Cross-lingual Discourse Relation Analysis: A corpus study and a semi-supervised classification system

Junyi Jessy Li[1]**, Marine Carpuat**[2] **and Ani Nenkova**[1]
[1] University of Pennsylvania, Philadelphia, PA 19104, USA
{ljunyi, nenkova}@seas.upenn.edu
[2] National Research Council Canada, Ottawa, ON K1A 0R6, Canada
marine.carpuat@nrc.gc.ca

Abstract

We present a cross-lingual discourse relation analysis based on a parallel corpus with discourse information available only for one language. First, we conduct a corpus study to explore differences in discourse organization between Chinese and English, including differences in information packaging, implicit/explicit discourse expression divergence, and discourse connective ambiguities. Second, we introduce a novel approach to learning to recognize discourse relations, using the parallel corpus instead of discourse annotation in the language of interest. Our resulting semi-supervised system reaches state-of-art performance on the task of discourse relation detection, and outperforms a supervised system on discourse relation classification.

1 Introduction

The analysis of the way spans of text semantically connect with each other to create a coherent text has a rich theoretical and empirical tradition (Mann and Thompson, 1988; Marcu, 1997; Di Eugenio et al., 1997; Allbritton and Moore, 1999; Schilder, 2002). Because of the difficulty in annotation, however, labelled datasets were rare and rather small.

The release of the Penn Discourse Treebank (PDTB) (Prasad et al., 2008) brought about a new sense of maturity in discourse analysis, finally providing a high-quality large-scale resource for training discourse parsers for English. Based on the PDTB, a number of studies have provided insightful analysis of the use of discourse connectives in English news text and have developed methods for the identification of discourse relations and their arguments (Wellner and Pustejovsky, 2007; Pitler et al., 2008; Pitler and Nenkova, 2009; Pitler et al., 2009; Lin et al., 2009; Prasad et al., 2010; Park and Cardie, 2012; Lin et al., 2014). Some have applied the insights and classifiers to standard natural language processing tasks such as assessing text coherence and text quality (Pitler and Nenkova, 2008; Lin et al., 2011), detecting causal dependencies of events (Do et al., 2011), and machine translation (Meyer and Popescu-Belis, 2012).

A resource like the PDTB is extremely valuable, and it would be desirable to have a similar resource in other languages as well. Following the release of the PDTB, smaller corpora annotated with discourse relations have been developed for Hindi (Oza et al., 2009), Turkish (Zeyrek and Webber, 2008), Arabic (Al-Saif and Markert, 2010), and the effort is on-going with Chinese (Zhou and Xue, 2012).

On the other hand, for the vast majority of languages, such well-annotated resource for discourse relations is not available. In our work we carry the valuable annotations in the PDTB over to another language—Chinese—using parallel corpora. Projecting information available in one language onto another has been explored in areas such as part-of-speech tagging (Yarowsky et al., 2001; Das and Petrov, 2011), grammar induction (Hwa et al., 2005; Ganchev et al., 2009) and semantic role labeling (Pado and Lapata, 2005; Johansson and Nugues, 2006; van der Plas et al., 2011). For discourse relations, prior work has shown that a parallel corpus is helpful for disambiguating certain explicit discourse connectives (Meyer et al., 2011). To the best of our knowledge, the work we present here is the first study that directly infers discourse relations using resources only available in another language.

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers,
pages 577–587, Dublin, Ireland, August 23-29 2014.

The goal of our work is not only to measure the accuracy with which discourse relations can be identified in another language without annotations beyond the PDTB, but also to catalog the differences in discourse relation realization across different languages, Chinese and English in our case. We show that the two languages vastly differ in how information is packaged into a sentence, which also leads to differences in the implicit/explicit expression of discourse relations and the ambiguities in discourse connectives. These differences challenge the currently accepted distinctions between syntax and discourse between the two languages for applications such as machine translation. Then we present our semi-supervised learning algorithm to recognize explicit discourse relations in Chinese, relying solely on discourse information available in English. For multiway classification, our system outperforms a supervised system trained on the existing pilot dataset of discourse relations in Chinese (Zhou and Xue, 2012). In the task of binary classification for identifying specific discourse relations, the performance of our system is within 4% accuracy of that of the supervised system for all but one relations.

