can be attributed to improvement in discourse relation translation.

Improvements in DC translation are not always sensitive to conventional evaluation metrics [90], since DCs make up to only a small portion of tokens in the source and target texts. Specialized metrics to assess DC translation are developed, based on bilingual word alignment and a dictionary of DCs [43, 44]. Still, evaluation on missing/additional DC (i.e. potential implicitation/explicitation of discourse relations) relies on manual analysis.

****

To summarize, automatic discourse parsing is a non-trivial task, and, similar to human discourse processing, sense classification of implicit relations is harder than relations marked by DCs. Machine translation can benefit from disambiguation of explicit discourse relations, but it is not yet clear if translation of implicit relations are learnt in existing models. Cross-linguistically, it is observed that languages differ in the level of discourse marking. A discourse relation can be explicit in one language but implicit in another. However, the transfer of discourse marking from the source language to target language is not yet specifically modeled or evaluated in current MT systems.

2.3 Information-theoretic frameworks for human language processing

This section provides background on two frameworks based on Information Theory, which are the foundation of the Discourse Marking Model described in the first part of the thesis. These frameworks, namely Bayesian pragmatic reasoning and uniform information density, have been applied in previously work, individually, to explain language comprehension and production.

2.3.1 Bayesian pragmatic reasoning and the rational speech act model

A growing body of evidence shows that human interpretation and production of natural language are inter-related [21, 106, 149, 150]. 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 motor control and sensory systems reason by *Bayesian inference* [25, 63], which is, at the same time, a popular formulation used in language technology.

For example, it is proposed that the brain's mirror neuron system recognizes a perceptual input by Bayesian inference [63]. 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 [32, 107].

Analogous to this principle of Bayesian language perception, a series of studies have developed the Grice's Maxims [38] based on game-theoretic approaches [56, 31, 37, 36, 9]. 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 successfully simulated experimental results concerning different aspects of human communicationat 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 (e.g. using pronouns or proper nouns to refer to an entity) [66, 37, 10, 61, 112, 67].

The RSA model [31] is a variation of these game-theoretic approaches in pragmatics. On top of reproducing experimental data, recent works also learn the RSA model from corpus data. For example, Orita et al.[101] applies RSA model to predict the choice of referring expressions in corpus data and Monroe and Potts [94] optimizes a 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).

Production and interpretation of discourse relations is also a kind of cooperative communication between speakers and listeners (or authors and readers). The Discourse Marking Model proposed in this thesis thus applies 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). Since discourse relations can be marked or unmarked, inference of discourse relations can be explained in a unified framework as scalar implicatures [2]. The proposed model is built upon the *speaker's model* of RSA to predict speaker's choice of explicit or implicit DCs. Details of the model are explained in Chapter 3 together with the proposed method.

2.3.2 Entropy Rate Constancy Uniform Information Density in natural language

Shannon's Information Theory states that the most efficient way of communication in a noisy channel is to send information at a constant rate [125]. Based on this theory, the principle of *Entropy Rate Constancy* argues that human language communication also obeys the Information Theory and produce language at a constant entropy rate [35], which is defined as the *surprisal* of a string. Analysis of written corpora as well as dialogues reveals that the entropy of a sentence, taken out of context, tends to increase with sentence number, providing support to the principle [35, 145] and correlate with processing effort as represented by eye-tracking data [62].

Grounding on the principle of *Entropy Rate Constancy*, the *Uniform information density* (UID) principle [71] states that speakers structure utterances by optimizing *information density*, which is the quantity of information (measured by *surprisal*) 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 [6], morphological [30] and syntactic [54] reductions, and also referring expressions [133]. The UID principle provides a theoretical basis that connects the use of DCs with other discourse relation signals. According to UID, information density rises when an 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 omitting linguistic markers. In the context of discourse relations, explicit DCs are omitted when the discourse relation is highly predictable.

Analyses of the PDTB in the literature show that Causal and Continuous senses are more often implicit, or marked by less specific DCs [3]. Indeed, these senses are presupposed by listeners according to linguistics theories [65, 70, 96, 119, 124]. In contrast, the DC *instead* is more often dropped for the discourse relation Chosen Alternative, if the first argument contains negation words, which are identified cues for this relation [5].

The corpus statistics presented in these analyses support the UID hypothesis 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 the first argument is negated. Although measures have been proposed to rate the implicitness of a relation sense [4, 59], these measures only quantify the general marking of each sense in the data (e.g., the contrast sense), but not the speaker's choice for each particular instance (e.g., the contrast sense, given particular arguments and context).

In contrast, the model proposed in this work incorporates an *information density predictor*, which specifically predicts the expectability of a given relation. In turn, the speaker's choice of discourse marking is biased based on the predicted degree of expectability. Instead of particular senses or cues in the corpus, UID is generally applied to model each relation instance of any relation sense in the corpus, in conjunction with other language production factors.

****

To summarize, speakers' representation of discourse relations can be explained by Bayesian pragmatic reasoning as well as the UID principle. However, explanation by the two frameworks appear contradictory. Bayesian pragmatic reasoning asserts that speakers try to be more informative by marking discourse relations, while the UID principle asserts that markers should be dropped to avoid trough in information density. Chapter 3 presents the first proposal to combine these two frameworks into one unified perspective, by considering the balance of information between the DCs and the discourse arguments.

Chapter 3

A Psycholinguistic Model on the Marking of Discourse Relations

The first part of this dissertation investigates the marking of discourse relation in spontaneous language produced by humans. The objective is to explain how human speakers choose the optimal level of marking in their utterances, either intentionally or subconsciously.

Speakers or authors produce informative utterances such that listeners or readers can understand the intended message. Grice's *Maxim of Quantity* states that human speakers communicate by being as informative as required, but no more [38]. If a speaker always tries to provide as much information as possible, the resulting utterance could become excessively long and tedious. Such utterance not only takes effort for the speaker to produce, but also contains redundant information that is not necessary for the listener.

This work combines two information-theoretic frameworks, namely the Rational Speech Acts (RSA) model [31] and the Uniform Information Density (UID) principle [71], into a psycholinguistic model that simulates how speakers reason the balance between ambiguity and redundancy.

On one hand, the RSA model formulizes the interwoven relation of language comprehension and production in terms of a listener model and speaker model. These models are grounded on the findings in human language processing that listeners simulate how an utterance is produced to guide comprehension and speakers consider the ease of comprehension when planning production [21, 63, 106]. Tthe RSA model quantifies the *informativeness* of the choice of discourse marking by the likelihood for the listerners to disambiguate the discourse relation.

On the other hand, the Uniform Information Density (UID) principle models show redundant utterances are avoided. The UID principle views language communication as a form of information transmission through a noisy channel, through which a constant rate of information flow is optimal according to, Shannon's Information Theory [35, 71, 125]. Speakers thus structure utterances by optimizing *information density*, which is the quantity of information (measured by *surprisal*) transmitted per *unit of utterance*, typically *a word*. In particular, a highly predictable utterance triggers a drop in *information density*, which has to be smoothed by choosing a more ambiguous utterance, such as by leaving out linguistic markers. Figures 3.1 and 3.2 demonstrates the application of UID to explain the preference of explicit or implicit DCs.

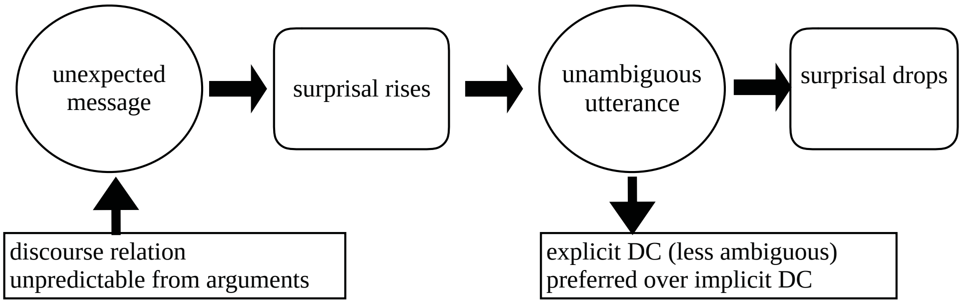

unexpected message
surprisal rises
unambiguous utterance
surprisal drops
discourse relation unpredictable from arguments
explicit DC (less ambiguous) preferred over implicit DC

Figure 3.1: Choice of an explicit DC based on UID

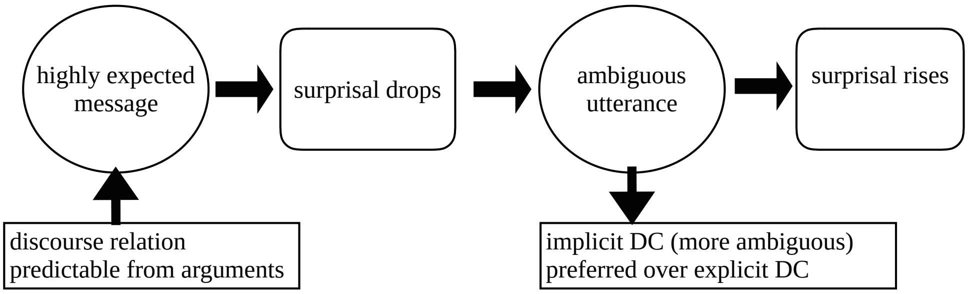

highly expected message
surprisal drops
ambiguous utterance
surprisal rises
discourse relation predictable from arguments
implicit DC (more ambiguous) preferred over explicit DC

Figure 3.2: Choice of an implicit DC based on UID

In short, this work presents a computational psycholinguistic model that implements Grice's *Maxim of Quantity* by computing how speakers try to be informative (using the RSA model), but not too informative (based on the UID principle). The model is applied to predict whether an explicit or implicit DC is used to express a discourse relation, given the context of the discourse relation and the discourse sense to be conveyed. Using the actual presence or absence of DCs in the PDTB as the gold standard for evaluation, the proposed model not only achieves a higher accuracy than previous work [105], but also provides an interpretable account of the various cognitive factors behind the predicted decision.

This chapter is organized as follows. Section 3.1 first introduces related studies on discourse relation marking. Section 3.2 explains the prerequisite of the RSA model, following the adaptation to model discourse marking. Experiments and evaluation using the corpus data of PDTB are presented in Section 3.3. Section 4 presents the evaluation of the model by a psycholinguistic experiment conducted through crowsdsourcing. A conclusion is drawn in Section 3.4.

3.1 Previous work on discourse marking prediction

The choice of discourse marking strategies has been studied in earlier works as a subtask for natural language generation [1, 23, 41, 95, 123, 128]. In the absence of large-scale resources, investigations are based on manually derived rules and lexicons or psycholinguistic experiments.

With the emergence of large corpora annotated with discourse relations, [105] presented a machine-learning approach to predict whether an explicit or implicit DC is used in the corpus for a particular discourse relation. They argue that while the choice is related to the ease of inference, it may also depend on other stylistic or textual factors. A classifier is trained to predict whether a *candidate DC* (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. Features include observable surface forms, such as presence of percentage and dollar signs, argument length, count of subject nouns, and content word ratio, as well as contextual discourse structures, such as the previous discourse relation and whether the relation is embedded or shared. The classifier is trained and tested on a subset of the most frequent relations from the PDTB, after screening away infrequent senses and DCs. An overall high classification accuracy of 86% is achieved and relation-level and discourse-level features are found to be more useful than argument-level features. Nonetheless, their approach does not provide theoretically grounded explanation of why an utterance is preferred by the speaker.

The Discourse Marking Model proposed in this chapter also predicts the use of explicit or implicit DCs in PDTB, as in [105]. However, instead of a data-driven approach that focuses on correctly replicating the occurrence of DCs in the corpus, the proposed model explains the speakers' option of marking from the viewpoint of human language production. Although the proposed model does not make use of the *candidate DC* as a feature, which is the result of the speaker's choice, if an explicit DC is preferred, it achieves higher accuracy than [105] when evaluated on the same test samples.

3.2 The Discourse Marking Model

This section describes the proposed method for modeling the speaker's choice of DC marking. Prerequisite of the RSA model is first explained, followed by the details of the proposed marking model, which predicts the marking of a discourse relation produced by a speaker based on the *speaker model* of RSA.

3.2.1 The RSA model

The RSA model describes the speaker and listener as rational agents who cooperate towards efficient communication. 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 3.1).

$$P_{listener}(s|w, C) \propto P_{speaker}(w|s, C)P(s) \tag{3.1}$$

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 soft-max optimizing the expected *utility* ($U(w; s, C)$) of the utterance (Equation 3.2),

$$P_{speaker}(w|s, C) \propto e^{\alpha \cdot U(w;s,C)} \tag{3.2}$$

where α is the decision noise parameter, which is set to 1 to represent the Luce's choice axiom [31], i.e. a rational decision without bias [1].

The speaker emulates the listener's interpretation and chooses an utterance s/he believes to be informative. Since 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 3.3).

$$U(w; s, C) = I(s; w, C) - D(w) \tag{3.3}$$

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 3.4).

$$I(s; w, C) = \ln P(s|w, C) \tag{3.4}$$

The Discourse Marking Model is based on the speaker's model of RSA. Section 3.2.2 explains how the RSA model is adapted to discourse presentation, followed by the details of each component (Sections 3.2.3 to 3.2.5).

3.2.2 RSA for discourse relation presentation

According to Equation (3.2), the probability for a speaker to use utterance w to convey his intended message s in context C is:

$$P(w|s, C) = \frac{e^{U(w;s,C)}}{\sum_{w' \in W} e^{U(w';s,C)}} \tag{3.5}$$

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

[1] $\alpha = 0$ means the decision is totally unrelated to pragmatic reasoning. $\alpha > 1$ suggests biased choices.

to convey s, the sense of discourse relation. Therefore, speaker's choice of DCs is predicted based on the following two probabilities:

$$\begin{aligned} P(exp|s,C) &= \frac{e^{U(exp;s,C)}}{e^{U(exp;s,C)} + e^{U(imp;s,C)}} \\ P(imp|s,C) &= \frac{e^{U(imp;s,C)}}{e^{U(exp;s,C)} + e^{U(imp;s,C)}} \end{aligned} \tag{3.6}$$

According to Equation (3.3), the *utility* U of an explicit DC equals to its *informativeness* I deducted by production cost D.

$$U(exp;s,C) = I(s;exp,C) - D(exp) \tag{3.7}$$

$I(s;exp,C)$ 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 the objective here is to quantify the informativeness of the DC only. Therefore, $I(s;exp,C)$ is defined 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*.

$$I(s;exp,C) = \ln P(s|exp,C) - \ln P(s|C) \tag{3.8}$$

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 3.2.3.

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 [5].

Therefore, according to UID, an implicit DC is preferred if the arguments are informative. The utility of an implicit DC is therefore raised 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 ($args$)[2].

$$e^{U(imp;s,C)} = e^{U(null;s,C)} + e^{U(args;s,C)} \tag{3.9}$$

$$U(null; s, C) = I(s; null, C) - D(null) \tag{3.10}$$

$$U(args; s, C) = I(s; arg, C) - D(args) \tag{3.11}$$

The amount of information that the null DC provides for the discourse relation is defined similarly as in Equation (3.8):

$$I(s; null, C) = \ln P(s|null, C) - \ln P(s|C) \tag{3.12}$$

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. $I(s; arg, C)$ is thus approximated by *the confidence of a discourse parser in predicting discourse senses from the arguments.* Details will be explained in Section 3.2.4.

Lastly, various psycholinguistically motivated measures are explored to approximate the prodcution cost $D(exp)$ in Section 3.2.5. In contrast, no effort is required to produce a *null* DC. Also, it is assumed 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:

$$e^{U(exp;s,C)} > e^{U(null;s,C)} + e^{U(args;s,C)} \tag{3.13}$$

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

3.2.3 Informativeness of DCs

This section explains how the informativeness in Equations (3.8) and (3.12) are estimated. In discourse production, the utterance lexicon, $W = \{exp, imp\}$ in Equation (3.5), and the set of speaker's intended messages (all possible discourse

[2] In turn, an explicit DC is preferred if the arguments are not informative. It is also plausible to 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 3.13.

relation senses) are always *valid*[3]. 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. These empirical distributions are extracted from the training portion of the corpus.

The context C can be defined by the surrounding discourse relations. Specifically, the discourse contexts (and their abbreviation in Table 3.2) 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', and the pair of sense and form (SF or TSF). In practice, 5 forms of discourse presentation are used, based on the definition in the PDTB: explicit DC, implicit DC, alternative lexicalization, entity relation and 'no relation'. 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).

It is hypothesized that the speaker also thinks ahead the coming discourse structures when planning the current ones. Various discourse contexts are compared in the experiment.

3.2.4 Informativeness of arguments

$I(s; arg, C)$ in Equation (3.11) refers to the amount of information in the arguments that contributes to the interpretation of the discourse sense. According to UID, information density[4] 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 [5].

Generalizing this idea to capture various cues in the arguments for various senses, the proposed model approximates $I(s; arg, C)$ 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 [108, 78, 103, 117]. If the arguments contain much informative features, the

[3] 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.

[4] This is opposite to *'informativeness'* in RSA, which is defined by *negative surprisal* (Equation 3.4).

parser will predict the sense more confidently.

Two methods are proposed, 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, the *negative surprisal* of the estimated probability P_p of the parser output sense s_{output} (Equation 3.14) is used to approximate $I(s; arg, C)$.

$$I(s; arg, C) \approx w_a \cdot \ln P_p(s_{output}) \tag{3.14}$$

At the same time, the probability distribution of all senses is less uniform if one sense is assigned a high probability. Therefore, alternatively, $I(s; arg, C)$ is approximated by the *negative entropy* of the probability distribution estimated by the parser (Equation 3.15). Note that although $I(s; arg, C)$ is approximated byinformation-theoretic measures, these approximations are not related to the formulation of RSA nor UID.

$$I(s; arg, C) \approx w_a \sum_{s_p \in O} P_p(s_p) \log P_p(s_p) \tag{3.15}$$

where O is the set of senses defined in the parser and w_a is a positive weight tuned on the dev set. The *general informativeness* of the arguments to imply *any* discourse senses is measured, so s_{output} does not necessarily equal s.

The implicit sense classifier from the winning parser [139] of the CoNLL shared task 2015 is used in modeling argument informativeness. This classifier is designed to identify a subset of 14 implicit senses plus the *entity relation*. This implicit DC classifier is trained by Naive Bayes based on a pool of proven features, including syntactic features, polarity, immediately preceding DC, and Brown cluster pairs, production rules, dependency rules, last word or argument 1, first 3 words of argument 2, presence of modality verbs and inquirer, Syntactic features are based on automatic parsing using Stanford CoreNLP [84]. The parser is trained on the same sections of the PDTB as the training set used in the experiment.

