1. Introduction

- Background
  - Discourse structure is fundamental to many text-based applications, such as summarization and question-answering.
  - Constructing discourse resources has been attracting more and more attention in recent years.

- Motivation
  - The general notion of discourse structure mainly consists of discourse unit, connective, structure, relation, and nuclarity. Previous studies on discourse failed to fully express these kinds of information.
  - The Rhetorical Structure Theory (RST) (Mann and Thompson, 1994) represents a discourse as a tree with phrases or clauses as elementary discourse units (EDUs). However, RST ignores the importance of connectives to a great extent.

- Penn Discourse Treebank (PDTB) (Prasad et al., 2008) adopts the predicate-argument view of discourse relation, with discourse connective as predicate and two text spans as its arguments (Carpenter and Bird, 2004). As a result, PDTB is not [partially] if market conditions turn sour.[5]

- The special characteristics of Chinese discourse structure
  - It is difficult to define EDU in Chinese due to the frequent occurrence of the ellipsis of subjects, objects, and predicates, and the lack of functional markers for EDU.
  - The connectives in Chinese omit much more frequently than those in English with about 82.0% vs. 54.9% (Zhou and Li, 2012).
  - The difference in classifying Chinese discourse relations from English (Xing, 2001; Huang and Liao, 2002).
  - The nucleus of a Chinese discourse relation is dynamically determined from the global meaning of a discourse.

2. Related Work

- English
  - Rhetorical Structure Theory Discourse Treebank (RST-DT) (Carlson et al., 2003)
    - 385 documents from the Wall Street Journal
    - 18 relation clauses with 78 finer grained rhetorical relations
  - Penn Discourse Treebank (Prasad et al., 2008)
    - 2160 documents from Penn Treebank-2
    - 18459 explicit and 16224 implicit relations
  - A three level hierarchy of relation

- Chinese
  - English Tradition
    - CUSH Discourse Treebank for Chinese (890 documents) - CTB5.0 - only annotate explicit relations
  - RST: Zhejiang University, financial concept corpus (Wang and Zhou, 2011)
    - The corpus of Chinese complex sentences
      - 600 thousand sentences
      - Only for explicit connective

4. Chinese Discourse Treebank

- Annotator Training
  - A Ph.D. in Chinese linguistics as the supervisor (senior manual annotators) and five undergraduate students in Chinese linguistics as annotators (two pairs).
  - The annotation is done in four phases

- Tagging Strategies
  - Employ a top-down strategy

- Consistency
  - EDU segmentation
  - Explicit or Implicit
  - Explicit connective identification
  - Implicit connective insertion
  - Mononuclear or Multinuclear
  - Nuclarity
  - Structure

- Corpus Statistics
  - 500 newswire articles from Chinese Treebank 6.0
  - 2542 paragraphs (1 CD for one paragraph)
  - 10650 EDUs with an average of 4-5 EDUs per tree
  - 7310 relations, of which 182 are explicit relations (24.8%) and 5498 are implicit relations (75.2%)
  - With the deepest level of 9, most (98.5%) of discourse relations occur in level 1 (2342), level 2/2342, level 3 (4520/4/712), and level 5 (424)
  - 5357 (48.7%) relations are mononuclear relations with 2110 nucleus ahead, while the remaining 3754 relations are multi-nuclear
  - 282 connects, among which 274 (140 can be deleted) appears as explicit connectives and 44 can be inserted in place of implicit connectives

5. Comparison with other Discourse Banks

<table>
<thead>
<tr>
<th>EDU</th>
<th>PDTB</th>
<th>CDUB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Connective - Mark explicit connectives and implicit connectives</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relation Abstract set of relation types, annotate the relation types</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Structure Complete tree, induce connective and bias argument</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nucleus Determine by certain theoretical relation</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Characteristic
  - Clear defined, start of combination, one relation per two EDUs
  - Predicator-argument tree, one relation can two arguments
  - Clear defined from these aspects, one relation has two or more EDUs
  - Mark whether connective can be deleted without changing the discourse relation, insert implicit connective with word insertion and bad argument

6. Preliminary Experimentation

- The adoption of tree structure conforms to traditional Chinese discourse theories and practice
- The hierarchical structure of connectives indicates the hierarchical structure of discourse units
- Discourse Relation
  - We adopt the connective itself to express the discourse relation
  - For the abstraction of discourse relations, we leave it in a later separate stage

- Nucleus and Satellite
  - Discourse relations may be either mononuclear or multi-nuclear
  - We adopt the connective itself and select the unit which can stand for the relationship with other discourse units in a discourse.

7. Conclusions

- Propose a Connective-driven Dependency Tree (CDT) structure as a representation scheme for Chinese discourse structure
- Describe CDT in detail from various perspectives, such as EDU, connective, structure, relation and nuclarity
- Annotate 500 CTBU corpus guided by CDT scheme
- Evaluation of the CDTB corpus on EDU recognition justifies the appropriateness of the CDT scheme to Chinese discourse structure and the usefulness of our CDTB corpus
- In the future work, we will focus on enlarging the scale of the corpus annotation and developing a complete Chinese discourse parser

Figure 2: CDT representation of Example (3)

- Discourse Structure
  - The adoption of tree structure conforms to traditional Chinese discourse theories and practice
  - The hierarchical structure of connectives indicates the hierarchical structure of discourse units

- Discourse Relation
  - We adopt the connective itself to express the discourse relation
  - For the abstraction of discourse relations, we leave it in a later separate stage

- Nucleus and Satellite
  - Discourse relations may be either mononuclear or multi-nuclear
  - We adopt the connective itself and select the unit which can stand for the relationship with other discourse units in a discourse.