2 Data

As our parallel corpus, we use the newswire portion of the GALE Chinese-English Word Alignment and Tagging Training corpus (parts 1 and 2). The corpus contains 2,175 newswire articles, corresponding to 6,255 translation segments with 248,999 Chinese characters. These articles were translated into English by human translators. Gold standard word alignments are available for this corpus. A *minimal match* alignment approach (Li et al., 2010) was adopted for creating the gold standard, namely, alignments are between an English word and only the necessary Chinese *characters*. We repurpose this resource created for machine translation research for our cross-lingual discourse analysis. The availability of manual alignments between Chinese discourse connectives and their English translation makes it possible to conduct a reliable analysis by focusing on actual cross-lingual divergences, without noise introduced by potential errors from automatic aligners. [1]

We use a highly accurate supervised classifier for English explicit discourse relations (Pitler and Nenkova, 2009)[2] to automatically annotate the English portion of the GALE parallel corpus. The classifier was trained on the PDTB to identify discourse relations explicitly signaled by a set of 100 discourse connectives such as *however, because, while* or *for example*. For each instance of the 100 words or expressions, the classifier predicts if the expression is used as a discourse connective or if the instance is a non-discourse connective sense of the phrase or word. For each instance predicted to be a discourse connective, the classifier identifies the discourse relation signaled by the connective: TEMPORAL, COMPARISON, CONTINGENCY or EXPANSION. In our work we predict the same five categories for Chinese expressions which can serve as discourse connectives.

For evaluation and the study of discourse connective ambiguities, we use a development set from the Chinese Discourse Treebank (CDTB) (Zhou and Xue, 2012) consisting of 170 documents[3]. In the CDTB, an annotation style similar to the PDTB is applied on the texts from the Chinese Treebank corpus (Xue et al., 2005). For a discourse connective, one of eight discourse relation senses is annotated. All of these classes are subsumed by the four top-level relations in the PDTB. We map them to the PDTB relation senses according to their definitions:

Alternative → Expansion; Causation → Contingency; Conditional → Contingency; Conjunction → Expansion; Contrast → Comparison; Expansion → Expansion; Purpose → Contingency; Temporal → Temporal.

3 Information packaging characteristics

The notion of sentence in Chinese is very different from that in English. Punctuation marks were introduced in the early 20^{th} century; sentences resemble more a collection of related information than structurally well-defined syntactic units as in English. In fact, commas are often ambiguous, signaling

[1] While cross-lingual projection could be directly applied to automatic word alignments, discourse relation analysis raises some specific challenges because the main target of analysis (discourse connectives) are function words, which do not have as much of an impact on the final analysis in applications focusing on content words. As a result, we exclusively use manual alignment links in this study, and will address issues raised by automatic alignments in future work.

[2] The classifier is available at http://www.cis.upenn.edu/∼epitler/discourse.html

[3] This is an on-going annotation project. We are grateful to the authors for providing us with their valuable development set.

	% data	avg-length	std-length
1-many	18.83	61.42	28.85
1-1	81.17	35.73	25.34

Table 1: Percentage, length and standard deviation of sentences for which one Chinese source sentence is translated into one (1-1) or multiple (1-many) English sentences. Length is calculated based on the number of Chinese characters.

either clausal subordination, coordination or end-of-sentence (as construed from an English-centric point of view). Automatic systems have been developed to disambiguate the function of commas (Jin et al., 2004; Xue and Yang, 2011). This is a rather interesting phenomenon for discourse processing, as the English equivalents of Chinese sentences are in fact multi-sentential discourses in English.

The GALE corpus allows us to examine how often this mismatch of discourse organization occurs. Here we look for Chinese source sentences that were translated into multiple English sentences by the human translators. Consider the following example in which the corresponding clauses on both sides are numbered and marked in square brackets:

source [近年来"救灾外交"、"救灾援助"等新名词不断出现]$_1$，[各国围绕救灾问题展开了暗中的竞争与较量]$_2$，[一些国家谋求以救灾为名成立各种国际联盟]$_3$。

ref [In recent years, new phrases such as "disaster relief diplomacy" and "disaster relief aid" have appeared constantly]$_1$. [In relation to the issue of disaster relief, all countries have been silently competing with one another and comparing offerings]$_2$. [Some countries are trying to establish various kinds of international alliance in the name of disaster relief]$_3$.

In this example, the Chinese sentence packed the following related content into a single sentence: the occurrence of the new phrases about disaster relief, the competition among the countries related to disaster relief, and alliances in the name of disaster relief. The phrases expressing this information are separated by the commas in the source Chinese sentence because they are about a single concept "disaster relief". However, this information needs to be partitioned into three different sentences, each with different subjects, when translated to English.