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. Although the performance of this state-of-the-art implicit DC classifier is still unsatisfactory (34.45%[5] on PDTB Section 23), the model only makes use of the probability estimation of the prediction.

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

The 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, specifically 28.45% vs. 51.30% on test set, matching at the 4 top level discourse sense and counting predictions of *entity relation* as *Expansion*. This supports the motivation to use the parser estimation as an information density predictor.

3.2.5 Cost function

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

Mean DC length: Production cost intuitively increases with word length. The mean DC length of a discourse relation is defined by 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. The word length of the *candidate DC* is not used, because speakers first decide to use an explicit DC or not, then decide which DC best expresses the relation.

DC/arg2 ratio: Similarly, another option is to 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. This version of cost function use sthe 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 [69, 13, 127].

Prime distance: This version uses 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 [39, 14, 55].

Distance from start: This version is the relative position of the relation within the article as the production cost. It is hypothesized that more effort is needed as the production proceeds.

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

$$D(exp) = w_c \cdot cost(exp) \tag{3.16}$$

3.3 Experiment

This section describes an experiment that applies 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 find out if the proposed model explains the factors affecting speaker's choice of DC marking and how the prediction performance compare with the state-of-the-art, i.e. Patterson and Kelher [105]. The details of the experimental data and settings are first described in the next section.

3.3.1 Setting

The experiment is based on the annotation of discourse relation senses and explicit/implicit DCs in the PDTB. This work focuses on the marking of discourse relations by discourse connectives, so only samples labelled *explicit* or *implicit* are used, while annotations of other forms of discourse relations, such as entity relations and attributions, are excluded. In addition, the proposed 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 [114]. In addition, as a result of the annotation procedure, implicit relations always occur *inbetween two arguments*. 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, only samples in the Arg1-DC-Arg2 order are used. In the testing phrase, excluded samples are counted as *explicit* by default.

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 samples, each labelled with one of the senses. However, as mentioned in Section 2.1.1 of Chapter 2, it is notable that the individual senses of a multi-sense relation are not disjoint and *having multiple senses* is *part of the sense* [4, 114]. Multi-sense is an important factor of a 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, all combinations of senses are treated as *individual senses*, each containing one to three joint sense labels. This results in a total of 122 senses. In fact, there is only one sample of three joint labels in the experimental dataset, although up to four joint labels are possible (two implicit DCs labeled two senses each).

The resulting experimental data set contains $5,201$ explicit and $16,049$ implicit relations, after 4 cases of intra-sentential implicit relations, due to sentence splitting errors of the PTB (single sentences wrongly splitted into two), are removed. Table 3.1 is a summary of the distribution in descending order of frequency. In fact, joint multi-senses are not rare: The most frequent multi-sense, *Expansion.Conjunction–Temporal.Synchrony*, is the 17th most frequent sense.

The experimental data are split in the same way as in previous work [105]: sections 2–22 are used as the training set, sections 0–1 as the development set; and sections 23–24 as the test set. In the training phrase of the experiment, probability distributions in the marking model are deduced from the training set of the experimental data. In the testing phrase, the model is applied to predict whether an explicit/implicit DC is likely to be used for each discourse relation in the development set and test set. During evaluation, the predictions are compared with the actual marking in the corpus. Parameters in the model (w_a and w_c) were selected to maximize the prediction accuracy on the development set and the same optimal parameters were used on the test set. For direct comparison

	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
...	...	...	...
20	Expansion	8	89
...	...	...	...
50	Contingency.Pragmatic cause.Justification #Expansion.Instantiation	0	6
...	...	...	...
122	Contingency	0	1
Total		$5,201$	$16,049$

Table 3.1: Sense distribution of explicit and implicit DCs in the experimental data.

with previous work, samples of infrequent DCs and relation senses were excluded from the development and test sets according to the same criteria as in previous work [105]. The resulting development and test sets contain $1,720$ and $1,878$ relations, respectively.

3.3.2 Results

The Discourse Marking Model is applied to predict the speaker's choice of DC marking on the dev and test sets. Table 3.2 shows the results under various settings, evaluated by accuracy and the harmonic mean of precision and recall for explicit and implicit relations respectively.

	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 [105]				–	–	–	.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	**.902**$^{++}$	**.903**	**.901**	.882^{+}	.883	.881
	TSF01	surprisal	prime frequency	.895^{++}	.898	.892	.889^{++*}	**.893**	.885
	TS01	entropy	prime distance	.895^{++}	.900	.889	**.890**$^{++*}$	.892	**.888**

Table 3.2: Accuracies and F1 scores of predicted DC marking. The best values are bolded. (abbreviations: S: full relation sense; TS: top-level sense; F: relation form; SF: sense and form; TSF: top sense and form; 10: previous relation; 20: previous 2 relations; 11: previous relation and next relation) $^{+}$/$^{++}$: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)

Row **BL** shows the results of the Discourse Marking Model without the cost function and argument informativeness component, and with constant context C. This setting is considered 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 marking is strongly related to the intended discourse sense.

Rows in (a) show 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 marking not only depends on surrounding discourse relations but also other contextual factors.

Rows in (b) show the contribution of the *argument informativeness* component, under constant discourse 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. Similar improvement is observed when adjustment with the estimated *argument informativeness* is applied only if the parser output sense is correct (matching at the top level sense).

Rowin (c) show 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 in (d) show 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: the proposed model explains the speaker's choice of DC marking in terms of DC and argument informativeness, as well as production cost, while contextual discourse structure is not a significant constraint on 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 two best combinations (89.03% and 88.92% vs. 86.60%).

Lastly, the results are compared 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 marking of the training set. The classifiers are built by LIBLINEAR [28]. When extracting the argument informativeness features from the training set, using the automatic discourse parser, the parser estimates of the *implicit* samples are penalized by a constant ratio, since the discourse parser is also trained on these samples. The classifier achieves an accuracy of 88.3% on the test set, which does not significantly outperform previous work. This suggests that the information-theoretic configuration is an advantage of the proposed Discourse Marking Model.

3.3.3 Discussion

This chapter describes a computational model that predicts discourse marking in human language production. The model is trained and evaluated using discourse annotation on corpus data as the gold standard. This section discusses the advantages and disadvantages of the methodology.

One advantage of learning the Discourse Marking Model from PDTB is the compatibility of the PDTB's annoation with the RSA framework. Previous applications of RSA focus on the pragmatic use of language, where the *intended message* and *lexicon of an utterance* largely depend on the context. In the task of referring expression generation, the sets of valid referents and referring expressions differ case by case. For example, *red* is an invalid option for referring to a *blue* ball; *he* is an invalid option for referring to a *woman*; and it is difficult to define a finite set of referents in the corpus. In contrast, the usage of DCs is generally universal across different contexts. A DC can be used or dropped to represent various discourse senses in various contexts.

In addition, the PDTB annotation scheme predefines the sets of DCs and discourse relation senses. In this way, the listener and speaker models of RSA can be derived statistically by counting the co-occurence of the DCs, sense labels, and contextual factors in the annotated corpus. Nonetheless, the proposed method to use surrounding discourse relations as context did not improve classification accuracy. Therefore, one direction to improve the proposed model is to make fuller use of the training data to learn a more expressive and general abstraction of the context governing the choice of discourse marking.

On the other hand, the pragmatic reasoning approach of RSA has been criticized for being unrealistic, because previous studies find that speakers tend to produce referring expressions that are overspecifying [7, 22, 27, 34]. In other words, while ideal pragmatic speakers should only focus on the minimum properties that help listeners to identify the referent, the referring expressions that speakers actually choose often include redundant properties that are not necessary for distinguishing the referent from other candidates. In the context of discourse marking, an utterance is overspecifying if an explicit DC is chosen even though an implicit DC is enough for the listeners to infer the discourse relation.

Nonetheless, the proposed model also benefits from an UID-inspired component, such that the prediction on discourse marking does not only rely on pragmatic reasoning of the speakers. The UID component penalizes the choice of explicit DCs when informative signals are present in the arguments.

In addition, learning RSA from corpus statistics allows the model to detect the general trend in the marking of a relation sense. Some relation senses are highly likely to be marked/unmarked irrespective of the presense of other signals, while for other relations, the presense of other discourse signals affects the choice, as illustrated in Examples (5) to (7). In these examples, the speaker probabilities (Equation 3.6) estimated by the best performing model (last row in Table 3.2) are shown along with the predicted marking choices. $P'_s(imp|s, C)$ and $P'_s(exp|s, C)$, which are the speaker probabilities without the UID bias (i.e., $e^{U(arg;s,C)} = 0$, in Equation 3.9), are shown for comparison.

(5) And market expectations clearly have been raised by the capital gains victory in the House last month (Implicit:**since**; CONTINGENCY.CAUSE.REASON) An hour before Friday's plunge, that provision was stripped from the tax bill. *(WSJ2429)*

(without UID) $P'_s(exp|s, C) = 0.042$ $P'_s(imp|s, C) = 0.958$
(with UID) $P_s(exp|s, C) = 0.024$ $P_s(imp|s, C) = 0.976$
Prediction= Implicit

(6) Boeing's offer represents the best overall three-year contract of any major U.S. industrial firm in recent history. **But** (Explicit; COMPARISON.CONTRAST.OPPOSITION) Mr. Baker called the letter ...very weak. *(WSJ2308)*

(without UID) $P'_s(exp|s,C) = 0.813$ $P'_s(imp|s,C) = 0.187$
(with UID) $P_s(exp|s,C) = 0.535$ $P_s(imp|s,C) = 0.465$
Prediction= Explicit

(6) Full-time residential programs ... are particularly expensive – more per participant than a year at Stanford or Yale. (Implicit:**but**; COMPARISON.CONTRAST) Non-residential programs are cheaper, ... *(WSJ2412)*

(without UID) $P'_s(exp|s,C) = 0.649$ $P'_s(imp|s,C) = 0.351$
(with UID) $P_s(exp|s,C) = 0.383$ $P_s(imp|s,C) = 0.617$
Prediction= Implicit

In Examples (5) and (6), the UID bias does not affect the prediction based on DC informativeness alone, since the CONTINGENCY.CAUSE.REASON sense is dominantly implicit (Example (5)) and the textscComparison.Contrast.Opposition sense is dominantly explicit (Example (6)), according to the probability distribution in the corpus. In these cases, argument informativeness has little effect on the RSA model. In contrast, in Example (7), the COMPARISON.CONTRAST sense could be expressed explicitly or implicitly, and the UID bias reverses the prediction based on DC informativeness. The model predicts that the speakers would not overspecify the discourse relation by a DC, since there are enough informative signals in the arguments (e.g., *expensive* vs. *cheaper* or *residential* vs. *non-residential*).

On the other hand, the proposed method approximates the argument informativeness based on the probability output of an automatic discourse parser, and therefore is limitted by the accuracy of the discourse parser, as shown in Example (8).

(7) "Jeux Sans Frontieres"... is a hit in France. (Implicit:**but**; COMPARISON.CONTRAST)
A U.S.-made imitation under the title "Almost Anything Goes" flopped fast. *(WSJ2361)*

(without UID) $P'_s(exp|s,C) = 0.762$ $P'_s(imp|s,C) = 0.238$
(with UID) $P_s(exp|s,C) = 0.624$ $P_s(imp|s,C) = 0.376$
Prediction= Explicit

The parser detects low informativeness in the arguments, and thus the model wrongly predicts that explicit marking is more likely. A possible explanation for this is that the constrast between *hit* and *flopped* is uncommon, and the parser fails to identify it as a discourse-informative signal. The performance of the discourse parser used in the experiment is not yet satisfactory. The accuracy of the marking model could be improved with a more accurate discourse parser.

Lastly, the classifier of the discourse parser may have poorly calibrated probabilities, which means the probability estimates of the parser may not be well associated with how well the parser detects discourse signals. The association, and thus the overall performance of the model, may be improved by an additional probabilistic calibration step on the parser output [98].

3.4 Conclusion

This chapter presents the Discourse Marking Model that predicts a speaker's choice of using an explicit DC or not given the discourse relation s/he wants to express. The model gives an cognitive account of the speaker's choice and also outperforms previous work on the same task.

This chapter presents a language production model that predicts whether a speaker will choose to use an explicit DC or not given the discourse relation they want to express. The model gives an cognitive account of the speakers' choice and its results outperform those of previous work on the same task.

Although the option of DC marking is a subtle preference in the absence of other grammatical constraints, the Discourse Marking Model tackles the option as a rational preference by the speaker. Using an information-theoretic approach, the model predicts a speaker's choice by balancing the advantage (informativeness) and disadvantages (production cost and redundancy) of using an explicit marker.

This is the first work to apply the RSA framework to discourse processing. The universal distribution of utterances and senses are adjusted based on the discourse context. Furthermore, the approach takes 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 a result, the UID principle is incorporated into the RSA framework into a unified model.

Chapter 4

Behavioural Study using Crowdsourcing

The experiment described in Section 3.3 evaluates the prediction ability of the model against the actual data in the PDTB. In other words, the marking of each discourse relation chosen by the writers of the *Wall Street Journa* is taken as the gold standard. However, it is possible that other writers would choose differently, given the same relation sense and context. Behavioral experiments can be designed to compare the judgment of multiple human speakers with the judgement of the annotators of PDTB and writers of the *Wall Street Jounal*, as well as with the model's predictions.

For example, [105] also use a readability judgement task to test speakers' choice of explicit or implicit DCs in PDTB. While their discourse marking classifier accurately (86.6%) predicts the choice of using an explicit or implicit DC given the discourse sense, their judgment study shows that human performance of the task is only 68% accurate, implying that both choices are acceptable in some cases.

This chapter describes an empirical experiment to investigate the marking preference of multiple speakers. A balanced sample of discourses are selected from PDTB and judgements on whether a DC should be dropped are collected from a large number of human subjects.

The primary purpose of this experiment is to evaluate the agreement on DC marking among a group of speakers. On one extreme, if the choice of explicit or implicit DC is totally arbitrary given the linguistic context and relation sense, then all subjects will have the same judgement for each sample. On the other extreme, if the preference of marking is independent of the linguistic context but subject to other external factors, then even distribution of judgements is expected .

Another purpose of the experiment is to compare the 'gold' marking choice in the corpus data of PDTB with the judgements made by the human subjects. If the choice of explicit / implicit DC chosen by the majority of the human subjects does not match with the actual data in PDTB, it implies the limitation of using corpus data to acquire and evaluate a model of speaker's choice of marking.

Last but not least, the predictions of the Discourse Marking Model are compared with the human judgements to further evaluate the performance of the Discourse Marking Model.

This chapter is organized as follows. Section 4.1 first introduces some related experiments on discourse processing . Section 4.2 describe the materials used as stimuli in the experiment and section 4.3 explains the detailed procedure to collect human judgments by crowdsouring. Experiment findings are presented in Section 4.4 and a conclusion is drawn in Section 4.5.

4.1 Related experimental studies on human discourse processing

There is a large body of psycholinguistic research on discourse relation interpretation based on comprehension tasks, e.g. [88, 15, 82, 81, 42, 121, 120, 60, 93]. In these studies, subjects are presented with discourse samples of various conditions, such as different signals or argument order, and asked to judge the relation sense or argument salience.

On the other hand, it is not trivial to design behavioral tasks to access the production of discourse relations. [128] shows subjects short stories in pictures and asks them to reproduce the stories in speech or written form, from which the choice of discourse markers are investigated. This design captures the natural production of discourses by the speakers but the procedure is resource consuming and it is not possible to employ a picture-to-text task in this study since the samples are drawn from PDTB, which is a written resource. It is also difficult to ask subjects to choose an explicit/implicit DC based on the abstract relation sense hierarchy defined in PDTB.

In [105] human judgements of discourse marking are collected by presenting subjects with two versions of discourse samples from PDTB. In one version, the two arguments are joined by an explicit DC, which is the DC actually occur in the PDTB data, if present, or the DC *'annotated'* as implicit DC. In the other version, the two arguments are shown as two sentences without a DC in between. Subjects are asked to judge which version is small natural, given that the sense to be conveyed is the same. The proportion of human judgements that match with the corpus data is reported to be 16% smaller than that of their system's prediction.

The design in [105] basically asserts a comprehension situation and may not be able to fairly judge the production of discourse relations. In fact, the results of their study suggest that it could be the case. Two-third of the judgements that do not match with the corpus data are false positives of explicit DCs, i.e. the subjects prefer an explicit DC while the DC is implicit in the corpus. Psycholinguistics studies show that explicit DCs facilitates the comprehension of discourse relations [88, 120, 60, 93]. Given that the two versions of the sample are of the same meaning, the subjects may tend to choose the explicit version since it is easier to understand.

The marking of discourse relations is also examined using a more production-oriented experimental design, such as the picture-to-language transcription task described in [128]. Another option is to use a cloze test, in which subjects are presented with the discourse arguments and relation sense to be conveyed and are asked to fill in a DC or leave the relation implicit.

4.2 Experimental materials

This section explains the behavioral experiment designed to collect humans' judgments on the preference of producing marked or unmarked discourse relations. Following the recent success in crowdsourced discourse annotation [116, 122], crowdsourcing is used to collect a large number of judgements psuch that a distribution of the marking preference can be obtained.

The proposed experiment is a two-step sentence completion task. In the first step, subjects are asked to complete the sentence with a DC. The purpose of this interpretation step is to induce a discourse relation in their minds. In the second step, which is the target of the investigation, the subjects are asked if the meaning is unchanged even without a DC. For comparison with the corpus data and the model predictions, where each marking choice is subject to the annotated/given discourse sense[1], a marking judgement is considered only if the interpreted sense is the 'correct' sense. All judgements are made in a 5-point scale to capture the production preference in finer granularity. Details of the experiment are described in the following subsections.

100 samples of inter-sentential discourse relations are selected from Sections 0,1,23,24 of PDTB, as shown in Table 4.1. The samples contain equal proportions of the four top categories of relation senses and multi-sense. Half of the samples are explicit DCs and half are implicit DCs. Also, half of the samples are correctly classified by the Discourse Marking Model, of the best performing configuration, and the other half are wrongly classified.

discourse relation sense	Explicit	Implicit	Total	Correct model predictions	Wrong model predictions	Total
Comparison	10	10	20	10	10	20
Contingency	10	10	20	10	10	20
Expansion	10	10	20	10	10	20
Temporal	10	10	20	10	10	20
Multi-sense	10	10	20	10	10	20
Total	50	50	100	50	50	100

Table 4.1: Distribution of discourse samples from PDTB used in the sentence completion task. Same proportion of explicit/implicit relations, relation senses, and instances that are correctly/wrongly classified by the marking model are randomly sampled.

Each sample consists of the arguments[2] of the discourse relation in question and the previous and next sentence as context. The DC of the relation, if any, is replaced by a *blank*. To guide the subjects to ignore punctuation errors, optional periods, commas, and capitalized sentence initials are displayed next to the blank.

Subjects are asked to fill in the blank with three choices of DCs or to leave

[1] This given discourse sense is defined as the *'correct' sense* in the rest of this paper. Relation senses other than the 'correct' sense are called *'wrong' senses*.

[2] Sources of attributions are included but supplementary arguments are excluded.

it blank. One of the DC choices is 'correct' and the other two are 'wrong', unless the relation is annotated with 2 implicit DCs. In the later case, there are two 'correct' choices and one 'wrong' choice. The order of the DC options is randomized. 'Correct' DC is the actually occurring explicit DC, if any, or the annotated implicit DC. 'Wrong' DC is a randomly selected DC always used for a sense different from the 'correct' sense at the top level sense hierarchy.

By mean of this simplified sense interpretation task, the subjects are guided to select the discourse sense to be produced, before they are asked to choose weather the sense should be expressed explicitly or implicitly. Intuitively, this procedure is closer to a production scenario. If only the 'correct' DC is shown, subjects may bias to choose the explicit DC since they know it correctly describes the discourse sense. In the analysis, only judgements in which the 'correct' sense is chosen in the first step are considered. An example of the stimuli is shown in Figure 4.1.

Strictly speaking, these youth are not performing service.
They are giving up no income, deferring no careers, incurring no risk (./,)________ (,) T/they believe themselves to be serving, and they begin to respect themselves (and others), to take control of their lives, to think of the future.
That is a service to the nation.

(a) but (b) since (c) ultimately (d) leave it blank

Figure 4.1: An example of stimulus shown to the human subjects on the crowd-sourcing platform.
The explicit DC 'but' originally occurs in the corpus data.

Subjects are asked to rate the correctness of each option from a 5-point scale, namely *'definitely correct', 'probably ok', 'cannot tell', 'not so good',* and *'definitely wrong'*. The task instruction explicitly tells the subjects to rate the option '(d) leave it blank' correct, if they think the passage has the same meaning with / without their chosen DC.

In the above example, rating of the 'leave it blank' option is collected only if '(a) but', the 'correct' sense, is rated as either 'definitely correct' or 'probably ok', and is rated higher than options (b) and (c)[3]. Without any explicit DCs, the 'correct' sense may not be obvious if the relation is originally explicit, as in the example of Figure 4.1. However, the performance of the sense prediction step is not a focus of this study, and is not comparable to the PDTB annotation since the comprehension conditions are altered.

Multiple judgements are collected for each sample by crowdsourcing. The procedure is described in the next subsection.

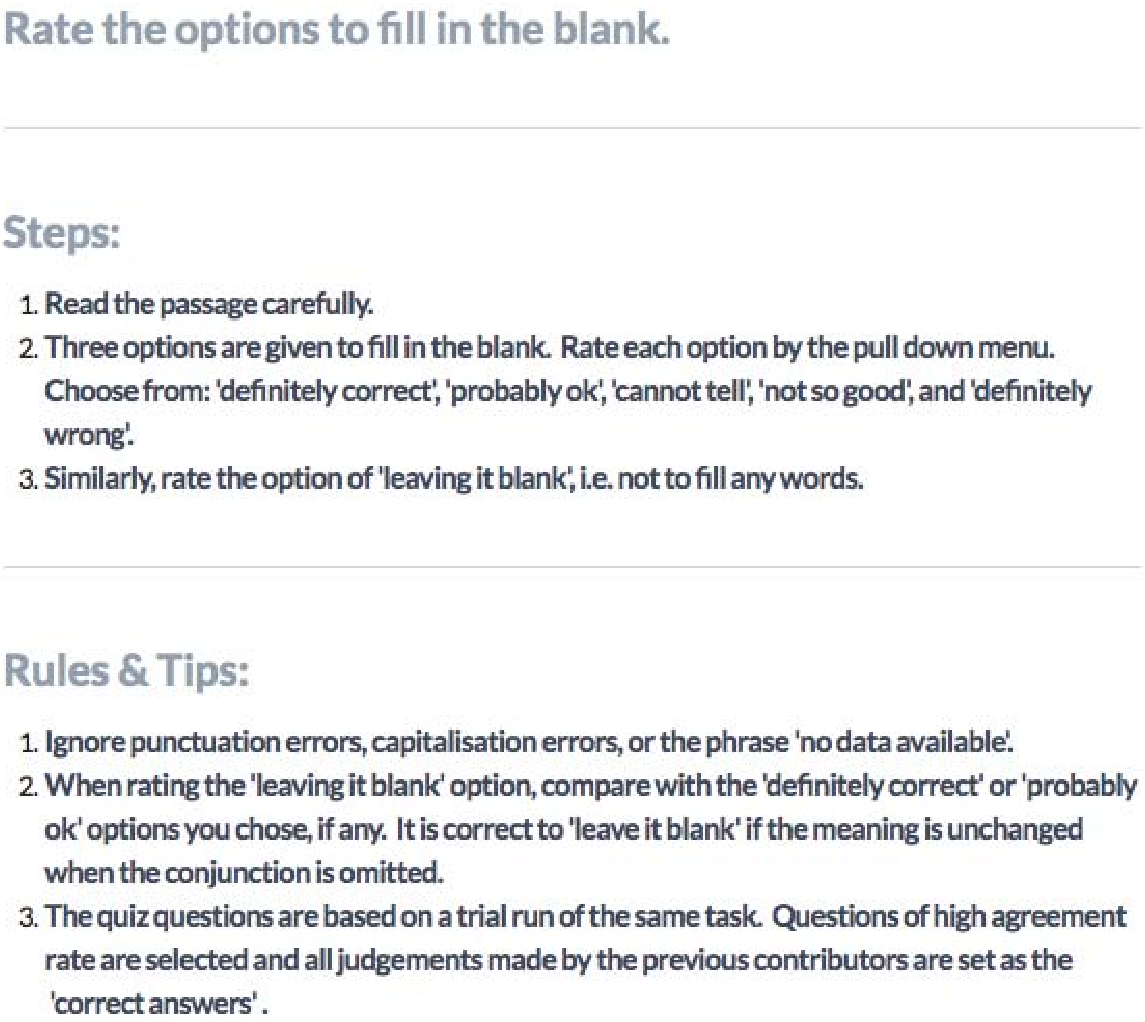
4.3 Procedure

This experiment is carried out on the CrowdFlower platform[4], where judgements are collected by crowdsourcing. In total, 151 English-speaking subjects are recruited, 99 of which are situated in United States and 52 in United Kingdom. Each subject is awarded USD 0.11 to USD 0.14 for each judgement question s/he completes (rating of 4 options). The subjects are not told that the task is a psycholinguistic survey. Figures 4.2 and 4.3 show the exact interface shown to the subjects as seen on a web browser.

Strategies are employed to exclude spams in the crowdsourced judgements. First, all the recruited subjects are 'Level 3 contributors', which means they have a record of nearly perfect performance in their previous tasks on the task platform. In addition, 17 'test questions' are randomly inserted in the task questions. Only judgements by subjects who give valid answers to 75% or more of the 'test questions' are trusted and collected for investigation. 50 trusted judgements per sample (5000 in total) are collected. On average, 56.7 seconds is spent on each trusted judgements. Findings from the judgements are analyzed in the next subsection.

[3] In case of multi-sense, the interpretation is considered correct if any of the multiple senses is rated the highest and positive. On the contrary, if two senses are rated the highest but there is only one 'correct' sense, the interpretation is considered wrong.

[4] https://www.crowdflower.com/



Rate the options to fill in the blank.

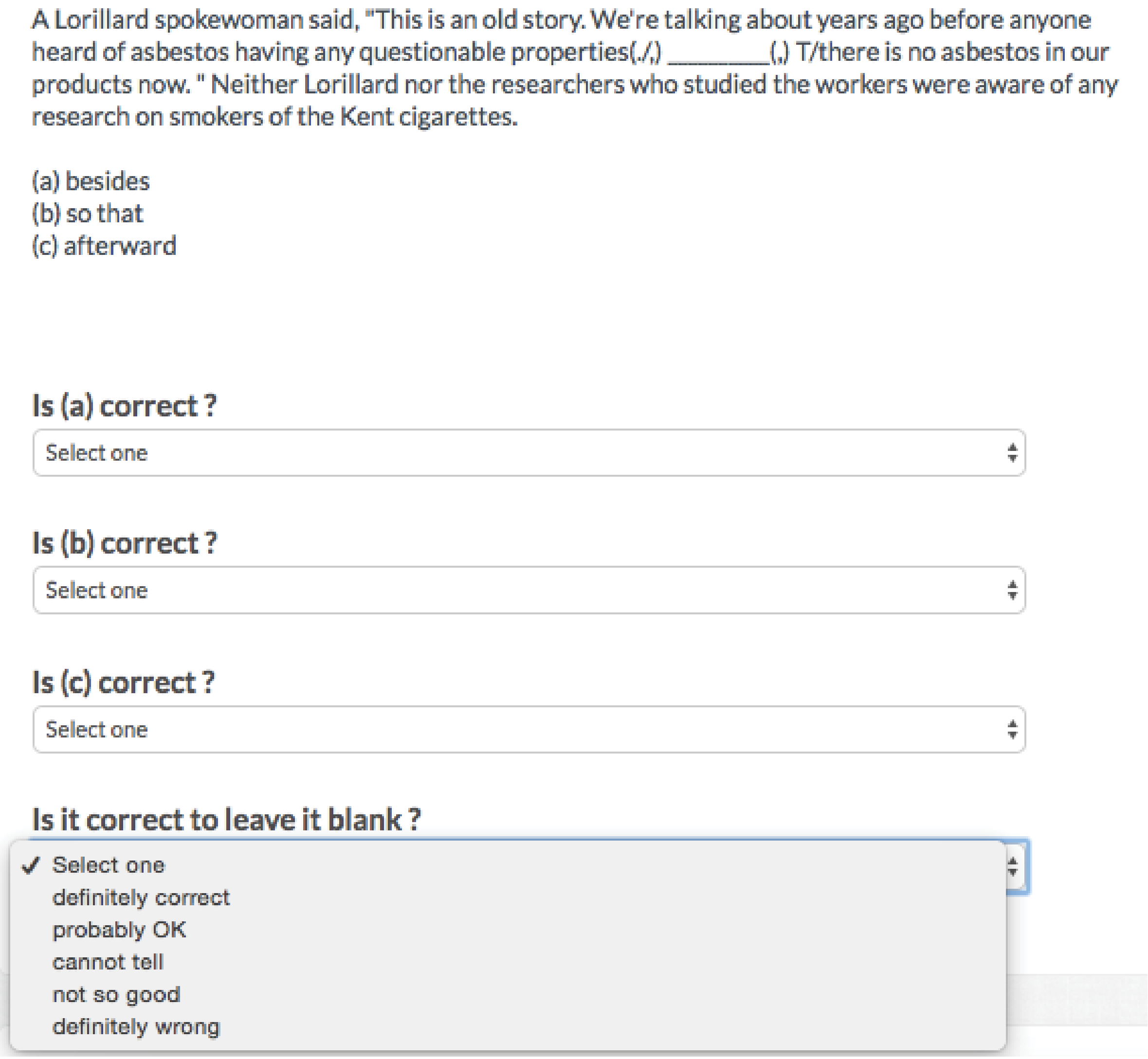
Steps:

1. Read the passage carefully.
2. Three options are given to fill in the blank. Rate each option by the pull down menu. Choose from: 'definitely correct', 'probably ok', 'cannot tell', 'not so good', and 'definitely wrong'.
3. Similarly, rate the option of 'leaving it blank', i.e. not to fill any words.

Rules & Tips:

1. Ignore punctuation errors, capitalisation errors, or the phrase 'no data available'.
2. When rating the 'leaving it blank' option, compare with the 'definitely correct' or 'probably ok' options you chose, if any. It is correct to 'leave it blank' if the meaning is unchanged when the conjunction is omitted.
3. The quiz questions are based on a trial run of the same task. Questions of high agreement rate are selected and all judgements made by the previous contributors are set as the 'correct answers' .

Figure 4.2: Task description as seen to subjects on a web browser.



A Lorillard spokewoman said, "This is an old story. We're talking about years ago before anyone heard of asbestos having any questionable properties(./,) _________(,) T/there is no asbestos in our products now. " Neither Lorillard nor the researchers who studied the workers were aware of any research on smokers of the Kent cigarettes.

(a) besides
(b) so that
(c) afterward

Is (a) correct ?

Select one

Is (b) correct ?

Select one

Is (c) correct ?

Select one

Is it correct to leave it blank ?

✓ Select one
definitely correct
probably OK
cannot tell
not so good
definitely wrong

Figure 4.3: Multiple choice question as seen to subjects on a web browser.

4.4 Results

The target results of the experiment are the ratings given to the 'leave it blank' option of each sample. As explained in the previous subsection, the marking judgement are considered only if the subject has chosen the 'correct' sense in the sense judgement step. Removing judgements with 'wrong' sense interpretation results in 2618 judgements in total, spanning across 100 samples. The accuracy of sense interpretation is considerably low. It is because when explicit DCs are absent, reading of other discourse relations can also be valid. Among the 2618 judgements, 48% are 'gold' explicit relations[5], and 52% are 'gold' implicit relations.

For easier analysis and reference, the rating labels of the 'leave it blank' option are mapped to a scale of DC marking, as per Table 4.2. Higher marking level means the subject prefers an explicit DC more than an implicit DC.

Judgement label 'Is it correct to leave it blank?'	marking level
definitely correct	-2
probably ok	-1
cannot tell	0
not so good	1
definitely wrong	2

Table 4.2: Mapping of the judgement ratings of the 'leave it blank' options to the marking level of DC chosen by the subjects

4.4.1 Agreement on DC marking among human subjects

Table 4.3 lists the counts of various marking levels chosen by the human subjects. About one-third and over half of the judgements choose marking levels -2 and -1 respectively, suggesting that an implicit DC is preferred in most of the cases. In fact, the majority choice of marking level per sample is either -2 and -1 for all the 100 samples, save 1. The averaged proportions of judgements

[5] Relations that are originally explicit/implicit in the PDTB are referred to as 'gold' explicit and 'gold' implicit relations respectively from now on.

choosing -2 and -1 are 53% and 30% respectively. In other words, for each sample, 83% of the subjects agree that the sense can be expressed implicitly.

marking level	'Gold' Explicit	'Gold' Implicit	Total	
-2	315	437	752	(29%)
-1	711	721	1432	(54%)
0	9	7	16	(1%)
1	117	92	209	(8%)
2	116	93	209	(8%)
Total	1268 (48%)	1350 (52%)	2618	(100%)

Table 4.3: Counts of marking ratings chosen by human subjects. Implicit DCs are chosen for most judgements.

The distribution of the marking judgments suggests that people generally agree on the marking level of a given relation. However, irrespective of the relation sense, an implicit DC is preferred by the majority of the subjects. One possible explanation is that the choice of marking level -1 ('probably OK' to drop the DC) actually suggests that both explicit and implicit DCs are acceptable. This accounts for about two-thirds of the total counts of marking level -1 and -2.

The judgements are next compared against the DC actually occurring in PDTB as well as the model predictions, taking each personal judgement as an individual sample.

4.4.2 Comparison of human judgments with corpus data and model predictions

This section presents the comparison between the subjects' judgements of DC marking and the choices of explicit/implicit DCs actually occur in PDTB, also with the predictions made by the Discourse Marking Model. The representativeness of the marking choice in PDTB against each instance of human judgement are examined, instead of the majority vote per sample, which is almost always 'implicit' according to the experimental results.

The marking level judgements are mapped to choices of explicit/implicit DCs directly. marking levels 1 and 2 are choices of explicit DCs and marking levels −1 and −2 are choices of implicit DCs[6]. marking level 0 is considered 'unknown'. The comparison is shown in Table 4.4. Originally, the discourse samples are selected such that half of them are correctly predicted by the Discourse Marking Model and half of them are not. After removing judgements of 'wrong' sense interpretations and counting each judgement individually, 49% of the judgements are correctly predicted by the model, as shown in Table 4.5. The 'accuracies' of different relations senses are shown in Table 4.7, and lastly, the agreement between the model predictions and each instance of human judgements is shown in Table 4.6.

Human Judgements	'Gold' Explicit	'Gold' Implicit	Total	Precision
Explicit	**233**	185	418	55%
Unknown	9	7	16	− − −
Implicit	1026	**1158**	2184	53%
Total	1268	1350	2618	
Recall	18%	86%		Accuracy 53%

Table 4.4: Comparison of marking judgements by the human subjects against the 'gold' corpus data. The true positives are bolded.

Comparison of Tables 4.4 and 4.5 shows that the human judgements and the Discourse Marking Model performs similarly in guessing the DC marking preference in the corpus data, while the 'accuracy' of human judgements outperforms by 4%. Comparing Tables 4.5 and 4.6, it is observed that the model predictions match more with the human judgements than with the corpus data. About 50% of the human judgements and model predictions match with the corpus data, while 66% of the model predictions match with the human judgements. However, these 'accuracies' do not reflect the general performance because the distribution of the samples is not based on the distribution of discourse relations in PDTB.

[6] Analysis is also carried out by mapping the marking levels to explicit/implicit choices by comparing the ratings with the ratings given to the 'correct' DCs. For example, the choice is considered 'unknown' if the rating for 'leave it blank' is positive but not more positive than the ratings to the 'correct' DC. About one-fourth of the implicit choices become 'unknown' under this mapping criteria. Still, the majority vote per sample is 'implicit'.

Model Predictions	‘Gold’ Explicit	‘Gold’ Implicit	Total	Precision
Explicit	**296**	374	670	44%
Implicit	972	**976**	1948	50%
Total	1268	1350	2618	
Recall	23%	72%		Accuracy 49%

Table 4.5: Comparison of automatic prediction by the *marking model* against the ‘gold’ corpus data. The true positives are bolded.