In the GALE corpus, we identified 1,178 (out of total 6,255) source sentences with reference translations containing more than one sentence. In other words, sentence/discourse mismatch between Chinese and English occurs for 18.83% of the data. Table 1 shows the portion of data involved in such mismatch, with percentage, mean and standard deviation of source sentence length. Not surprisingly, Chinese sentences that require multiple sentences in their English translation are much longer. These long sentences are fairly common, which suggests that the difference in information packaging is highly prevalent and could potentially affect key applications such as machine translation, where systems are trained on a sentence to sentence basis.

We will return to the discussion of this mismatch later, when we discuss how English and Chinese also appear to differ in the way discourse relations are signaled. Briefly, the issue is that relations that are explicit in one language may become implicit in the other, easily inferred by the reader but not marked by a discourse connective. Also, there is an increase in the sense ambiguity of discourse connectives related to EXPANSION relations in Chinese.

4 Implicit and explicit relations

In this section, we present two other differences between the two languages related to discourse organization. One is the need for a discourse relation expressed implicitly in one language to be expressed explicitly in another. The other is the difference of the ambiguity of discourse connectives across the two languages. Before the discussion of these interesting asymmetries, we first present the method for direct projection of discourse relations using the GALE gold standard alignments, which we use to gather a set of explicit discourse connectives in Chinese.

4.1 Direct projection

Thus far we have available a parallel Chinese/English corpus, discourse connectives automatically tagged with their senses on the English side and manual alignments of atomic units between English and Chi-

	Comparison	Contingency	Expansion	Temporal
CH/EN mismatch	63	109	360	195
all	551	469	1198	885
% data	11.43	23.24	30.05	22.03

Table 2: Numbers and percentages of Chinese/English implicit/explicit mismatches.

nese. So for each discourse connective in an English sentence, it is straightforward to identify the corresponding expressions in the Chinese sentence following the gold standard alignments. Then the aligned Chinese expression can be assigned a discourse tag—non-discourse use or one of the four main discourse relation types—which is the same as in the English translation. We call the resulting annotation on the Chinese sentences *discourse projection*.

Further we discard potential expressions of Chinese connectives if they occurred with the same part of speech only once in the entire corpus. The result is a list of a total of 118 Chinese discourse connectives harvested using direct projection.

4.2 Implicit or Explicit?

A discourse relation can be expressed either with an explicit connective (e.g. *however, since*), or implicitly without a connective, in which case the relation would have to be inferred by the reader. Languages may differ in how they express discourse relations.

We investigate such implicit/explicit mismatch using direct projection. Specifically, we study the cases in which an English discourse connective is not aligned to any part of its corresponding Chinese sentence. In this case, the human translator explicitly expressed a discourse relation that was implicitly conveyed in the corresponding Chinese sentence.

The following four examples illustrate a Chinese/English implicit/explicit mismatch for each of the TEMPORAL, COMPARISON, CONTINGENCY and EXPANSION relation, respectively. On the Chinese side we also mark the position of the inserted English connective.

source [当地时间4月27日]$_1$，[阿富汗首都喀布尔举行的抗击苏联入侵胜利阅兵式遭到袭击]$_2$，when$_{\text{TEMPORAL}}$ [阿富汗总统卡尔扎伊和其他政要慌忙撤离现场]$_3$。
ref [On april 27 local time]$_1$, [Afghan president Karzai and other important officials were forced to flee the scene]$_3$ [when$_{\text{TEMPORAL}}$ a military parade in Kabul, Afghanistan commemorating victory in the fight against the soviet invasion was attacked]$_2$.

source [但在目前普遍使用的十种语言中]$_1$，[阿拉伯语仅名列第四位]$_2$，while$_{\text{COMPARISON}}$[英语名列第一]$_3$。
ref [However, of the ten commonly-used languages today]$_1$, [Arabic only ranks fourth]$_2$, [while$_{\text{COMPARISON}}$ English ranks first]$_3$.

source [“中华航空”上海代表处首席代表董大伟告诉记者]$_1$：“[现在的两岸包机还不是真正意义上的‘直航’]$_2$，since$_{\text{CONTINGENCY}}$[还需要经过香港飞行情报区]$_3$。”
ref [Tung Ta-Wei, head representative for China Airlines in Shanghai, told reporters]$_1$, “[presently, the cross-strait charter flights are still not ‘direct flights’ in the true sense of the term]$_2$, [since$_{\text{CONTINGENCY}}$ they still have to pass through the hong kong flight information region]$_3$. ”

source [柳斌杰说]$_1$，[中国出版业下一个发展的重点将是参与国际竞争]$_2$，and$_{\text{EXPANSION}}$[今后双方可就此加大合作力度]$_3$。
ref [Liu Binjie said]$_1$, [a key area of development for the Chinese publishing industry will be participating in international competition]$_2$, [and$_{\text{EXPANSION}}$ in the future the two sides can strengthen their cooperation in this area]$_3$.