Model Prediction	Subject Explicit	Subject Implicit	Subject Unknown	Total	Precision
Explicit	**109**	558	3	670	16%
Implicit	309	**1626**	13	1948	83%
Total	418	2184	16	2618	
Recall	26%	74%	—		Accuracy 66%

Table 4.6: Comparison of automatic prediction by the *marking model* against human judgements. The true positives are bolded.

	senses					
	Comparison	Contingency	Expansion	Temporal	multi-sense	Overall
Human	56%	49%	52%	45%	58%	53%
Model	48%	50%	53%	44%	50%	49%

Table 4.7: Comparison of ‘accuracy’ of human judgements and model predictions across relations of various senses.

Among the 5 categories of senses, human judgements for relations of the ‘Comparison’ senses and multi-senses particularly outperform the model predictions. It is found that most discrepancies come from extra false positives of explicit DCs predicted by the model for these two senses. A possible explanation is that humans detect discourse signals in the arguments more accurately than the model, which relies on an automatic implicit relation classifier. Based on the detected signals in the arguments, the subjects, but not the model, judge that the relation can be interpreted correctly without inserting a DC.

4.5 Conclusion

To summarize, the experiment results show that the human subjects generally favour to use an implicit DC to present a discourse relation. However, looking the finer ratings of the choice, most subjects only agree that it is ‘probably OK’ to use an implicit DC, suggesting that both explicit and implicit DCs are acceptable in about two-third of the cases. On the other hand, the human judgements match with the corpus data to the similar extent as the model predictions match with the corpus data. In addition, the model predictions match more with the human judgements than with the corpus data, further supporting the appropriateness of the proposed model.

Chapter 5

A Pragmatic Model for the Comprehension of Discourse Relations

Chapter 3 presents the Discourse Marking Model that simulates humans preference of discourse marking in language production. This chapter extends the framework to model how humans interpret the sense of a discourse relation.

According to the RSA model, or the Bayesian Pragmatic framework in general, speakers prefer an utterance that is informative and not costly to produce. In turn, listeners emulate the language production process of the speakers, and expect the speakers to have chosen an informative expression.

Discourse relations have a mixture of semantic and pragmatic properties [138, 73]. The sense of a discourse relation is encoded in the semantics of a DC, yet the interpretation of polysemic DCs and implicit DCs relies on the pragmatic context. During discourse interpretation, listeners expect speakers to have chosen an informative discourse marking strategy, as shown in the examples presented in Chapter 1.

(1a) It was a great movie, **but** I did not like it.

(1b) It was a great movie. **Therefore**, I liked it.

(1c) It was a great movie. I liked it.

(1d)* It was a great movie. I did not like it.

The explicit DCs *but* and *therefore* are less ambiguous and thus more informative, but an implicit DCs (Example (1c)) is also informative enough to express the '**therefore**' sense – the rational listener emulates that the rational speaker thinks with this logic and interprets Example (1c) as a causal relation. This chapter extends the Discourse Marking Model to find out if Bayesian pragmatic approaches are applicable to human comprehension of discourse relations. The RSA model is applied in opposite direction to DC interpretation using a discourse-annotated corpus, the Penn Discourse Treebank. In addition, the proposed model is integrated with a state-of-the-art automatic discourse parser to improve automatic discourse sense classification. It is hypothesized 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.

The model is explained in in Section 5.1, followed by description of experiments in Section 5.2 and conclusion in Section 5.3.

5.1 Model

This section explains how the interpretation of discourse relations is modeled by Bayesian pragmatics, based on the listener model of the RSA model. As described in Chapter 3, the RSA model describes the speaker and listener as rational agents who cooperate towards efficient communication. It is composed of a speaker model (Equation 5.1) and a listener model (Equation 5.2).

$$P_S(w|s,C) \propto \exp(\alpha \cdot U(w;s,C)) \tag{5.1}$$

$$P_L(s|w,C) \propto P_S(w|s,C)P_L(s) \tag{5.2}$$

The speaker chooses an utterance of high utility, i.e. informative and easy to produce. The listener infers the speaker's intended meaning by considering how likely, s/he thinks, the speaker uses that utterance ($P_S(w|s,C)$). The inference is also related to the *salience* of the meaning ($P_L(s)$), a private preference of the listener.

Theoretically, the speaker and listener emulate the language processing of each other in unlimited iterations (i.e. the speaker thinks the listener thinks the speaker thinks..), as shown in Figure 5.1.

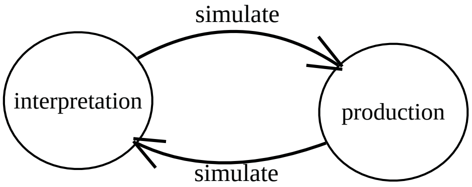


Figure 5.1: Unlimited iterations between speakers and listeners.

However, the inference is grounded on literal interpretation of the utterance due to constraints in lexical semantics. Figure 5.2 illustrates the direction of pragmatic inference between the speaker and listener *in their minds*. Pragmatic listeners/speakers reason for 1 or more levels, but not the literal listener/speaker.

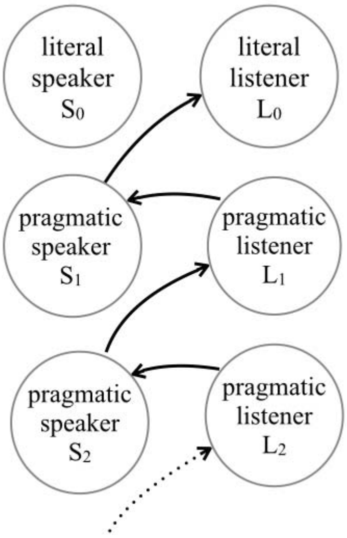


Figure 5.2: Iterative inference grounded on literal interpretation.

Experiments presented in this chapter compare 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 [66, 37].

Specifically, given the DC w and context C in a text, the listener's interpreted relation sense s_i is the sense that maximizes $P_L(s|w, C)$. s_i is defined as

$$s_i = \arg\max_{s \in S} P_L(s|w, C) \tag{5.3}$$

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 PDTB.

$$P_{L_0}(s|w, C) = \frac{count(s, w, C)}{count(w, C)} \tag{5.4}$$

As shown in Figure 5.2, the pragmatic speaker S_1 estimates the utility of a DC by emulating the comprehension of the literal listener L_0 (Equation 5.1). The probability for the pragmatic speaker S_n to use DC w to express meaning s is estimated as:

$$P_{S_n}(d|s, C) = \frac{\exp(\ln P_{L_{n-1}}(s|d, C) - cost(d))}{\sum\limits_{d' \in D} \exp(\ln P_{L_{n-1}}(s|d', C) - cost(d'))} \tag{5.5}$$

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

The cost function in Equation 5.5, *cost*(d), measures the production effort of the DC. The cost of producing *any explicit DC* is simply defined 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. 3.1). The probability for the pragmatic listener L_n to assign meaning s to DC d is estimated as:

$$P_{L_n}(s|d, C) = \frac{P_{S_n}(d|s, C) P_L(s)}{\sum\limits_{s' \in S} P_{S_n}(d|s', C) P_L(s')} \tag{5.6}$$

where $n \geq 1$ and S is the set of defined sense. The salience of a relation sense in Equation 5.6, $P_L(s)$, is defined by the frequency of the sense in the corpus.

$$P_L(s) = \frac{count(s)}{\sum\limits_{s' \in S} count(s')} \tag{5.7}$$

Lastly, the context variable C is defined by the the immediately previous discourse relation to resemble incremental processing. It is hypothesized 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.

5.2 Experiment

This section describes experiments that evaluate the model against discourse-annotated corpus. The purpose of the experiments is 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?

Similar to the Discourse Marking Model, the experiment is based on the annotation of PDTB, and samples labelled with 'No Relation' are excluded. The rest of the samples are distinguished between *explicit DCs* (samples that are labeled *Explicit*) and *non-explicit DCs* (samples that are labeled *implicit*, *alternative lexicalization*, or *same entity*).

Since some relation senses occur very sparsely, the original sense labels (42 distinct senses) are mapped into 15 sense labels (first column of Table 5.2), following the mapping convention of the CoNLL shallow discourse parsing shared task 2015[143]. 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 5.1.

	Train	Test	Total
	Sec.2-22	Sec.23	Sec.0,1,24
Explicit	15,402	3,057	18,459
Non-Exp	18,569	3,318	21,887
Total	33,971	6,375	40,346

Table 5.1: Sample count per data set

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

Table 5.2: 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.
The sum 6525 does not match with Table 5.1 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.

5.2.1 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, the DC interpretation by the literal listener L_0 is compared with that of the 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 $P_L(s|w, C)$. $P_{L_0}(s|w, C)$ is simply based on co-occurrences in the training data (Eq. 5.4). $P_{L_1}(s|w, C)$ and $P_{L_2}(s|w, C)$ are calculated by Eq. 5.5 and 5.6, in which the salience of each sense is also extracted from the training data (Eq. 5.7).

	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 5.3: 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 5.3 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 as in Figure 5.2.

5.2.2 Insights on automatic discourse parsing

Next experiment investigates 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 [143]. 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, the output of an automatic discourse parser is integrated with the probability prediction by the pragmatic listener L_1. As in the argument informativeness component of the Discourse Marking Model, the winning parser of the CONLL shared task [139] is used. The parser is also trained on Sections 2-22 of PDTB, and thus does not overlap with the test set.

For each test sample, the parser outputs a probability estimate for each sense. These estimates are used to replace the *salience* measure ($P_L(s)$) (in Eq. 5.7) and deduce $P'_{L_1}(s|w, C)$, where C is the previous relation form.

$$P'_{L_1}(s|w, C) = \frac{P_{S_1}(w|s, C)P_{parser}(s)}{\sum\limits_{s' \in S} P_{S_1}(w|s', C)P_{parser}(s')} \tag{5.8}$$

Table 5.2 compares the performance of the original parser output and the prediction based on P'_{L_1}. Significant improvement in classification accuracy is achieved and the F1 scores for most senses are improved. This confirms the application potential of information-theoretic approach on automatic discourse parsing.

5.3 Conclusion

This chapter applies the Bayesian pragmatic framework, in opposite direction to the Discourse Marking Model in Chapter 3, to model the interpretation of discourse relations. Experimental results support the applicability of the Bayesian framework on human DC comprehension, that listeners emulate speakers' choice of discourse marking during interpretation. A variation of the experiment also demonstrates the applicability on automatic discourse parsing.

******

The finding in this chapter concludes our understanding of discourse relation marking from the perspective of human language processing, which is the theme of the first half of this dissertation. The second half of the dissertation investigates the marking of discourse relation from a cross-lingual perspective, in the application of machine translation.

Chapter 6

Cross-lingual annotation of discourse relations

The second half of this dissertation examines the discourse marking in translation, which can be viewed as language production in a cross-lingual setting. Strategies to represent discourse relations vary across languages. It is thus a challenging task to correctly translate discourse relations.

In particular, the translation of implicit discourse relations is a noticeable problem when translation from Chinese since implicit discourse relations are abundant in Chinese. According to the statistics of PDTB and the Chinese Discourse Treebank (CDTB), the propotion of implicit DCs in Chinese is signifantly higher than that in English, as shown in Figure 6.1. This implies that certain Chinese implicit DCs are translated to explicit DCs in English. In fact, explicitating discourse relations when translating from Chinese to English is a popular technique used among human translators, but not exploited in current MT system.

A reasonable attempt to learn discourse-relation-aware translation rules is a knowledge-based approach based on an annotated corpus. This chapter describe an effort to annotate discourse relation on the Chinese side and cross-lingually aligning the relations in a Chinese-English translation corpus. Motivated by the characteristics of long Chinese sentences with multiple discourse segments, a novel scheme is proposed to annotate Chinese discourse in sequence instead of the traditional hierarchical structure.

The sequential annotation on the source Chinese corpus is presented in Section 6.1. This is followed by the description of the strategy to align of DCs from Chinese to English in Section 6.2. The annotation statistics are analyzed in Section

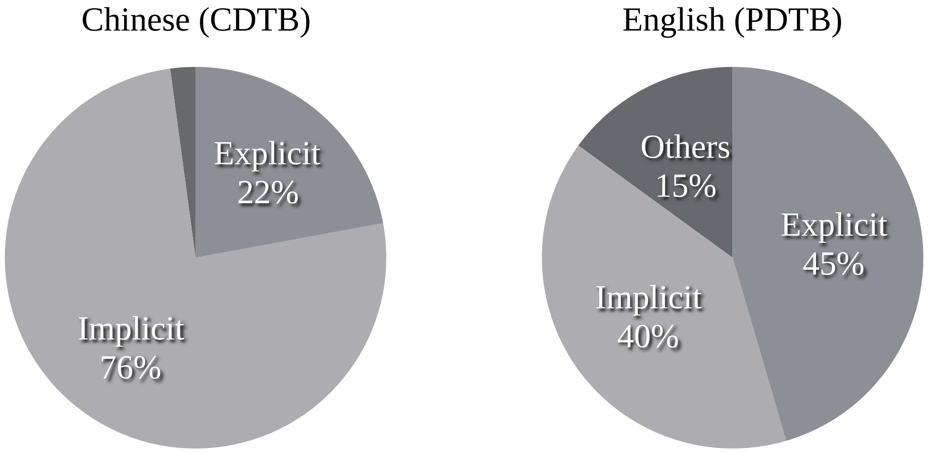


Figure 6.1: Proportions of explicit and implicit DCs as annotated in the PDTB and the CDTB.

6.3, followed by the conclusion in Section 6.4.

6.1 Sequential annotation of Chinese discourse

This section proposes a linguistically driven approach to represent discourse relations in Chinese text as *sequences*. An annotation effort on 325 articles in the Chinese Treebank is conducted. It is observed that certain surface characteristics of Chinese texts, such as the order of clauses, are overt markers of discourse structures, yet existing annotation proposals adapted from formalism constructed for English do not fully incorporate these characteristics.

Section 6.1.1 summarizes the characteristics of Chinese discourse that motivate the design of the annotation scheme, which is described in Section 6.1.2. Lastly, Section 6.1.3 demonstrates an end-to-end discourse chunker is built using this annotation, based on a cascade of classifiers to demonstrate that the annotation is consistent and machine-learnable.

6.1.1 Characteristics of Chinese discourse

Interpretation of discourse relations, as of other linguistic structures, is subject to the surface form of the text. It is noticed that Chinese discourse structures are expressed by certain surface features that do not exist in English.

Paratatic sentence structure

Chinese sentences are sequences of clauses, typically separated by punctuations, and complex Chinese sentences can be as long as paragraphs. This is known as paratatic sentence structure, which are used to represent the temporal or reasoning order or related events, or simply to achieve consistent rhythmic patterns. In contrast, syntactical constraint is prominent in English and this kind of 'paratactic' structures only occur as occasional rhetorical measures.

Each punctuation-separated segment of this kind of 'running sentence' can be considered as an discourse units [146, 154, 76, 102]. On the other hand, discourse structure provides clues to split the source sentence. It is because some DCs only relate discourse units within the same sentences (e.g. *'but', 'because'*) while some only relate with the previous sentence (e.g. *'however', 'in addition'*).In addition, above the clause level, Chinese sentences (marked by the period '。') are larger units of discourse [19]. When presented with texts where periods and commas are removed, native Chinese speakers disagree with where to restore them [12]. The actual sentence segmentation of the text thus represents the spans of discourse arguments intended by the writer and should be taken into account.

Therefore, translation in units of sentences is thus not always preferable in Chinese-English translation. In current SMT models, however, sentence splitting is the result of the language model or translation rules containing periods or sentence initial markers. A long Chinese sentence is typically translated to one English sentence with 'comma splices' (ungrammatical commas between complete sentences without connecting by conjunctions).

Abundant implicit DCs and paired DCs

On top of ambiguous discourse connectives as in other languages, Chinese documents contain abundant implicit connectives. In particular, the sequence of clauses in a long complex sentence is usually separated by commas alone without explicit connectives.

Annotation statistics of in Chinese and English monolingual corpora reveal that the distribution of explicit and implicit DCs are largely different between the two languages. In particular, Chinese discourse units are typically clauses separated by commas, so DCs are often implicit. Explicit and implicit DCs account for 45% and 40% of the DCs annotated in the Penn Discourse Treebank (PDTB) [113] respectively, while in the Chinese Discourse Treebank (CDTB), they account for 22% and 76% respectively [155].

In the annotation procedure of PDTB, implicit DCs are annotated only between sentences, after all explicit DCs are annotated. Since implicit DCs in Chinese also occur within a sentence, it is more effective to annotate both types in one procedure.

In addition, parallel DCs are frequent in Chinese discourse, yet usually either one DC of the pair occurs to signify the same relation [152]. For example, (1) and (2) are grammatical alternatives to (1).

(1) 虽然 (*suiran*, although) $\boxed{\textit{Arg1}}$, 但是 (*danshi*, but) $\boxed{\textit{Arg2}}$.

(2) $\boxed{\textit{Arg1}}$, 但是 (*danshi*, but) $\boxed{\textit{Arg2}}$.

Instead of viewing '虽然 (*suiran*, although) - 但是 (*danshi*, but)' as a pair of parallel DCs, they can be regarded individually as a forward-linking (fw-linking) DC and a backing linking (bw-linking) DC. A fw-linking DC relates its attached discourse unit to a later coming unit, while a bw-linking DC relates its attached discourse unit to a previous unit.

Findings in linguistic studies also show that fw-linking DCs only link discourse units within the sentence boundary. On the other hand, bw-linking DCs can link a discourse unit to a preceding unit within or outside the sentence boundary, except when it is paired with a fw-linking DC [26].

Word order

Syntactical structure is presented by word order in Chinese - so is discourse. While the *Arg1* can occur before or after *Arg2* in English, arguments predominantly occur in fixed order in Chinese, depending on the logical relation. For example, the same concession relation can be expressed by both constructions (3) and (4) in English, but only construction (3) is acceptable in Chinese.

(3) 虽然 (*suiran*, although) $\boxed{\mathit{Arg2}}$, $\boxed{\mathit{Arg1}}$.

(4) $\boxed{\mathit{Arg1}}$, 虽然 (*suiran* ,although) $\boxed{\mathit{Arg2}}$.

According to Chinese linguistics, adjunct clauses and discourse adverbials always precede the main clauses [33, 20]. The clauses are semantically arranged in a topic-comment sequence following the writer's conceptual mind [132, 12]. When the arguments are not arranged in the standard order, the sense of the DC is altered.

For example, when '虽然' (*suiran*, although' is used in construction (2), it represents an 'expansion' relation [49]. Exceptional'inversion' of the order is explained in linguistic literature as 'supplementary materials' or 'after thoughts' in spontaneous speech or stylistic highlights in westernized writing [26, 77]. Therefore, discourse relations should be defined given the order of the arguments, instead of annotating *Arg1* and *Arg2*.

[154] defines *Arg1* and *Arg2* by their semantics, such as *Arg1* for 'reason' and *Arg2* for 'result' in a causal relation. In the resulting Chinese Discourse TreeBank 0.5 (CDTB 0.5), only 2.7% of the relations have *Arg2* preceding *Arg1*.

To summarize, in contrast with the ambiguous arguments in English, punctuations and limitations on DC usage actually mark certain discourse structure in Chinese. An annotation scheme described in the next section is designed based on these characteristics.

6.1.2 Annotation scheme

This section describes an annotation scheme that follows the natural discourse chains in Chinese] According to the scheme, discourse structure is annotated as a sequence of alternating arguments and DCs. This section highlights the main differences of the proposed scheme compared with other frameworks.

Arguments

The main difference of this annotation scheme is that the the order of the arguments for each DC is defined by default. Each clause separated by punctuations except quotation marks is treated as a candidate argument. Clauses that do not

function as discourse units are classified into 3 types - *attribution*, *optional punctuation* and *non-discourse adverbial*.
Since the arguments of a particular discourse relation occur in fixed order and are always adjacent, each argument is related to the immediately preceding argument by a bw-linking DC. In turn, the DC in the first clause of a sentence links the sentence to the previous one, preserving the 2 layer structure denoted by punctuations. An implicit bw-linking DC is inserted if the clause does not contain an explicit DC.

Another characteristic of this annotation effort is that 'parallel DCs' are annotated separately as one fw-linking DC and one bw-linking DC. Implicit bw-linking DCs are inserted , if possible, even the relation is already marked by a fw-linking DC in the previous argument[1].

In other words, duplicated annotation of one relation is allowed. This helps create more valid samples to capture various combinations of Chinese DCs. When an argument spans more than one discourse units, a fw-linking DC is used to mark the start of the span. Similarly, an implicit DC is inserted if necessary.

Connectives

There is a large variety of DCs in Chinese and their syntactical categories are controversial. [49] reports a lexicon of 808 DCs, 359 of which are found in the data. Since many DCs signal the same relation, the proposed approach adopts a functionalist approach to label DC senses.

In the current approach, a DC does not limit to any syntactical category. Annotators are asked to perform a linguistic test by replacing a candidate expression with an unambiguous and preferably frequent DC of similar sense, which is named a 'fine sense' of DC. If the replacement is acceptable, then the expression is identified as a DC and the sense is categorized under the 'fine sense'.

For example, '尤为' and '特别是' (*youwei, tebieshi*, in particular / especially) are categorized under '尤其 ' (*youqi*, in particular), if the annotator agrees that they are interchangeable in the context. Based on the assigned 'fine sense', each DC instant is categorized into the 4 'coarse senses' defined in PDTB: *contingency*, *comparison*, *temporal*, and *expansion*.

[1] Temporal relations are often marked by one fw-linking DC alone and it is not acceptable to insert an implicit bw-linking DC. In this case, the 'redundant' tag is used.

The discourse and syntactical limitations of the DCs are considered in the replaceability test. For example, the following pairs are not labeled the same ‘fine sense’ even the signaled discourse relation is the same:

- Fw v.s. bw-linking DCs:
  虽然 (*suiran*, although), 但是 (*danshi*, but)
- *Cause-result* v.s. *result-cause* order:
  因为... 所以... (*yinwei...suoyi...*, because...therefore...) and
  之所以... 是因为... (*zhisuoyi...shiyinwei...*, the reason why...is because...)
  The two pairs are treated as four different DCs.
- Placed before v.s. after subject:
  却 (*Que* but) and 但是 (*danshi* but)

The list of ‘fine senses’ is not pre-defined but is constructed in the course of annotation; an expression is registered as another ‘fine sense’ if it cannot be replaced. Note that expressions that are considered as ’alternative lexicalizations’ in PDTB or CDTB are also categorized as explicit connectives, if they pass the replaceability test. Otherwise, an implicit DC, chosen from the list of ‘fine senses’, is inserted.

Annotation results

Materials of the corpus are raw texts of 325 articles (2353 sentences) from the Chinese Treebank [102] The annotation is carried out by the MAE annotation tool [131]. Errors that affect the annotation process, namely punctuation errors that lead to wrong segmentation, have been corrected. The annotation is openly released with mapping to the Chinese Treebank[2].

227 DCs are identified in the data, of which 66 are fw-linking DCs. The DCs are categorized into 74 ‘fine senses’ and 22 have ambiguous senses (labelled with more than one ‘fine senses’). The distribution of the tags is shown in Table 6.1. Note that some of the *‘implicit’* relations defined belongs to *‘explicit’* in other annotation schemes since ‘double annotation’ occurs in the current annotation effort.

[2] http://cl.naist.jp/nldata/zhendisco/

	contingency	comparison	temporal	expansion	total
Explicit	380	248	521	683	1832
Implicit	1551	446	164	3022	5183
	adverbial	attribution	optional punctuation		total
Non-discourse	630	783	336		1749

Table 6.1: Distribution of various tags in the annotated corpus

6.1.3 End-to-end discourse chunker

The proposed linguistically driven annotation of discourse structure takes the surface discourse features as ground truth. In particular, discourse relations are defined based on default argument order and span. To demonstrate its learnability, a discourse chunker is designed in the form of a classifier cascade as used in English discourse parsing [79]. Features are extracted from the default arguments of each relation. The accuracy of each component and the overall accuracy of the final output are evaluated, based on classification up to the 4 main senses.

The pipeline consists of 5 classifiers, as shown in Figure 6.2, each of which is trained with the relevant samples, e.g. only arguments annotated with explicit DCs are used to train the explicit DC classifier. 289 and 36 articles are used as training and testing data respectively.

Features include lexical and syntactical features (bag of words, bag of POS, word pairs and production rules) that have been used in classifying implicit English DCs [108, 79], and probability distribution of senses for explicit DC classification. The extraction of features is based on automatic parsing by the Stanford Parser [72]. The surrounding discourse relations are also used as features, based on the hypothesis that certain relation sequences are more likely than others. The classifiers are trained by SVM with a linear kernel using the LIBSVM package[18]. Table 6.2 shows the accuracies of individual classifiers tested on relevant samples. Results based on predictions by the most frequent class are listed as baseline (BL). As expected, implicit relations (IMP) are much harder to classify than explicit relations (EXP). The classification result of non-discourse-unit segments (Non-dis or not) is similar to the preliminary report of [76](averaged F1 88.8, accuracy 89.0%).

The classifiers are then run from Steps 1-5. After Step 1, identified non-

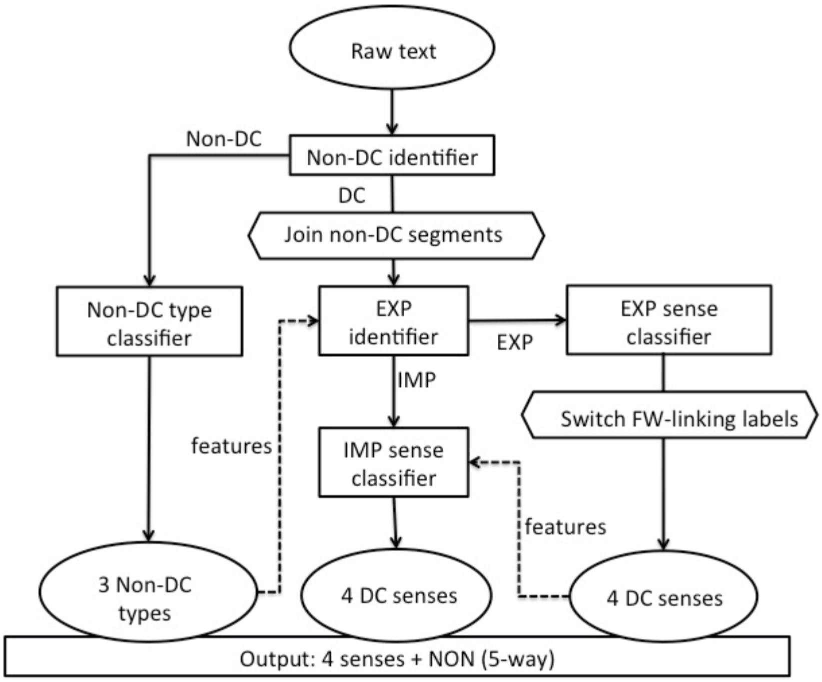


Figure 6.2: Cascade of discourse relation classifiers.

discourse-unit segments are joined as one argument and features are updated. The discourse context features are also updated after each step based on last classifier's output. The tag of a fw-linking DC is switched to the next segment, as a relation connecting the next segment to the current one. The current segment is thus passed to the implicit classifier, given that there is not any bw-linking DCs.

For applications that need discourse, it may not be necessary to distinguish between explicit and implicit relations. Thus, the outputs of the explicit and implicit classifiers are combined when evaluating the end-to-end outputs. Specifically, the pipeline outputs one of the 4 discourse senses or 'non-discourse-unit' across a segment boundary, while the reference can be more than one, since duplicated annotation is allowed. The system prediction is considered correct if it

Step	classifiers	Test F1/Acc	BL F1/Acc
1	Non-dis or not	.91/.94	.44/.80
2	EXP identifier	.92/.93	.39/.65
3	EXP 4 senses	.90/.92	.15/.58
4	Non-dis 3 types	.86/.88	.17/.35
5	IMP 4 senses	.41/.61	.18/.58

Table 6.2: Accuracies of individual classifiers on 'gold' test samples. F1 is the average of the F1 for each class.

is included in the gold tag set. The combined outputs are evaluated in terms of accuracy.

Table 6.3 shows the classification accuracies evaluated by the above principle under different error propagation settings. For example, given gold identification of non-discourse segments (Step 1) and explicit DC classifier (Step 2), classification of the 4 main explicit sense reaches accuracy of 0.854, but is dropped to 0.800 if step 1 and step 2 are automatic[3]. After step 2, the evaluation checks if the segment is correctly classified to one of the 3 types (exp/imp/non-dis). Steps 3-5 refer to the accuracies when counting only the subset of explicit, non-discourse-unit and implicit relations respectively, determined according to gold annotations.

It is observed that errors are generally propagated along the pipeline. Similar to the finding in English [108], the discourse context as predicted by earlier classifiers does not affect the later steps - the results are the same based on gold or automatic outputs. The end-to-end accuracy of the proposed pipeline is 65.7% and the baseline (classify all as 'expansion') is 50.0%.

Finally, different variations of the pipeline are compared, as shown in Table 6.4. The best result (70.1% accuracy), is obtained by classifying implicit DCs and non-discourse units in one step. For comaprison, Huang and Chen [50] reports an accuracy of 88.28% on 4-way classification of inter-sentential discourse senses, and Huang and Chen [51] reports an accuracy of 81.63% on 2-way classification of intra-sentential contingency vs comparison senses.

[3] Note that the results under the complete gold settings do not necessarily echo the results of the individual components, where duplicated outputs are counted individually.

	Accuracies					
	non-dis	exp/imp	explicit	non-dis	implicit	over
	or not	/non-dis	senses	types	senses	-all
Step	2-way	3-way	4-way	3-way	4-way	5-way
4	Gold	Gold	Gold	Gold	.670	.706
3	Gold	Gold	Gold	.879	.670	.706
2	Gold	Gold	.854	.879	.670	.703
1	Gold	.888	.800	.865	.665	.697
-	.862	.847	.800	.836	.657	.657

Table 6.3: Accuracies at each stage under different error propagation settings.

Pipeline variations	Overall 5-way acc.
steps 1-5	.657
combine steps 1-5	.549
switch steps 1 & 2	.697
switch steps 1 & 2 + combine steps 4&5	**.701**

Table 6.4: 5-way accuracies of modified pipelines

To summarize, the sequential annotation is learnable. In addtion, the result is much degraded if one 5-way classifier is trained to classify all relations. This shows that explicit and implicit DCs ought to be treated separately, even though we do not concern about distinguishing them in the final output.

6.2 Alignments of discourse connectives

To investigate how DCs are translated from Chinese to English, the source Chinese DCs are aligned to their translations on a parallel corpus. The target DCs are also annotated with their nature and senses. This section describes the strategy and findings of this annotation. As mentioned in the last section, the corpus comes from 325 newswire articles (2353 sentences) of the the Chinese Treebank, which are actually the 325 articles with English translation [102, 11].

Using the translation spotting technique [91], both the explicit and DCs are aligned cross-lingually. Annotation is carried out on the raw texts. The *explicit* and *implicit* labels used on the source annotation are also used to tag the target English texts. In addition, two other labels are used in the cross-lingual annotation, as defined in the following:

- **Redundant**: The 'redundant' tag is used when it is not grammatically acceptable to insert an implicit DC. Typically, it is annotated on either side of a DC alignment. For example, either half of a pair of parallel Chinese DCs (e.g.'因为'*because*...'所以'*therefore*) is aligned to 'redundant', as it is not grammatical to use both DCs in English. The tag is also used on the Chinese side when a bw-linking DC cannot be inserted, typically for the *temporal* relation.

- **AltLex**: 'AltLex' refers to the 'Alternative lexicalization' of a discourse relation that cannot be isolated from context as an explicit DC, e.g. '*it was followed by*' for a *Temporal* relation. Prepositions that mark discourse relations are also labeled 'AltLex', such as '*through*' for a *Contingency* relation. This label is defined on English side only.

The discourse sense annotation and DC alignment are carried out at one pass by below procedure:

1. Explicit DCs are identified in the source Chinese sentence, and labeled with sense tags.

2. The English translation of the DC is spotted, aligned to the Chinese DC and labeled with sense tags.

3. If the Chinese DC is not translated to an English DC, the annotator first looks for 'Alt-Lex'. If no 'AltLex' can be identified, an implicit DC is inserted. If insertion is not grammatical, the DC is aligned to 'redundant'.

4. On the Chinese side of the corpus, implicit DCs are inserted between two discourse units if they are not related by an explicit DC. Each component of a paired DC is treated independently: when only half of a paired DC occurs explicitly, the other half is inserted as an implicit DC. The implicit DC is aligned following the strategy in Step 3.

5. Any explicit DCs on the English side that are not aligned are identified. Further implicit DCs are inserted to the Chinese side for alignment. If insertion of implicit DCs is ungrammatical, they are aligned to ‘redundant’.

Each pair of aligned DCs are thus tagged with 8 labels. The meaning of the *nature* labels are summarized in Table 6.5, and some annotation examples are shown below.

Example 1

中国必须对国有企业进行改革, [**1**] 加强本身的竞争力。
China must implement reforms on state-owned enterprises so as to [**1**] improve its own competitiveness. .

	Chinese	English
[**1**]nature:	implicit	explicit
actual DC:	*nil*	so as to
fine sense:	来	in order to
coarse sense:	*Contingency*	*Contingency*

Example 2

[**1**] 在投 资项目上比上年减少四百四十四件,但 [**2**]投 资金额却 [**3**]比上年增加一点三亿多美元。

[**1**] The number of investment projects dropped by 444 as compared with last year, but [**2**] the value of investments [**3**] rose by more than 130 million as compared with last year.

	Chinese	English
[**1**]nature:	implicit	implicit
actual DC:	*nil*	*nil*
fine sense:	其实	in fact
coarse sense:	*Expansion*	*Expansion*
[**2**]nature:	explicit	explicit
actual DC:	但	but
fine sense:	但是	but
coarse sense:	*Comparison*	*Comparison*
[**3**]nature:	explicit	redundant
actual DC:	却	*nil*
fine sense:	却	*nil*
coarse sense:	*Comparison*	*nil*

Nature tags for aligned 'DC'		
Chinese	English	
Explicit	**Explicit**	explicit DC identified
Implicit	**Implicit**	implicit DC insertable
-	**AltLex**	expressions alternative to DC
Redundant	**Redundant**	ungrammatical to insert DC

Nature tags for Non-EDU Chinese segments	
Attribution	source of attribution
Adverbial	adverbial initialized
Optional	optional comma for a rhythmic pause

Table 6.5: Tags for Chi-Eng DC annotations

6.3 Corpus analysis

In total, 7266 pairs of discourse relations are aligned. 227 Chinese and 152 English DCs, and 74 Chinese and 75 English fine senses are identified.

Table 6.6 shows the number of unique DCs and fine senses that are identified in the annotation process. A smaller variety of DCs are used in the English translation than the Chinese source. The number of fine senses recognized in implicit DCs is smaller than that of explicit DCs, implying that some fine senses are only expressed explicitly.

Table 6.7 and 6.8 shows the distribution of coarse DC senses on the two sides of the corpus respectively. Similar to the findings in PDTB and CDTB, there are more implicit DCs than explicit DCs on the Chinese side but they are of similar proportion in English.

Comparison, *Contingency*, and *Expansion* relations are more often expressed by implicit DCs than explicit DCs in Chinese. On the other hand, *Contingency* and *Expansion* relations are more often expressed by implicit DCs than explicit DCs in English. Similar tendency is found in the PDTB. In CDTB, among the 9 coarse senses, *Causation*, *Entailment*, *Expansion* and *Conjunction* relations are more often implicit than explicit.

Exp.	COM	CON	EXP	TEM	Total
Chi.	30(11)	63(18)	72(26)	62(19)	227(74)
Eng.	20(11)	41(13)	55(23)	40(14)	156(61)
Imp.	COM	CON	EXP	TEM	Total
Chi.	−(9)	−(15)	−(17)	−(13)	−(54)
Eng.	−(7)	−(11)	−(12)	−(9)	−(39)

Table 6.6: DCs and DC fine senses (in brackets)
DCs and fine senses that have multiple course senses are counted as different DCs/senses. (If counted only once, the total numbers of unique DCs and DC fine senses (in brackets) are: explicit-Chinese: 200(70); explicit-English: 139(56); implicit-Chinese: (52); implicit-English: (38))

Chinese	Explicit	Implicit	Total
Comparison	248 (36%)	446 (64%)	694 (9.9%)
Contingency	379 (20%)	1551 (80%)	1930 (27.5%)
Expansion	683 (18%)	3022 (82%)	3705 (52.8%)
Temporal	522 (76%)	165 (24%)	687 (9.8%)
Total	1832 (26%)	5184 (74%)	7016

Table 6.7: Proportion of various DCs per coarse sense. On top of above, there are 250‘redundant’ cases.