The first example is particularly interesting from a discourse point of view as it combines information ordering considerations along with the implicit/explicit expression of discourse relations: not only is the connective *when* missing in Chinese but the two arguments of the connective appeared in reverse order in the English translation of the sentence, with the comma omitted.

In Table 2, we show the numbers and percentages of Chinese/English implicit/explicit mismatches for each relation. We also list the ten connectives that are most frequently associated with the mismatch (i.e., were added to the reference translation), in the format of *connective (# mismatches)* below:

and (341), when (120), while (45), if (37), so that (29), but (23), after (22), so (22), as (21), then (18)

This analysis reveals that the EXPANSION relation is more likely to be implicitly expressed in Chinese, although in other relations this phenomenon is also present.

Connective	Senses	Connective	Senses
而	COMPARISON (7) EXPANSION (2)	又	CONTINGENCY (1) EXPANSION (1)
则	COMPARISON (1) EXPANSION (2)	在...同时	TEMPORAL (3) EXPANSION (2)
如	CONTINGENCY (1) EXPANSION (3)	同时	TEMPORAL (1) EXPANSION (1)

Table 3: Ambiguous Chinese connectives, according to manual annotations in the development CDTB.

A similar mismatch also happens when an English discourse connective is aligned to a punctuation mark in Chinese, illustrated in the following example, where the comma underlined in the source sentence was translated to *and*, thus to an explicit EXPANSION in English:

source [这个总队所属驻边境线中队]$_1$，[大都驻地偏远]$_2$，[自然条件艰苦]$_3$，[信息化建设比较滞后]$_4$。
ref [Most of the contingent's squadrons garrisoned along the border]$_1$ [are stationed in remote areas]$_2$ [where the natural conditions are rough]$_3$ [and$_{\text{EXPANSION}}$ the construction of informatization relatively lags behind]$_4$.

The insertion of the explicit discourse connective *and* makes the use of punctuation between "rough conditions" and "informatization" unnecessary in English. Through our direct projection we found 136 such implicit to explicit transformations with commas and 5 with semicolons. All of them are of the relation EXPANSION, further highlighting the differences in information packaging between the two languages.

4.3 Ambiguity of connectives

Although most of the English discourse connectives identified in the PDTB are not ambiguous, some of the most frequently used ones are (Pitler et al., 2008; Miltsakaki et al., 2008). For example, *while* can signal both TEMPORAL and COMPARISON relations; *since, as* can signal both TEMPORAL and CONTINGENCY relations. Discourse connectives in different languages have different ambiguities; prior work has shown that it is easier to disambiguate the sense of an ambiguous connective when parallel corpora are available (Meyer et al., 2011). The two languages analyzed in Meyer et al. (2011), English and French, are closely related European languages; here we investigate such differences in ambiguities between English and Chinese connectives.

Specifically, using the connectives collected from direct projection, we inspect the relations annotated for these connectives in the Chinese Discourse Treebank development set, and extract connectives such that the majority sense they signal constitutes less than 90% of their total occurrences. Unlike in English where the vast majority of ambiguities are between TEMPORAL and some other sense, we find that all such connectives in Chinese are ambiguous between some relation and EXPANSION. An example of ambiguity between TEMPORAL and EXPANSION is shown below:

source 这样杜伊才能在拿足所有合同内工资的同时$_{\text{TEMPORAL}}$，又乐得清闲，冷眼旁观。
ref Only in this way can Dujkovic sit back and do nothing and look on others disinterestedly when$_{\text{TEMPORAL}}$ getting his full salary per contract.

source 在减少开车出行的同时$_{\text{EXPANSION}}$，还往汽油里掺上从餐馆回收来的食油。

ref While$_{\text{EXPANSION}}$ reducing driving time, they are also mixing gasoline with cooking oil recycled from restaurants.

In the first case, there is a synchrony relation between Dujkovic's "sitting back and doing nothing", and "getting his full salary". In the second case, "reducing driving time" and "mixing gasoline with cooking oil" are a list of methods for saving gasoline.

In Table 3 we list these ambiguous Chinese connectives, their senses and the frequency with which they were annotated. The ambiguities we see here are very different from those in English where the TEMPORAL—CONTINGENCY and COMPARISON—CONTINGENCY ambiguities are most prominent.

5 Predicting discourse relation sense in Chinese

Our analysis so far has revealed considerable differences in the expression of discourse relations in Chinese and English. We now show that projected annotations can be used to disambiguate Chinese discourse connectives despite these differences.