English.	Explicit	Implicit	AltLex	Total
Comparison	287 (51%)	274 (48%)	6 (1%)	567 (9.3%)
Contingency	308 (25%)	584 (47%)	338 (27%)	1230 (20.3%)
Expansion	1545 (42%)	1927 (52%)	218 (6%)	3690 (60.8%)
Temporal	408 (70%)	108 (19%)	63 (11%)	579 (9.5%)
Total	2548 (42%)	2893 (48%)	625 (10%)	6066

Table 6.8: Proportion of various DCs per coarse sense. On top of above, there are 1200 ‘redundant’ cases.

Table 6.9 shows the number of alignments between discourse relations of different nature. Among the 5184 implicit DCs in Chinese, about 70% are not explicitly translated in English (2812 aligned to implicit DCs and 775 to 'redundant'). The rest 30% are translated to explicit DCs or other explicit lexicalization in English. The crosslingual alignment of discourse senses will be further examined in in Section 7.3.2 of Chapter 7.

English / Chinese	Explicit	Implicit	Redundant	Total
Explicit	1332	1193	23	2548
Implicit	81	2812	0	2893
Redundant	198	775	227	1200
AltLex	221	404	0	625
Total	1832	5184	250	7266

Table 6.9: Number of alignments between discourse relations of different nature

6.4 Conclusion

This chapter presents a novel scheme to annotate Chinese discourse structure in sequence and to align 7266 discourse relations of different nature from Chinese to English in a translation corpora. The statistics shows the divergence in DC usage between Chinese and English. It suggests that certain implicit Chinese DCs are explicitated in the English translation. To correctly model the translation of implicit relations, do we need a discourse parser that classifies an implicit source DC to its fine sense or coarse sense? Or will SMT robustly handle implicit-to-explicit DC translation without any discourse preprocessing? We seek to answer these questions in the next chapter.

Chapter 7

Machine Translation of Implicit Discourse Relations

Using the annotated resource presented in the Chapter 6, this chapter investigates how discourse relations should be tackled in MT systems, in particular, how implicit discourse relations should be translated from Chinese to English.

The corpus analysis in Chapter 6 reports that the marking of discourse relations varies largely between Chinese and English languages. Comparing with other language pairs, such as Arabic and English, it is found that discourse factors impact MT quality more in Chinese-to-English translation, especially when translating discourse relations that are expressed implicitly in one language but explicitly in the other [74].

When translating from Chinese to English, implicit DCs are explicitated when necessary. For example, a causal relation can be inferred between the 2 clauses of the Chinese sentence below. In the English translation, the 2 clauses should be connected by an explicit DC, such as 'thus'.

(1) ^{1}[出口快速增长] ,
export grows rapidly
^{2}[成为推动经济增长的重要力量。]
become important strength in promoting the economy to grow.

An open question in discourse for SMT is how best to handle cases where DCs are implicit in the source (e.g. Chinese) but explicit in the target (e.g. English). This chapter investigates how implicit DCs are translated in a translation corpus, and if explicitating implicit DCs in the source can improve MT.

With an automatic discourse parser, a discourse-tree-to-string translation model can be built. Nonetheless, state-of-the-art accuracy of implicit discourse sense classification is still low for downstream application. To examine the MT of implicit relations without bias on discourse parsing performance, oracle experiments are designed to evaluate the MT of implicit DCs assuming that the gold discourse sense is given.

The experiment setting is explained in Section 7.1, followed by results and analysis in Sections 7.2 and 7.3. A conclusion of this chapter is drawn in Section 7.4.

7.1 Methodology

In the cross-lingual annotation, implicit DCs senses are defined by DCs that are identified during explicit DC annotation. In other words, the implicit DCs are represented by explicit DC that acturally occur in Chinese discourse. It is hypothesized that explicitating implicit DCs in the source based on manual annotation will improve implicit-to-explicit DC translations and thus the overall MT result.

Hence, the annotated corpus is used as the *test set* for the MT experiments. The source input is preprocessed based on the manual DC annotations. A number of variations of the preprocess are compared:

- **Implicit fine sense** : The annotated lexicalized fine sense is inserted to the source text. For example, referring to Example 2 in Section 6.2, '其实 ('in fact') ' is inserted at position [**1**] in the source sentence.

- **Implicit coarse sense**: Classification up to the coarse discourse sense could be helpful enough to translate the implicit DCs. The most frequent fine sense of the annotated coarse sense is inserted to the source text[1]. Referring to the same example, '而且' ('and') is inserted at position [**1**] because it is the most frequent fine sense under the coarse sense *Expansion.*

- **Most explicitated DCs** : According to findings in translation studies, explicitation of DCs is DC-dependent [157]. The input source text is thus

[1] The top frequent DCs per coarse sense for *Expansion, Comparison, Contingency* and *Temporal* relations are '而且' ('and'), '但' ('but'), '然后' ('then'), and '从而' ('thus') respectively.

preprocessed by explicitating only the N most frequently explicitated implicit DCs (implicit in source but explicit in target) according to the manual annotation[2]. Referring to the same example, no DC is inserted at position [**1**] because the annotated fine sense '其实' ('in fact') is not within the top four.

- **Same DC for all implicit relations**: To evaluate the effect of inserting explicit DCs to the source text independent of the discourse sense, the most frequently explicited DC, '而且' ('and') are homogenrously inserted to all positions where an implicit DC is annotated in the source text. Therefore, '而且' is inserted to position [**1**] of both Example 1 and Example 2 in Section 6.2 under this setting.

Four kinds of preprocessing are compared to see what kind of explication of implicit DCs could improve MT. For each of the 4 kinds of preprocessing, there is also an additional variant 'implicit-to-explicit only' (imp-exp), which restrictively explicitate only those DCs that are actually aligned to explicit target DCs. This is to evaluate the importance of identifying which implicit DC has to be explicitly translated. Referring to Example 2, no DC is inserted to position [**1**] since it is not an 'implicit-to-implicit' alignment. These various versions of source texts are decoded by SMT systems.

Baseline MT systems are trained with 2.5 million sentences of bitexts through the LDC[3], including newswire, broadcast news and law genres. To see if there is any bias of DC translation to certain framework, three types of SMT systems are built with default settings: a phrase-based model and a hierarchical model using MOSES [64], and a tree-to-string model using TRAVATAR [97]. All models use a 5-gram language model trained on the English Gigaword [104] and are tuned by MERT [99]. GIZA^{++} [100] is used for automatic word alignment and the Stanford Parser [72] to parse the source text for tree-to-string MT training. Tuning and testing with the newswire portions of OpenMT08 and OpenMT06 respectively, the phrase-based, Hiero and tree-to-string systems yield BLEU scores of 26.7, 26.1 and 20.4 respectively, evaluating against 4 reference translations.

These SMT models are used to translate the source text in which implicit DCs

[2] The 4 most often explicitated fine senses are used, which are '而且' ('and'), '而' ('whearas'), '和' ('and'), '并' ('also').

[3] LDC2004T08, LDC2005E47, LDC2005T06, LDC2007T23, LDC2008T08, LDC2008T18, LDC2012T16, LDC2012T20, LDC2014T04, LDC2014T11, LDC2014T15

are explicitated by the methods described above in this section. 1178 sentences and 1175 sentences of the manually annotated parallel corpus are used as the tuning and test sets respectively. The systems are tuned with the tuning set preprocessed by the *implicit fine sense* method.

Note that the SMT training data is not discourse annotated and thus the translation models are not trained with any discourse markups. Nonetheless, the source side of the training data contains abundant examples of both implicit and explicit DCs. It is believed that the translation model will contain translation rules for both natures. The question is whether explicitating implicit DC senses in the source input will the improve final performance.

7.2 Results

Figure 7.1 shows the BLEU and METEOR scores of the SMT outputs resulting from various preprocessed test sets. Explicitation of implicit DCs in the source input generally results in evaluation scores comparable to that of the unprocessed input. Similar results are produced by the 3 SMT frameworks. Only the **SAM** preprocess results in higher evaluation scores using Hiero SMT.

Unexpectedly, disambiguating the implicit discourse sense up to the fine sense does not yeild better translation comparing with disambiguation up to the coarse sense. In turn, homogenously inserting '而且' ('and') without sense disambiguation yeilds even better result. Similar scores are produced by explicitating only the most frequently explicitated implicit DCs. The 'implicit-to-explicit only' restriction generally produces higher scores, suggesting that it is crucial to identify which DCs should be explicitated in translation and which should not.

Results of the oracle MT experiment show that MT performance is hardly improved by explicitating implicit DCs even based on manual annotation. It will be more difficult to improve MT based on predicted implicit discourse senses.

7.3 Analysis

The negative MT results could be due to the following possibilities: (1) Improvement of DC translation is not captured by automatic evaluation scores. (2) The sense of the implicit DCs that requires explicitation is unevenly distributed,

	PBMT		Hiero		T2S	
	B	M	B	M	B	M
original	**15.6**	**24.5**	15.6	24.4	**12.6**	**22.7**
implicit fine sense	15.5	24.4	15.3	24.4	12.3	22.6
implicit fine sense+imp-exp	**15.6**	24.4	15.6	24.4	12.4	22.6
Implicit coarse sense	15.4	**24.5**	15.4	24.4	12.4	**22.7**
Implicit coarse sense+imp-exp	15.5	24.4	15.5	24.4	12.5	22.6
Most explicitated DCs	**15.6**	**24.5**	15.6	24.5	12.5	22.6
Most explicitated DCs +imp-exp	**15.6**	24.4	15.6	24.4	12.5	**22.7**
Same DC	15.4	**24.5**	**15.7**	**24.6**	12.4	**22.7**
Same DC+imp-exp	15.5	24.4	15.5	24.4	12.4	**22.7**

Table 7.1: BLEU (B) and METEOR (M) scores of MT outputs resulting from various DC insertions. Highest scores of each SMT system are bolded

such that disambiguating the sense has limited effect. (3) The context in which a discourse relation is expressed explicitly in the source largely differs from the context in which it is expressed implicity. As a result, translation rules of actual explicit DCs cannot correctly translate artificially expliciated DCs.

These possibilities are analyzed in the following sections.

7.3.1 Is the translation of implicit-to-explicit DCs improved?

Since DCs contribute to a small portion of word counts in the MT output, the difference in DC translation is not sensitive to global n-gram-based evaluation metrics. Translation of DCs can be actually improved while BLEU scores remain similar [90].

100 sentences of the baseline Hiero output, the reference translation, as well as the Hiero MT outputs produced by the preprocesses (most explicited DC method, with and without 'imp-exp' restriction) are manually analized It is done by spotting how each implicit source DC is translated - to which explicit DC or not translated as explicit DC. Table 7.2 shows the proportion of different DC alignments produced by different MT systems and the reference translation.

Part (1) of Table 7.2 compares the rate in which implicit source DCs are explic-

(1)	**implicit-to-explicit rate**
Ref.	19%
Original	23%
Most explicitated DCs	73%
Most explicitated DCs+imp-exp	33%

(2)	**correct**	**incorrect**
Original	22%	78%
Most explicitated DCs	23%	77%
Most explicitated DCs+imp-exp	48%	52%

(3)	**insert=explicit**	**nil=explicit**
Most explicitated DCs	90%	10%
Most explicitated DCs+imp-exp	44%	56%

(4)	**correct**	**incorrect**	**correct**	**incorrect**
Most explicitated DCs	25%	75%	6%	94%
Most explicitated DCs+imp-exp	97%	3%	9%	91%

Table 7.2: Comparison of implicit DC translations in different preprocessing schemes

itated in the translation outputs. As expected, more implicit DCs are translated explicitly in the output of the preprocessed source text than that of the original source text. However, the original output already explicitates more implicit DCs than the reference does.

Part (2) of the table shows how much of the target DCs aligned to (originally) implicit source DCs are correct translation. The explicit target DC is considered **correct** if it matches with the explicit DC in the reference translation, and **incorrect** if the explicit DC is different from the reference DC or the relation is not translated as an explicit DC in the reference. It is seen that the preprocess (23%) hardly improves the accuracy compared with the original output (22%), unless only source DCs that are known to be explicitly translated are explicitated (48%).

Part (3) of the table shows how often explicitating source DCs actually produces explicit DC translations. '**insert=explicit**' means the target explicit DC is aligned to a source explicit DC inserted by preprocess. '**nil=explicit**' means the target explicit DC is not aligned to any source DCs (inserted or not). It is observed that implicit DCs are sometimes explicitly translated by the MT systems even without source explicitation, yet the translation accuracy is low, comparing with translation from explicitated source DCs, as shown in Part (4) of the table.

Result of this analysis supports the hypothesis that the improvement in implicit-to-explicit DC translation is not captured by MT evaluation metrics. Although the MT outputs under comparison have similar scores, implicit-to-explicit DC translation is improved under the *Most explicitated DCs+imp-exp* setting, but not under the other settings. In addition, the result suggests that certain implicit-to-explicit DC translation is captured by SMT even without source explicitation preprocessiing.

7.3.2 Which senses are more common in implicit-to-explicit aligments?

On average, 18.5 Chinese and 15.25 English fine senses are identified under each of the 4 coarse senses. Nonetheless, the oracle MT experiment suggests that classifying the implicit discourse senses more precisely does not improve MT more. A possible explanation is that the senses of implicit-to-explicit DCs only limit to a small set of senses that are already captured by coarse sense classification.

Among the 7266 aligned relations, there are 1193 implicit-explicit alignments (refer to Table 6.9). Table 7.3 shows the sense distribution of these pairs. While the sense distribution on the Chinese side is comparable to the overall sense distribution (refer to Table 6.7), over 80% of which are translated by explicit DCs that signal an *Expansion* sense. In fact, 88% of the implicit source DCs are aligned to the explicit target DC '*and*'.

Coarse sense	Chinese		English	
Comparison	131	11.0%	90	7.5%
Contingency	300	25.1%	109	9.1%
Expansion	715	59.9%	958	80.3%
Temporal	47	3.9%	36	3.0%
Total	1193		1193	

Table 7.3: Sense distribution of imp.-exp. DC

source implicit fine sense	target explicit DC	count	(coverage)
而且 'and'	and	203	(17%)
而 'whearas'	and	117	(15%)
和 'and'	and	139	(12%)
并 'also'	and	81	(11%)
从而 'thus'	and	61	(7%)
所以 'therefore'	and	46	(5%)
来 'in order to'	and	26	(4%)
因此 'therefore'	and	23	(3%)
然后 'and then'	and	18	(2%)
即 'which is'	and	18	(2%)

Table 7.4: Top 10 frequent imp.-exp. alignments

Table 7.4 lists the top 10 frequent implicit-explicit alignments. It shows that '*and*' is used to explicitate a range of discourse relations. On the other hand, although '*and*' ambiguously signal various senses, non-*Expansion* senses only occur marginally in PTDB, as shown in Table 7.5. The distinct discrepancy suggests that DC usage differs between spontaneous writing and translation.

sense of explicit *‘and’*		count	(coverage)
Conjunction	(expansion)	2543	(85%)
result	(contingency)	38	(1%)
Conjunction and *result*		138	(5%)
others		281	(9%)
sense of implicit *‘and’*		count	(coverage)
Conjunction	(expansion)	891	(70%)
List	(expansion)	346	(27%)
others		35	(3%)

Table 7.5: Sense distribution of DC *‘and’* in PDTB.

Analysis of the implicit-explicit alignments explains why more precise sense disambiguation of the source relations does not improve MT. It is because the reference translation uses *‘and’* as the *‘wild card’* to translate most implicit DCs ‘explicitly’, but without explicitating the discourse sense. This finding is similar to the analysis based on word-aligned Chinese-English translation corpus, which also reports that *‘and’* is the most frequently added DC to the reference translation [75]. Therefore, to improve implicit-to-explicit DC translation, an additional task should be defined to identify whether a source implicit DC is kept implicit, explicitly translated to an ambigous DC such as *‘and’*, or explicitly translated to other unambiguous DCs.

Generally, it is pragmatically correct to use *‘and’* to translate an implicit discourse relation, or to keep the relation implicit as in the source. Nonetheless, repetatively using this stragegy will result in excessively long sentences, as in the Example (2) below. In this case, insertion of explicit DCs to the target text is desirable, instead of duplicating the source writing style.

Example (2) - Source

^{1}[天津港保税区投入运行五年来，] ^{2}[已建成了中国第一货物分拨中心，] ^{3}[具备了口岸 关的功能，] ^{4}[开通了天津港保税区经西安、兰州到新疆阿拉山口口岸的铁路专用线；]^{5}[建立了一批集仓储、运输、销售于一体的大型物流配给中心，] ^{6}[开办了铁路和国 际集装箱多式 联运，] ^{7}[月接卸集装箱能力达六千标准箱；] ^{8}[形成了七千门程控 电话的装机能力，] ^{9}[供 电能力达二点五万千伏、日供水能力一万吨。]

Example (2) - Reference

^{1}[Since being put into operation five years ago,] ^{2}[the Tianjin Port Bonded Area has completed the construction of China's first goods distribution center,] ^{3}[functioned like a customs port,] ^{4}[opened up the special use the railway line from the Tianjin Port Bonded Area passing Xi'an and Lanzhou to arrive at Xinjiang's Allah Mountain pass customs port,]^{5}[established a number of large-scale materials circulation distribution and supply centers integrating storage, transportation and sales,] ^{6}[opened multiple railway and international container joint-operations] ^{7}[with a monthly loading and unloading capacity reaching 6,000 standard containers.] ^{8}[It has built up an installation capacity of 7,000 sets of program-controlled telephones,] ^{9}[with a power supply capacity of 25,000 kilovolts, and a daily water supply capacity of 10,000 tons.]

7.3.3 Contexts of explicit/implicit DC usage

Lastly, the contexts in which a particular sense is expressed explicitly or implicitly in the source are compared. If the contexts are distinctly different, it suggests that artificially explicitated source implicit DCs cannot be captured by a translation model trained only with naturally occuring explicit DCs.

In addition, the contexts in which a source implicit DC is translated into an explicit DC or by other means (by implicit DC or alternative lexicalization) are compared. If the contexts are similar, it suggests that the translation strategy could be an option independent of the context.

Following Rutherford and Xue [118], the context of a discourse relation is defined as the unigram distribution of words in the 2 arguments connected by the relation. The context of a particular discourse usage is thus the sum of the unigram distributions of all discourse relations associated with that usage. The Jensen-Shannon Divergence (JSD) is used to evaluate the similarity of the contextual distributions [118, 53, 68]. This metric compares 2 distributions with the average. If both distributions are close to the average, it means they are close to each other as well. The metric value ranges from 0 (identical) to $\ln 2$.

Table 7.6 shows the difference between the context of each source sense against the context of other senses, when the discourse relation is expressed implicitly (Column [1]) and explicitly (Column [2]). The difference suggests that implicit and explict DCs are used in different contexts, supporting the hypothesis. In particular, the difference between the context of each sense against others is smaller in implicit usage, thus making implicit relations harder to disambiguate.

Comparing with the difference in context between implicit and explicit usage (Column [3]), the context of source implicit relations that are explicitated in the target is similar to the context of source implicit relations that are kept implicit (Column [4]). This suggests that to explicitate the implicit DC or not in translation is independent of the local context to certain extent.

	$JSD(q, r)$			
source fine sense	[1] 1 sense vs all imp	[2] 1 sense vs all exp	[3] exp vs imp	[4] imp-imp vs imp-exp
而且 'and'	.025	.149	.142	.059
而 'whearas'	.052	.111	.124	.076
和 'and'	.066	.166	.186	.106
并 'also'	.064	.052	.068	.110
从而 'thus'	.052	.182	.189	.094
所以 'therefore	.051	.238	.239	.142
来 'in order to	.053	.126	.124	.178
因此 'therefore'	.039	.164	.164	.119
然后 'and then'	.154	.286	.316	.218
即 'which is'	.131	.321	.393	.205

Table 7.6: Jensen-Shannon Divergence (JSD) of various discourse usage of the top imp-exp DCs

Example (3) below shows the optionality of DC translation. It is taken from the test data of OpenMT 06. The implicit relations between the 3 discourse units in the source are translated by different DC usage in the target. For example, the relation between Unit 1 and Unit 2 is translated to a *Temporal* DC *‘as’* in Reference 1, while translated to a *Contingency* DC *‘so that’* in Reference 3. In Reference 2, 4, it is kept implicit. This suggests that multiple reference are necessary for evaluation of DC translation.

Example (3) - Source:

^{1}[这厚重的历史回声, 通过电视台”连线”大 陆和香港],^{2}[南京市民与香港同胞”天涯共此时 ” ,] ^{3}[共同庆祝香港回 归祖国十周年。]

Example (3) - Reference 1:

^{1}[This rich echo of history connected the mainland and Hong Kong via television,] ^{2}[***as*** the citizens of Nanjing and Hong Kong compatriots ”shared the same occasion from the far corners of the earth”]^{3}[***and*** celebrated together the tenth anniversary of Hong Kong’s reversion to the motherland.]

Example (3) - Reference 2

^{1}[This echo of profound historical significance ”connected” the Mainland and Hong Kong through television;]^{2}[citizens of Nanjing and their fellow countrymen in Hong Kong ”shared this moment with the entire world” together]^{3}[***celebrating*** the 10th anniversary of Hong Kong’s handover to the motherland]

Example (3) - Reference 3

^{1}[The sophisticated echo of history ”connected” the mainland and Hong Kong through a TV channel,] ^{2}[***so that*** Nanjing citizens and Hong Kong compatriots ”shared the moments across the land”]^{3}[***to*** celebrate together the 10th anniversary of Hong Kong’s return to the motherland.]

Example (3) - Reference 4

[1][The heavy historical echo ”connected” the Mainland with Hong Kong through television station.][2][Residents of Nanjing shared the moment with Hong Kong compatriots from afar][3][***to*** celebrate the 10th Anniversary of the return of Hong Kong to its motherland together.]

7.4 Conclusion

Motivated by the difference in DC usage between Chinese and English, the translation of implicit to explicit DCs given the gold crosslingual DC senses is investigated. To simulate the incorporation of implicit DC information to MT, the implicit DCs in the input source text are explicitated based on manual annotation, and decode the preprocessed input by baseline, non-discourse-aware SMT models. Results show that artificially explicitating source implicit DCs in the input text alone does not improve the MT performance significantly.

Further analysis by translation spotting suggests that discourse usage as well as sense disambiguation can be subject to a certain level of optionality. In the annotated corpus, explicitation of implicit source DCs in translation is suppressed, either by translation not using an explicit DC, or by translation using an ambiguous, sense-neutral explicit DC.

Chapter 8

Conclusion

This dissertation investigates the marking of discourse relations from two major perspectives. The first perspective is from the viewpoint of human language processing, in the monolingual dimension. The Discourse Marking Model is proposed to predict whether or not speakers will produce an explicit marker given the discourse relation they wish to express. The model combines, for the first time, two well-known and successful information-theoretic frameworks: (1) the Rational Speech Acts model [31] and (2) the Uniform Information Density theory [71].

The second part of the dissertation investigates the marking of discourse relations from the viewpoint of machine translation application. A bilingual resource of discourse relations is constructed, and the translation of discourse relations by humans and machine are compared.

Contributions of this dissertation are summarized by below response to the research questions proposed in Section 1.4.

8.1 Summary

8.1.1 Can speakers' choice of discourse marking be explained by RSA and UID?

Chapter 3 presents the Discourse Marking Model, which combines RSA and UID. On one hand, the RSA model models the pragmatic interaction between language production and interpretation by Bayesian inference. On the other hand, the UID principle advocates that speakers adjust linguistic redundancy to maintain a uniform rate of information transmission.

If the hypotheses of the model is appropriate, each component in the model should contribute to the prediction accuracy, and indeed such improvement is observed in the experimental results. Therefore, this study provides solid support that speakers' choice of discourse marking be explained by RSA and UID.

8.1.2 How does the Discourse Marking Model compare with previous work?

The Discourse Marking model quantifies the *utility* of using or omitting a DC based on the expected surprisal of comprehension, cost of production, and availability of other signals in the rest of the utterance. Experiments based on the PDTB show that the model outperforms the state-of-the-art performance at predicting the presence of DCs [105], in addition to giving an explanatory account of the speaker's choice.

In addition, Chapter 5 extends the framework to model human comprehension of discourse connectives. Following the Bayesian pragmatic paradigm, 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. Experimental results demonstrates the superiority of pragmatic inference over literal comprehension.

8.1.3 How is discourse marking ***reproduced*** cross-lingually in human Chinese-English translation?

From the machine translation perspective, Chapter 6 proposes a linguistically driven approach to represent discourse relations in Chinese text as *sequences*, and then alignment the relations to their English translation. The annotation scheme tackles surface characteristics of Chinese texts, such as the order of clauses and overt markers of discourse structures, that are not fully incorporated in existing annotation proposals adapted from formalism constructed for English.

An annotated resource consisting 7266 pairs of discourse relations in 325 articles of the translated Chinese treebank is constructed and released openly. A thorough analysis of the corpus statistics confirms that many Chinese implicit relations are indeed translated explicitly in English, but the tendency is relation-dependent.

8.1.4 Can MT be improved when source implicit discourse relations are *pre-explicitated*

Chapter 7 examines how implicit (omitted) DCs in the source text impacts various machine translation (MT) systems, and whether a discourse parser is needed as a preprocessor to explicitate implicit DCs. Based on the manually aligned discourse relations, various preprocessing step that inserts explicit DCs at positions of implicit relations are evaluated.

Results show that, without modifying the translation model, explicitating implicit relations in the input source text has limited effect on MT evaluation scores. In addition, translation spotting analysis shows that it is crucial to identify DCs that should be explicitly translated in order to improve implicit-to-explicit DC translation. On the other hand, further analysis reveals that the disambiguation as well as explicitation of implicit relations are subject to a certain level of optionality, suggesting the limitation to learn and evaluate this linguistic phenomenon using standard parallel corpora

****

To summarize, this dissertation produces a new state-of-the-art on the task of predicting discourse marking, and, at the same time, presents a cognitively plausible model to explain the choice from the viewpoint of human language processing. In addition, the study deepens our understanding on the cross-lingual transfer of discourse marking in the context of human and machine translation.

8.2 Future work

The current study on the marking of discourse relations serves as the the basis of a number of future work in terms of theory, methodology and application. The section discusses the future directions to which this work can be extended.

8.2.1 Theoretical research directions

The Discourse Marking Model presented in Chapter 3 combines the RSA model and the UID principle to explain speaker's preference in discourse marking. The experiment results support the significance of the model prediction, but also raises other new questions. These questions should be tackled in future work to refine the Discourse Marking Model in order to improve prediction accuracy.

The most fundamental question is to what extend RSA and UID account for speakers' choice of discourse markedness. Although significant improvement was observed using corpus data of natural distribution of senses and markedness, the model predictions only mildly matched with the majority judgment of human raters, using samples with even distribution of senses and markedness. In fact, the variation among human judgments suggests that DC and context informativeness and production cost might not be the only factors behind the choice of discourse marking. Further investigation into other factors is necessary to extend our understanding on human discourse production.

Another core question is weather the current model the best formulation to combine RSA and UID. The current method is basically built on the RSA model by estimating the informativeness of a DC, including implicit DC, in context. A bias to maintain UID according to the presence of cues in context is formulated as argument informativeness and added to the DC informativeness. Since both DC and argument informativeness use context information in the argument, it would be more elegant to combine the two informativeness modules, and directly estimate the informativeness of the DC, given the discourse cues. This could be done by a more expressive formulation of the discourse context C, which is defined by previous discourse marking and discourse relations in this work.

On the other hand, findings of the analysis on discourse marking in machine and human translation also introduce new questions to existing theories. It is known that human translators do explicitate and implicitate discourse relations in the translation process [8, 157, 92]. The Chinese-English translation corpus analysis in this study further identifies the option of explicitating a discourse connective without explicitating the relation sense, which is to translate an implicit discourse relation using an ambiguous discourse connective. Therefore, in machine translation, the open question is weather it is appropriate to explicitate discourse relation with an ambiguous DC, such as 'and'. Investigation in other language pairs and domains is essential to generalize this finding using the 'Chinese Treebank Translation Corpus' used in this study.

8.2.2 Extension in methodology

This study uses a computational psycholinguistic model to predict discourse marking choice made by speakers. Distributions of the DC and discourse relation sense annotation of the PDTB are used to approximate the informativeness of DCs. Evaluation on a held-out set of the corpus supports the effectiveness of the model predictions, but this methodology does not directly prove the psychological reality of each component of the model. For example, are the more expected discourse marking actually more informative to the listeners? Do the proposed cost functions correctly model speaker's production load? And most importantly, do speakers actually chose an utterance that is informative for the listeners and can smooth information density? The last question is related to another open question: is the utterance choice based on RSA or UID a conscious decision or not. Clearly, these questions can not be answered by a data-driven computational model, but require further behavioral and physiological experiments.

The methodology used in the crowd-sourcing behavioral study in this work can also be further compared with other settings. For example, in the current method, subjects were asked to rate the if it is 'correct to leave it (the DC) blank'. This question might have facilitated a bias towards 'implicit DC', and indeed the majority feedback was 'probably OK'. Further evaluation using other experimental setting, such as asking for a 2-way judgment, could lead to other provide more insights on the variation of discourse marking choice between different individuals.

Another methodology issue is about the choice of material in this study. Discourse presentation differs across genres [140] and mediums [134]. This work utilizes news articles in the Wall Street Journal for modeling training and testing. In news genre, conveying information correctly to the listeners (readers) is a primary purpose. In an argumentative article, people may choose to keep a relation implicit, and thus ambiguous, as a strategy to 'leave space for retrieval' [73]. In other words, instead of asserting the relation explicitly, it is more persuasive to leave space for the listeners to infer the meaning. Similarly, the cross-lingual annotation and analysis is based on written-text in the news domain, while the discrepancy of Chinese-English DC usage is different in conversation dialogues and other domains [130]. The suppression in explicitation of implicit DC could be due to the fact that subjective interpretation is avoided in news report. A future direction is thus to exploit data from other domains, and to identify implicit

DC relations that require explicitation in translation.

8.2.3 Future work on applications

Simulation on human's choice of discourse marking can be applied to dialogue systems to generate human-like conversations that are not excessively explicit or ambiguously implicit. There is also possibility to use the model to generate an informative and yet non-redundant DC given the intended discourse relation. On the other hand, as mentioned in the previous subsection, discourse representation differs across genres and media, the model can be applied to predict the explicitation of discourse relations from, for example, news articles to spoken dialogues.

This study has demonstrated that applying the model in reverse direction, to simulate human's interpretation of discourse relations, has the potential to improve automatic discourse parsing. A larger picture is to design a full, incremental discourse parsing algorithm that is motivated by the psycholinguistic reality of human discourse processing, which in turn contributes to a comprehensive model on human language processing.

Last but not least, findings on the both perspectives of discourse marking in this work can be combined. Towards an discourse-relation-aware approach of machine translation, the discourse marking of the target English text can be predicted using the Discourse Marking Model, and based on the automatic classified sense on the Chinese side. To this end, the possibility to replace the conventional sentence-to-sentence machine translation to discourse-unit-to-discourse-unit machine translation also worth investigating.

8.3 Closing remark

Discourse relation is a challenging but interesting research topic in NLP as it is the top level of linguistic structures. A text containing grammatical errors or wrong use of words can be understood since the human language processing system is robust to erroneous input [29] as long as the discourse relations make sense. However, a grammatically perfect text is hard to read if the author's intention cannot be perceived from the organization of the discourse relations.

The primitive motivation of this study is human-level artificial intelligence -

not just to passively understand human languages but also to manipulate people by producing language with a clear intention. To achieve this goal, it is worth to examine how humans achieve this and formalize the findings by a computational model. I hope to continue research that applies, and evaluates, theoretical findings in related disciplines, such as linguistics or psychology, on NLP tasks in future.

Bibliography

[1] D. Allbritton and J. Moore. Discourse cues in narrative text: Using production to predict comprehension. In *AAAI Fall Symposium on Psychological Models of Communication in Collaborative Systems*, 1999.

[2] N. Asher. Implicatures and discourse structure. *Lingua*, 132:13–28, 2013.

[3] F. T. Asr and V. Demberg. Implicitness of discourse relations. *Proceedings of the International Conference on Computational Linguistics*, pages 2669–2684, 2012.

[4] F. T. Asr and V. Demberg. On the information conveyed by discourse markers. In *Proceedings of the Annual Workshop on Cognitive Modeling and Computational Linguistics*, pages 84–93, 2013.

[5] F. T. Asr and V. Demberg. 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, 2015.

[6] M. Aylett and A. Turk. 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, 2004.

[7] P. Baumann, B. Clark, and S. Kaufmann. Overspecification and the cost of pragmatic reasoning about referring expressions. *Proceedings of the Annual Conference of the Cognitive Science Society*, 2014.

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

[9] A. Benz, G. Jager, R. Van Rooij, and R. Van Rooij. *Game theory and pragmatics*. Springer, 2016.

[10] L. Bergen, R. Levy, and N. D. Goodman. Pragmatic reasoning through semantic inference, 2014.

[11] A. Bies, M. Palmer, J. Mott, and C. Warner. English chinese translation treebank v1.0 (ldc2007t02). *Linguistic Data Consortium*, 2007.

[12] M. Bittner. Topic states in mandarin discourse. *Proceedings of the North American Conference on Chinese Linguistics*, 2013.

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

[14] K. Bock, G. S. Dell, F. Chang, and K. H. Onishi. Persistent structural priming from language comprehension to language production. *Cognition*, 104(3):437–458, 2007.

[15] B. K. Britton, S. M. Glynn, B. J. Meyer, and M. Penland. Effects of text structure on use of cognitive capacity during reading. *Journal of Educational Psychology*, 74(1):51, 1982.

[16] L. Carlson, M. E. Okurowski, and D. Marcu. Rst discourse treebank, 2002.

[17] B. Cartoni, S. Zufferey, and T. Meyer. Annotating the meaning of discourse connectives by looking at their translation: The translationspotting technique. *Dialogue & Discourse*, 4(2), 2013.

[18] C. Chang and C. Lin. Libsvm : a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27, 2011.

[19] C. C. Chu. *A discourse grammar of Mandarin Chine*. P. Lang., 1998.

[20] C. C. Chu and Z. Ji. *A Cognitive-Functional Grammar of Mandarin Chinese*. Crane, 1999.

[21] H. H. Clark. Using language. *Journal of Linguistics*, 35(1):167–222, 1999.

[22] R. Dale and E. Reiter. Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive science*, 19(2):233–263, 1995.

[23] L. Danlos. Linguistic ways for expressing a discourse relation in a lexicalized text generation system. *Workshop of Discourse Relations and Discourse Markers*, pages 50–53, 1998.

[24] D. Das, M. Taboada, and P. McFetridge. Rst signalling corpus, 2015.

[25] K. Doya. *Bayesian brain: Probabilistic approaches to neural coding.* MIT press, 2007.

[26] H. Eifring. *Clause Combination in Chinese.* BRILL, 1995.

[27] P. E. Engelhardt, K. G. Bailey, and F. Ferreira. Do speakers and listeners observe the gricean maxim of quantity? *Journal of Memory and Language*, 54(4):554–573, 2006.

[28] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874, 2008.

[29] F. Ferreira and M. W. Lowder. Prediction, information structure, and good-enough language processing. 65:217–247, 2016.

[30] A. Frank and T. F. Jaeger. 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, 2008.

[31] M. C. Frank and N. D. Goodman. Predicting pragmatic reasoning in lanugage games. *Science*, 336(6084):998, 2012.

[32] B. Galantucci, C. A. Fowler, and M. T. Turvey. The motor theory of speech perception reviewed. *Psychonomic bulletin and review*, 13(3):361–377, 2006.

[33] H.-D. Gasde and W. Paul. Funcional categories, topic prominence, and complex sentences in mandarin chinese. *Linguistics*, 34, 1996.

[34] A. Gatt, R. P. van Gompel, K. van Deemter, and E. Krahmer. Are we bayesian referring expression generators. In *Proceedings of the Workshop on Production of Referring Expressions: Bridging the Gap between Cognitive and Computational Approaches to Reference*, 2013.

[35] D. Genzel and E. Charniak. Entropy rage constancy in text. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 199–206, 2002.

[36] N. D. Goodman and D. Lassiter. *Probabilistic semantics and pragmatics: Uncertainty in language and thought*. Wiley-Blackwell, 2014.