5.1 Learning with unlabeled data

The main idea of learning by projection across parallel corpora is to use a classifier to annotate the English portion of the data, then project the discourse relation sense labels onto the corresponding Chinese sentences. Then a classifier can be trained using features gathered on the Chinese portion of the data.

However, labels gathered from direct projections are not suitable for learning systems without extra processing. If an English connective is aligned to one of the Chinese connectives, we can transfer its label from English to the Chinese connective. However, it is highly likely that a Chinese connective appears in the source sentence but the reference translation used an alternative expression or paraphrase rather than the 100 identified connectives in the PDTB. It is difficult to distinguish through direct projection if an explicit discourse connective in Chinese was expressed implicitly in English or if the Chinese expression was used in a non-discourse sense.

The possibilities described above imply that in our work, we cannot assume that through direct projection we have a fully labeled dataset for discourse connective senses in Chinese. Instead we have a mixture of data with labeled positive examples (when an explicit English connective was aligned to the phrase) and unlabeled examples (where there was no explicit discourse connective in English, so the Chinese expression is either used in a non-discourse sense or is expressed implicitly or using alternative expressions in English, and thus the label is unknown).

Luckily, learning from positive and unlabeled examples, especially for binary classification, is a fairly well studied problem in machine learning (Lee and Liu, 2003; Liu et al., 2003; Elkan and Noto, 2008). We adopt such methods as part of our semi-supervised learning system.

In this work, we propose the following components for relation classification:

(Noisy) data labeling Classify each instance of a possible connective on the English side of the corpus into either *non-discourse use*, or one of TEMPORAL, CONTINGENCY, COMPARISON or EXPANSION. If the English connective signals one of the four relations, transfer the labels to the connectives expressed in the corresponding Chinese sentences through alignments, as described in Section 4.1.

Train sense classifier This classifier is trained only on the Chinese expressions labeled as one of the four main classes of discourse relation. We can train either a binary classifier to predict if a connective expresses a particular relation, or a 4-way classifier which assigns the most probable sense to each connective. The potentially problematic labels for the *non-discourse* class are not used in this stage.

Train discourse use classifier This classifier has to use the potentially problematic data, where we cannot distinguish negative examples from untagged positive examples. The problem is solved as a cascade of classifiers, an approach developed in Elkan and Noto (2008). The idea is to train a noisy classifier that produces a soft score for the data—a probability of being in the class rather than a strict class assignment.

Let y be the *true* discourse use class to be predicted: $y = 1$ for examples of discourse use, and $y = 0$ for examples of non-discourse use. Let l indicate whether the example is *labeled* as discourse use ($l = 1$), or *unlabeled* ($l = 0$, unknown or non-discourse use). First, we use a logistic regression classifier LR to estimate $P(l = 1|y = 1)$. Let's call this estimate e. Using LR, e can be estimated as $\sum_{x \in P} LR(x)/|P|$, where P is the set of the original positively labeled examples, $LR(x)$ is the probability of expression x to be labeled positively. We then use the estimator e to calculate the estimated value of $P(y = 1|l = 0)$, the probability of an expression being discourse use from the original *unlabeled* examples:

$$w = \frac{LR(x)}{e} / \frac{1 - LR(x)}{1 - e}$$

In the second stage, each of the *unlabeled* examples are duplicated, once as a positive example with weight w and once as a negative example with weight $1 - w$. Our second stage classifier—linear-kernel SVM with weights for each example—is trained on the combined set of positive examples (discourse use) and the duplicated version of the unlabeled examples (unknown and non-discourse use class). When w is close to 0.5, the example is practically noise (with labels 0.5 and -0.5) and does not affect the learning of parameters much. Weights closer to 1 practically reassign the originally non-discourse use example to the discourse use class (labels 1 and 0); a weight close to 0 leaves the example as one of the non-discourse use instances (with labels 0 and -1).

Test phase In testing, first the second-stage SVM model for discourse vs. non-discourse use is applied. For only the expression predicted to be discourse connectives (discourse use), we run the sense classifier to do binary or multiway relation classification. Binary classification labels whether a connective signals a particular relation; multiway classification labels one of the five possible classes: *non-discourse use*, TEMPORAL, COMPARISON, CONTINGENCY and EXPANSION. This series of classifiers results in a system that can assign the same labels as the classifiers trained for English.

To complete our presentation of the approach, we now turn to describe the features used to represent instances of potential discourse connectives.

5.2 Features

The following set of features for each expression we need to classify are extracted solely from the Chinese part of the corpus[4]. The syntactic parse trees were obtained automatically (Levy and Manning, 2003).