[37] N. D. Goodman and A. Stuhlmuller. Knowledge and implicature: modeling language understanding as social cognition. *Topics in cognitive science*, 5(1):173–184, 2013.

[38] H. P. Grice. Logic and conversation. *Syntax and Semantics*, 3:41–58, 1975.

[39] S. T. Gries. Syntactic priming: A corpus-based approach. *Journal of psycholinguistic research*, 34(4):365–399, 2005.

[40] B. J. Grosz and C. L. Sidner. Attention, intentions, and the structure of discourse. *Computational linguistics*, 12(3):175–204, 1986.

[41] B. Grote and M. Stede. Discourse marker choice in sentence planning. In *Proceedings of the International Workshop on Natural Language Generation*, pages 128–137, 1998.

[42] K. Haberlandt. Reader expectations in text comprehension. *Advances in Psychology*, 9:239–249, 1982.

[43] N. Hajlaoui and A. Popescu-Belis. Translating english discourse connectives into arabic: a corpus-based analysis and an evaluation metric. *Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages*, 2012.

[44] N. Hajlaoui and A. Popescu-Belis. Assessing the accuracy of discourse connective translations: Validation of an automatic metric. *Computational Linguistics and Intelligent Text Processing*, 2013.

[45] C. Hardmeier. Discourse in statistical machine translation. a survey and a case study. *Discours. Revue de linguistique, psycholinguistique et informatique.*, (11), 2012.

[46] J. R. Hobbs. *On the coherence and structure of discourse*. CSLI, 1985.

[47] J. Hoek, J. Evers-Vermeul, and T. J. Sanders. The role of expectedness in the implicitation and explicitation of discourse relations. *Proceedings of the Workshop on Discourse in Machine Translation*, pages 41–46, 2015.

[48] J. Hoek and S. Zufferey. Factors influencing the implicitation of discourse relations across languages. In *Proceedings the Joint ACL-ISO Workshop on Interoperable Semantic Annotation*, pages 39–45. TiCC, Tilburg center for Cognition and Communication, 2015.

[49] H.-H. Huang, T.-W. Chang, H.-Y. Chen, and H.-H. Chen. Interpretation of chinese discourse connectives for explicit discourse relation recognition. *Proceedings of the International Conference on Computational Linguistics*, 2014.

[50] H.-H. Huang and H.-H. Chen. Chinese discourse relation recognition. *Proceedings of the International Joint Conference on Natural Language Processings*, 2011.

[51] H.-H. Huang and H.-H. Chen. Contingency and comparison relation labeling and structure predictuion in chinese sentences. *Proceedings of the Annual Meeting of SIGDIAL*, 2012.

[52] H.-H. Huang, C.-H. Yu, T.-W. Chang, C.-K. lin, and H.-H. Chen. Analyses of the association between discourse relation and sentiment polarity with a chinese human-annotated corpus. *Proceedings of the Linguistic Annotation Workshop and Interperability with Discourse*, 2013.

[53] B. Hutchinson. Modelling the similarity of discourse connectives. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 2005.

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

[55] T. F. Jaeger and N. Snider. Implicit learning and syntactic persistence: Surprisal and cumulativity. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, page 827, 2008.

[56] G. Jager. *Game theory in semantics and pragmatics*, volume 3. Mouton de Gruyter, 2012.

[57] Y. Ji. *Semantic representation learning for discourse processing*. PhD thesis, Georgia Institute of Technology, 2016.

[58] Y. Ji, G. Haffari, and J. Eisenstein. A latent variable recurrent neural network for discourse relation language models. *arXiv preprint arXiv:1603.01913*, 2016.

[59] L. Jin and M.-C. de Marneffe. The overall markedness of discourse relations. In *Proceedings of the Conference on Empirical Methods on Natural Language Processing*, pages 1114–1119, 2015.

[60] J. Kamalski, T. Sanders, and L. Lentz. Coherence marking, prior knowledge, and comprehension of informative and persuasive texts: Sorting things out. *Discourse Processes*, 45(4-5):323–345, 2008.

[61] J. T. Kao, J. Y. Wu, L. Bergen, and N. D. Goodman. Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences*, 111(33):12002–12007, 2014.

[62] F. Keller. The entropy rate principle as a predictor of processing effort: An evaluation against eye-tracking data. *Proceedings of the conference on empirical methods in natural language processing*, 2004.

[63] J. M. Kilner, K. J. Friston, and C. D. Frith. Predictive coding: an account of the mirror neuron system. *Cognitive processing*, 8(3):159–166, 2007.

[64] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. Moses: Open source toolkit for statistical machine translation. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2007.

[65] G. R. Kuperberg, M. Paczynski, and T. Ditman. Establishing causal coherence across sentences: An erp study. *Journal of Cognitive Neuroscience*, 23(5):1230–1246, 2011.

[66] D. Lassiter and N. D. Goodman. Context, scale structure, and statistics in the interpretation of positive-form adjectives. *Semantics and Linguistic Theory*, 23:587–610, 2013.

[67] D. Lassiter and N. D. Goodman. Adjectival vagueness in a bayesian model of interpretation. *Synthese*, pages 1–36, 2015.

[68] L. Lee. On the effectiveness of the skew di- vergence for statistical language analysis. *Artificial Intelligence and Statistics*, 2001.

[69] W. J. Levelt and S. Kelter. Surface form and memory in question answering. *Cognitive psychology*, 14(1):78–106, 1982.

[70] S. C. Levinson. *Presumptive meanings: The theory of generalized conversational implicature*. MIT Press, 2000.

[71] R. Levy and T. F. Jaeger. Speakers optimize information density through syntactic reduction. *Advances in neural information processing systems*, pages 849–856, 2006.

[72] R. Levy and C. Manning. Is it harder to parse chinese, or the chinese treebank. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2003.

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

[74] J. J. Li, M. Carpuat, and A. Nenkova. Assessing the discourse factors that influence the quality of machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 283–288, 2014.

[75] J. J. Li, M. Carpuat, and A. Nenkova. Cross-lingual discourse relation analysis: A corpus study and a semi-supervised classifacation system. *Proceedings of the International Conference on Computational Linguistics*, 2014a.

[76] Y. Li, W. Feng, J. Sun, F. Kong, and G. Zhou. Building chinese discourse corpus with connective-driven dependency tree structure. *Proceedings of the Conference on Empirical Methods on Natural Language Processing*, 2014b.

[77] J. Lin. Syntactic structures of complex sentences in mandarin chinese. *Nanzan Linguistics*, 3, 2006.

[78] Z. Lin, M. Kan, and H. T. Ng. Recognizing implicit discourse relations in the penn discourse treebank. *Proceedings of the Conference on Empirical Methods on Natural Language Processing*, 2009.

[79] Z. Lin, H. T. Ng, and M. Kan. A pdtb-styled end-to-end discourse parser. Technical report, National University of Singapore, 2010.

[80] Z. Lin, H. T. Ng, and M. Kan. Automatic evaluating text coherence using discourse relations. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2011.

[81] N. L. Loman and R. E. Mayer. Signaling techniques that increase the understandability of expository prose. *Journal of Educational psychology*, 75(3):402, 1983.

[82] R. F. Lorch Jr and E. P. Lorch. On-line processing of summary and importance signals in reading. *Discourse Processes*, 9(4):489–496, 1986.

[83] W. C. Mann and S. A. Thompson. Rhetorical structure theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281, 1988.

[84] C. Manning, M. Surdeanu, J. Bauer, J. Finkey, S. J. Bethard, and D. McClosky. The standord corenlp natural language processing toolkit. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, 2014.

[85] D. Marcu, L. Carlson, and M. Watanabe. The automatic translation of discourse structures. *Proceedings of the North American Chapter of the Association of Computational Linguistics*, 2000.

[86] D. Marcu and A. Echihabi. An unsupervised approach to recognizing discourse relations. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2002.

[87] B. J. Meyer. *The organization of prose and its effects on memory*, volume 1. North-Holland Publishing Co., 1975.

[88] B. J. Meyer, D. M. Brandt, and G. J. Bluth. Use of top-level structure in text: Key for reading comprehension of ninth-grade students. *Reading research quarterly*, pages 72–103, 1980.

[89] T. Meyer and L. Polakova. Machine translation with many manually labeled discourse connectives. *Proceedings of the Discourse in Machine Translation Workshop*, 2013.

[90] T. Meyer, A. Popescu-Belis, and N. Hajlaoui. Machine translation of labeled discourse connectives. *Proceedings of the Biennial Conference of the Association for Machine Translation in the Americas*, 2012.

[91] T. Meyer, A. Popescu-Belis, S. Zufferey, and B. Cartoni. Multilingual annotation and disambiguation of discourse connectives for machine translation. *Proceedings of the Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2011.

[92] T. Meyer and B. Webber. Implicitation of discourse connectives in (machine) translation. In *Proceedings of the Workshop on Discourse in Machine Translation*, pages 19–26, 2013.

[93] K. K. Millis and M. A. Just. The influence of connectives on sentence comprehension. *Journal of Memory and Language*, 33(1):128–147, 1994.

[94] W. Monroe and C. Potts. Learning in the rational speech acts model. *arXiv preprint arXiv:1510.06807*, 2015.

[95] M. Moser and J. D. Moore. Investigating cue selection and placement in tutorial discourse. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 130–135. Association for Computational Linguistics, 1995.

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

[97] G. Neubig. Travatar: A forest-to-string machine translation engine based on tree transducers. *Proceedings of the Annual Meeting of the Association for Computational Linguistics (Demonstration Track)*, 2013.

[98] K. Nguyen and B. O'Connor. Posterior calibration and exploratory analysis for natural language processing models. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, pages 1587–1598, 2015.

[99] F. J. Och. Minimum error rate training in statistical machine translation. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2003.

[100] F. J. Och and H. Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 2003.

[101] N. Orita, E. Vornov, N. H. Feldman, and H. D. III. Why discourse affects speakers' choice of refering expressions. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 1639–1649, 2015.

[102] M. Palmer, F.-D. Chiou, N. Xue, and T.-K. Lee. Chinese treebank 5.0 (ldc2005t01). *Linguistic Data Consortium*, 2005.

[103] J. Park and C. Cardi. Improving implicit discourse relation recognition through feature set optimization. *Proceedings of Annual Meeting on Discourse and Dialogue*, 2012.

[104] R. Parker, D. Graff, J. Kong, K. Chen, and K. Maeda. English gigaword fifth edition ldc2011t07. Linguistic Data Consortium, 2011.

[105] G. Patterson and A. Kehler. Predicting the presence of discourse connectives. In *Proceedings of the Conference on Empirical Methods on Natural Language Processing*, pages 914–923, 2013.

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

[107] M. J. Pickering and S. Garrod. An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, 36(04):329–347, 2013.

[108] E. Pitler, A. Louis, and A. Nenkova. 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*, pages 683–691, 2009.

[109] E. Pitler and A. Nenkova. Using syntax to disambiguate explicit discourse connectives in text. *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*, pages 13–16, 2009.

[110] E. Pitler, M. Raghupathy, H. Mehta, A. Nenkova, A. Lee, and A. Joshi. Easily identifiable discourse relations. *Technical Report, University of Pennsylvania*, 2008.

[111] A. Popescu-Belis, T. Meyer, J. Liyanapathirana, B. Cartoni, and S. Zufferey. Discourse-level annotation over europarl for machine translation: Connectives and pronouns. *Proceedings of the Language Resource and Evaluation Conference*, 2012.

[112] C. Potts, D. Lassiter, R. Levy, and M. C. Frank. Embedded implicatures as pragmatic inferences under compositional lexical uncertainty. Manuscript, 2015.

[113] R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi, and B. Webber. The penn discourse treebank 2.0. *Proceedings of the Language Resource and Evaluation Conference*, 2008.

[114] R. Prasad, B. Webber, and A. Joshi. Reflections on the penn discourse treebank, comparable corpora, and complementary annotation. *Computational Linguistics*, pages 921–950, 2014.

[115] J. P. Rickards, B. R. Fajen, J. F. Sullivan, and G. Gillespie. Signaling, notetaking, and field independence–dependence in text comprehension and recall. *Journal of educational psychology*, 89(3):508, 1997.

[116] H. Rohde, A. Dickinson, C. Clark, A. Louis, and B. Webber. Recovering discourse relations: Varying influence of discourse adverbials. In *Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, page 22, 2015.

[117] A. Rutherford and N. Xue. 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*, pages 645–654, 2014.

[118] A. Rutherford and N. Xue. Improving the inference of implicit discourse relations via classifying explicit discourse connectives. *Proceedings of the North American Chapter of the Association of Computational Linguistics*, pages 799–808, 2015.

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

[120] T. J. Sanders and L. G. Noordman. The role of coherence relations and their linguistic markers in text processing. *Discourse processes*, 29(1):37–60, 2000.

[121] T. J. Sanders, W. P. Spooren, and L. G. Noordman. Toward a taxonomy of coherence relations. *Discourse processes*, 15(1):1–35, 1992.

[122] M. C. Scholman, J. Evers-Vermeul, and T. J. Sanders. Categories of coherence relations in discourse annotation: Towards a reliable categorization of coherence relations. *Dialogue & Discourse*, 7(2):1–28, 2016.

[123] D. Scott and C. S. de Souza. Getting the message across in rst-based text generation. *Current research in natural language generation*, 4:47–73, 1990.

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

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

[126] X. She, P. Jian, P. Zhang, and H. Huang. Leveraging hierarchical deep semantics to classify implicit discourse relations via mutual learning method. In *International Conference on Computer Processing of Oriental Languages*, pages 349–359. Springer, 2016.

[127] M. Smith and L. Wheeldon. Syntactic priming in spoken sentence production–an online study. *Cognition*, 78(2):123–164, 2001.

[128] C. Soria and G. Ferrari. Lexical marking of discourse relations-some experimental findings. In *Proceedings of the ACL-98 Workshop on Discourse Relations and Discourse Markers*, pages 36–42, 1998.

[129] J. H. Spyridakis and T. C. Standal. Signals in expository prose: Effects on reading comprehension. *Reading Research Quarterly*, pages 285–298, 1987.

[130] D. Steele and L. Specia. Divergences in the usage of discourse markers in english and mandarin chinese. *Text, Speech and Dialogue*, 2014.

[131] A. Stubbs. Mae and mai: lightweight annotation and adjudication tools. pages 129–133, 2011.

[132] J. H. Tai. Temporal sequence and chinese word order. *Iconicity in Syntax*, 1985.

[133] H. Tily and S. Piantadosi. Refer efficiently: Use less informative expressions for more predictable meanings. *Proceedings of the workshop on the production of referring expressions*, 2009.

[134] S. Tonelli, G. Riccardi, R. Prasad, and A. K. Joshi. Annotation of discourse relations for conversational spoken dialogs. 2010.

[135] R. Trivedi and J. Eisenstein. Discourse connectors for latent subjectivity in sentiment analysis. *Proceedings of the North American Chapter of the Association for Computational Linguistics*, 2013.

[136] M. Tu, Y. Zhou, and C. Zong. A novel translation framework based on rhetori- cal structure theory. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2013.

[137] M. Tu, Y. Zhou, and C. Zong. Enhancing grammatical cohesion: Generating transitional expressions for smt. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2014.

[138] T. A. Van Dijk. The semantics and pragmatics of functional coherence in discourse. In *Speech act theory: Ten years later*, pages 49–66. Versus, 1980.

[139] J. Wang and M. Lan. A refined end-to-end discourse parser. *CoNLL 2015*, pages 17–24, 2015.

[140] B. Webber. Genre distinctions for discourse in the penn treebank. pages 674–682, 2009.

[141] B. Webber, M. Stone, A. Joshi, and A. Knott. Anaphora and discourse structure. *Computational linguistics*, 29(4):545–587, 2003.

[142] F. Wolf, E. Gibson, A. Fisher, and M. Knight. The discourse graphbank: A database of texts annotated with coherence relations, 2005.

[143] N. Xue, H. T. Ng, S. Pradhan, R. P. C. Bryant, and A. T. Rutherford. The conll-2015 shared task on shallow discourse parsing. In *Proceedings of the CoNLL-2015 Shared Task on Shallow Discourse Parsing*, pages 1–16, 2015.

[144] N. Xue, H. T. Ng, A. Rutherford, B. Webber, C. Wang, and H. Wang. Conll 2016 shared task on multilingual shallow discourse parsing. *Proceedings of the CoNLL-16 shared task*, pages 1–19, 2016.

[145] X. Yang and D. Reitter. Entropy converges between dialogue participants: Explanations from an information-theoretic perspective. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2016.

[146] Y. Yang and N. Xue. Chinese comma disambiguation for discourse analysis. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2012.

[147] Y. Yoshida, J. Suzuki, T. Hirao, and M. Nagata. Dependency-based discourse parser for single-document summarization. *Proceedings of the Conference on Empirical Methods on Natural Language Processing*, 2014.

[148] F. Yung, K. Duh, and Y. Matsumoto. Crosslingual annotation and analysis of implicit discourse connectives for machine translation. In *Proceedings of the Workshop on Discourse in Machine Translation*, pages 142–152, 2015.

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

[150] H. Zeevat. *Perspectives on Bayesian Natural Language Semantics and Pragmatics*, pages 1–24. Springer, 2015.

[151] L. J. Zhou, W. Gao, B. Li, Z. Wei, and K.-F. Wong. Cross-lingual identification of ambiguous discourse connectives for resource-poor language. *Proceedings of the International Conference on Computational Linguistics*, 2012.

[152] L. J. Zhou, B. Li, Z. Wei, and K.-F. Wong. The cuhk discourse treebank for chinese: Annotating explicit discourse connectives for the chinese treebank. *Proceedings of the Language Resource and Evaluation Conference*, 2014.

[153] Q. Zhou. Annotation scheme for chinese treebank. *Journal of Chinese Information Processing*, 2004.

[154] Y. Zhou and N. Xue. Pdtb-style discourse annotation of chinese text. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2012.

[155] Y. Zhou and N. Xue. The chinese discourse treebank: a chinese corpus annotated with discourse relations. *Language Resources and Evaluation*, 49(2):397–431, 2015.

[156] Z.-M. Zhou, Y. Xu, Z.-Y. Niu, M. Lan, J. Su, , and C. L. Tan. Predicting discourse connectives for implicit discourse relation recognition. *Proceedings of the International Conference on Computational Linguistics*, 2010.

[157] S. Zuffery and B. Cartoni. A multifactorial analysis of explicitation in translation. *Target*, 26(3), 2014.