Connective The connective expressions themselves. The vast majority of connectives (at least in English) are unambiguous, so using the identity of the connective is a hard-to-beat baseline for sense prediction (Pitler et al., 2008).

Categories The syntactic category of the expression itself, as well as that of its parents, and its left and right siblings (if any). These features are adapted from Pitler and Nenkova (2009).

Depth Depth of the expressions's syntactic category in the parse tree for the sentence.

POS bigram Bigram of part-of-speech tags of the entire sentence.

Production pairs Parent-child node category pairs, gathered from subtrees of two ancestors starting from the parent of the expression's self-category. For example, a subtree IP→NP VP would yield the features (IP NP) and (IP VP). Production rules have shown to be effective for implicit discourse relation classification (Lin et al., 2009; Park and Cardie, 2012). This is a less sparse adaptation of such features.

Punctuation This class corresponds to two features. The first feature takes one of the three possible values: if the expression starts a sentence, if there is a punctuation to the immediate left of the expression, or none of above. The second feature has two values corresponding to whether there is a punctuation to the expression's immediate right.

Sequence pairs Left-to-right sequence pairs of node categories, gathered from subtrees of two ancestors starting from the parent of the expression's self-category. For example, a subtree IP→NP VP PU would yield the features (NP VP) and (VP PU).

Size of ancestor nodes The number of children a node has, calculated with three ancestors starting from the parent of the expression's self-category.

# characters The number of Chinese characters in the connective expression.

5.3 Classification results

In this section, we demonstrate the effectiveness of learning discourse relations through parallel data projection and semi-supervised learning. We use the GALE corpus for training and the Chinese Discourse Treebank development set (CDTB-dev) for testing. There are 5,136 training instances and 490 testing instances. In addition, we compare performance with 10-fold cross validation results over CDTB-dev. We obtain predictions for each fold and evaluate on the combined data from all folds, instead of averaging performance for each fold. In this way the results from 10-fold validation and those from the semi-supervised classifier trained on projected data are directly comparable. The LIBLINEAR package (Fan et al., 2008) was used for binary classification (including the discourse use classifier[5]), and SVM-Multiclass (Tsochantaridis et al., 2004) with linear kernel was used for multiway classification.

[4] As a reminder, the list of possible connectives was derived from direct projection after pruning items that occurred only once with a particular part-of-speech. There is a total of 118 such expressions for Chinese.

[5] http://www.csie.ntu.edu.tw/∼cjlin/libsvmtools/#weights_for_data_instances

	Baseline			Cascade			Supervised		
	A	F	P/R	A	F	P/R	A	F	P/R
connective (C)	67.62	51.23	75.45/38.79	70.29	65.88	66.35/65.42	77.96	75.00	75.00/75.00
C+tree depth	69.39	55.36	77.50/43.06	69.80	66.36	65.18/67.59	78.57	75.86	75.34/76.39
C+categories	66.94	52.63	71.43/41.67	70.82	66.97	66.82/67.13	83.67	82.61	77.87/87.96
C+size of ancestor	67.96	52.57	75.65/40.28	71.22	67.59	67.12/68.06	76.12	72.73	73.24/72.22
C+POS bigram	70.00	58.12	75.56/47.22	74.29	70.42	71.43/69.44	81.84	80.35	76.79/84.26
C+punctuation	67.96	52.85	75.21/40.74	73.67	70.48	69.68/71.30	82.65	80.81	78.85/82.87
above, combined	70.61	60.00	75.00/50.00	**75.10**	**71.63**	**71.96/71.30**	82.04	80.00	78.57/81.48

Table 4: Accuracy, F-measure and precision/recall for classifying discourse/non-discourse use of connective expressions, for top features and for the combined feature set.

	5-way Baseline	Projection	5-way Supervised
connective (C)	0.6332	0.6434	0.6114
C+tree depth	0.5959	0.6367	0.6384
C+punctuation	0.6224	0.6776	0.6425
C+size of ancestor	0.5939	0.6469	0.6073
C+categories	0.5837	0.6633	0.6359
C+POS bigram	0.6469	0.6980	0.6714
above, combined	0.6245	**0.7020**	0.6355

Table 5: Multiway discourse relation classification accuracies, for top and the combined features.

Discourse vs. non-discourse To demonstrate the cascade learning component in our system, we first show results from the intermediate stage of the discourse vs. non-discourse prediction task. We compare three systems: our cascade approach for handling noisy labels for non-discourse use, a baseline trained only on the original noisy non-discourse labels (this corresponds to the hard-label performance of the first stage classifier in our approach) and a supervised system trained on CDTB-dev (where predictions are obtained in 10-fold cross validation fashion).

In Table 4 we show the accuracy, precision/recall and F measure for each system, using connective expressions themselves and the five features that gave the best performance on the test set.

Cascade learning achieved a strong boost over the baseline with significant improvements on recall, although it does not perform as well as the fully supervised system. The features most useful for this task are POS bigrams and punctuations; syntactic category features are very useful for the supervised system, but not as useful for the cascade system.

Multiway classification Now we show how our system performs for the complete task of multiway classification of discourse relations for Chinese, recognizing each expression either as *non-discourse use* or one of the four discourse relation senses. We compare our semi-supervised multiway classification system against: *(i)* a baseline system that performs 5-way classification with the noisy labels from direct projection in the GALE data (again corresponding to the hard-label performance of the first stage classifier in our approach); *(ii)* a supervised system for 5-way classification trained on CDTB-dev (where predictions are obtained in 10-fold cross-validation fashion).

Table 5 records the accuracies for the connective expression and the five features performed best for this task. The top features for multiway relation classification, in addition to connectives, are part-of-speech bigrams, punctuations, and syntactic categories.

Notably, without any annotated data on the Chinese side, the projected semi-supervised system outperforms the 5-way supervised system for all but one of the features, and is significantly better when the top features are combined (70.2% vs. 63.55%). This finding justifies the idea and feasibility of using parallel corpora for discourse relation classification.

Binary classification Finally, we present results and the most informative features for binary classification of each relation sense individually. The semi-supervised projection system is compared against a fully supervised binary classification system over 10-fold CDTB-dev, with accuracies and F scores

	Projection		Supervised		Feature set
	A	F	A	F	
COMPARISON	94.49	59.70	96.33	57.14	Connective, categories, size of ancestor, # characters, POS bigram
CONTINGENCY	92.65	41.94	96.33	70.97	Connective, production pairs
EXPANSION	85.10	69.20	87.96	77.20	Connective, categories, production pairs, sequence pairs, POS bigram
TEMPORAL	88.37	48.65	94.08	60.47	Connective, categories, production pairs, sequence pairs

Table 6: Accuracy and F measure for binary classification for each relation, including features that significantly improves performance beyond the identity of the connective itself.

shown in Table 6. The feature sets included are the ones that significantly improve the F measure of a relation compared to that when using the connective expressions alone.

For accuracies, the semi-supervised system is only slightly (1.8-3.7%) below that of the supervised system for three of the four relations. On the other hand, F measures of the semi-supervised system are not as good as the supervised system except for the COMPARISON relation. The feature categories indicate that for Chinese discourse connectives, different feature sets are appropriate for different relations.

6 Conclusion

We investigated the tasks of discourse analysis and recognition without manual annotation. Instead, we used parallel corpora to project automatic annotations available on one side (English) to the other (Chinese). First, we conducted a corpus study which demonstrates the differences in information packaging and discourse organization between English and Chinese. We highlighted the existence of long sentences in Chinese that correspond to multiple sentences in English, mismatches between discourse expressions that are implicit vs. explicit in the two languages, and differences in the ambiguity of discourse connectives. Second, we presented a semi-supervised system that learns to predict discourse relations from the noisy annotations derived from parallel corpora. On the multiway discourse relation classification task, our system outperforms a fully supervised system trained using clean gold-standard annotation in the targeted language.

References

Amal Al-Saif and Katja Markert. 2010. The Leeds Arabic Discourse Treebank: Annotating discourse connectives for Arabic. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.

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

Dipanjan Das and Slav Petrov. 2011. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*, pages 600–609.

Barbara Di Eugenio, Johanna D Moore, and Massimo Paolucci. 1997. Learning features that predict cue usage. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 80–87.

Quang Xuan Do, Yee Seng Chan, and Dan Roth. 2011. Minimally supervised event causality identification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 294–303.

Charles Elkan and Keith Noto. 2008. Learning classifiers from only positive and unlabeled data. In *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 213–220.

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

Kuzman Ganchev, Jennifer Gillenwater, and Ben Taskar. 2009. Dependency grammar induction via bitext projection constraints. In *Proceedings of the Joint Conference of the Annual Meeting of the ACL and the International Joint Conference on Natural Language Processing of the AFNLP (ACL)*, pages 369–377.

Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Natural Language Engineering*, 11(3):311–325, September.

Meixun Jin, Mi-Young Kim, Dongil Kim, and Jong-Hyeok Lee. 2004. Segmentation of Chinese long sentences using commas. In *Proceedings of the SIGHAN Workshop on Chinese Language Processing*, pages 1–8.

Richard Johansson and Pierre Nugues. 2006. A FrameNet-based semantic role labeler for Swedish. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 436–443.

Wee Sun Lee and Bing Liu. 2003. Learning with positive and unlabeled examples using weighted logistic regression. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 3, pages 448–455.

Roger Levy and Christopher D. Manning. 2003. Is it harder to parse Chinese, or the Chinese Treebank? In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 439–446.

Xuansong Li, Niyu Ge, Stephen Grimes, Stephanie Strassel, and Kazuaki Maeda. 2010. Enriching word alignment with linguistic tags. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.

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

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2011. Automatically evaluating text coherence using discourse relations. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*, pages 997–1006.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2014. A PDTB-styled end-to-end discourse parser. *Natural Language Engineering*, 20:151–184, 4.

Bing Liu, Yang Dai, Xiaoli Li, Wee Sun Lee, and Philip S Yu. 2003. Building text classifiers using positive and unlabeled examples. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, pages 179–186.

William C. Mann and Sandra Thompson. 1988. Rhetorical Structure Theory: toward a Functional Theory of Text Organization. *Text*, 8(3):243–281.

Daniel Marcu. 1997. The rhetorical parsing of natural language texts. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 96–103. Association for Computational Linguistics.

Thomas Meyer and Andrei Popescu-Belis. 2012. Using sense-labeled discourse connectives for statistical machine translation. In *Proceedings of the Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation and Hybrid Approaches to Machine Translation (ESIRMT-HyTra)*, pages 129–138.

Thomas Meyer, Andrei Popescu-Belis, Sandrine Zufferey, and Bruno Cartoni. 2011. Multilingual annotation and disambiguation of discourse connectives for machine translation. In *Proceedings of the Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 194–203.

Eleni Miltsakaki, Livio Robaldo, Alan Lee, and Aravind Joshi. 2008. Sense annotation in the Penn Discourse Treebank. In *Proceedings of the International Conference on Computational Linguistics and Intelligent Text Processing (CICLing)*, pages 275–286.

Umangi Oza, Rashmi Prasad, Sudheer Kolachina, Dipti Misra Sharma, and Aravind Joshi. 2009. The Hindi Discourse Relation Bank. In *Proceedings of the Third Linguistic Annotation Workshop*, pages 158–161. Association for Computational Linguistics.

Sebastian Pado and Mirella Lapata. 2005. Cross-linguistic projection of role-semantic information. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT-EMNLP)*, pages 859–866.

Joonsuk Park and Claire Cardie. 2012. Improving implicit discourse relation recognition through feature set optimization. In *Proceedings of the Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 108–112.

Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 186–195.

Emily Pitler and Ani Nenkova. 2009. Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of the ACL-IJCNLP Conference: Short Papers*, pages 13–16.

Emily Pitler, Mridhula Raghupathy, Hena Mehta, Ani Nenkova, Alan Lee, and Aravind Joshi. 2008. Easily identifiable discourse relations. In *Proceedings of the Conference on Computational Linguistics (COLING): Posters*, page 87–90.

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

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

Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2010. Realization of discourse relations by other means: Alternative lexicalizations. In *Proceedings of the Conference on Computational Linguistics (COLING): Posters*, pages 1023–1031.

Frank Schilder. 2002. Robust discourse parsing via discourse markers, topicality and position. *Natural Language Engineering*, 8(3):235–255.

Ioannis Tsochantaridis, Thomas Hofmann, Thorsten Joachims, and Yasemin Altun. 2004. Support vector machine learning for interdependent and structured output spaces. In *Proceedings of the International Conference on Machine Learning (ICML)*, page 104.

Lonneke van der Plas, Paola Merlo, and James Henderson. 2011. Scaling up automatic cross-lingual semantic role annotation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*, pages 299–304.

Ben Wellner and James Pustejovsky. 2007. Automatically identifying the arguments of discourse connectives. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 92–101.

Nianwen Xue and Yaqin Yang. 2011. Chinese sentence segmentation as comma classification. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT): Short Papers*, pages 631–635.

Naiwen Xue, Fei Xia, Fu-dong Chiou, and Marta Palmer. 2005. The penn chinese treebank: Phrase structure annotation of a large corpus. *Natural Language Engineering*, 11(2):207–238, June.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the First International Conference on Human Language Technology Research (HLT)*, pages 1–8.

Deniz Zeyrek and Bonnie Webber. 2008. A discourse resource for Turkish: Annotating discourse connectives in the METU corpus. In *Proceedings of the 6th Workshop on Asian Language Resources*, pages 65–72.

Yuping Zhou and Nianwen Xue. 2012. PDTB-style discourse annotation of Chinese text. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 69–77.