Detection, Disambiguation, and Argument Identification of Chinese Discourse Connectives

Yong-Siang Shih
Advisor: Hsin-Hsi Chen, Ph.D.

1. **Introduction**
2. Related Work
3. Datasets
4. Methods & Experiments
5. Conclusion

Introduction

- Discourse analysis is important for language understanding.

儘管浦東新區制訂的法規性文件有些比較“粗”，有些還只是暫行規定，有待在實踐中逐步完善，

但這種法制緊跟經濟和社會活動的做法，受到了國內外投資者的好評，他們認為，到浦東新區投資辦事有章法，講規矩，利益能得到保障。

讓步關係

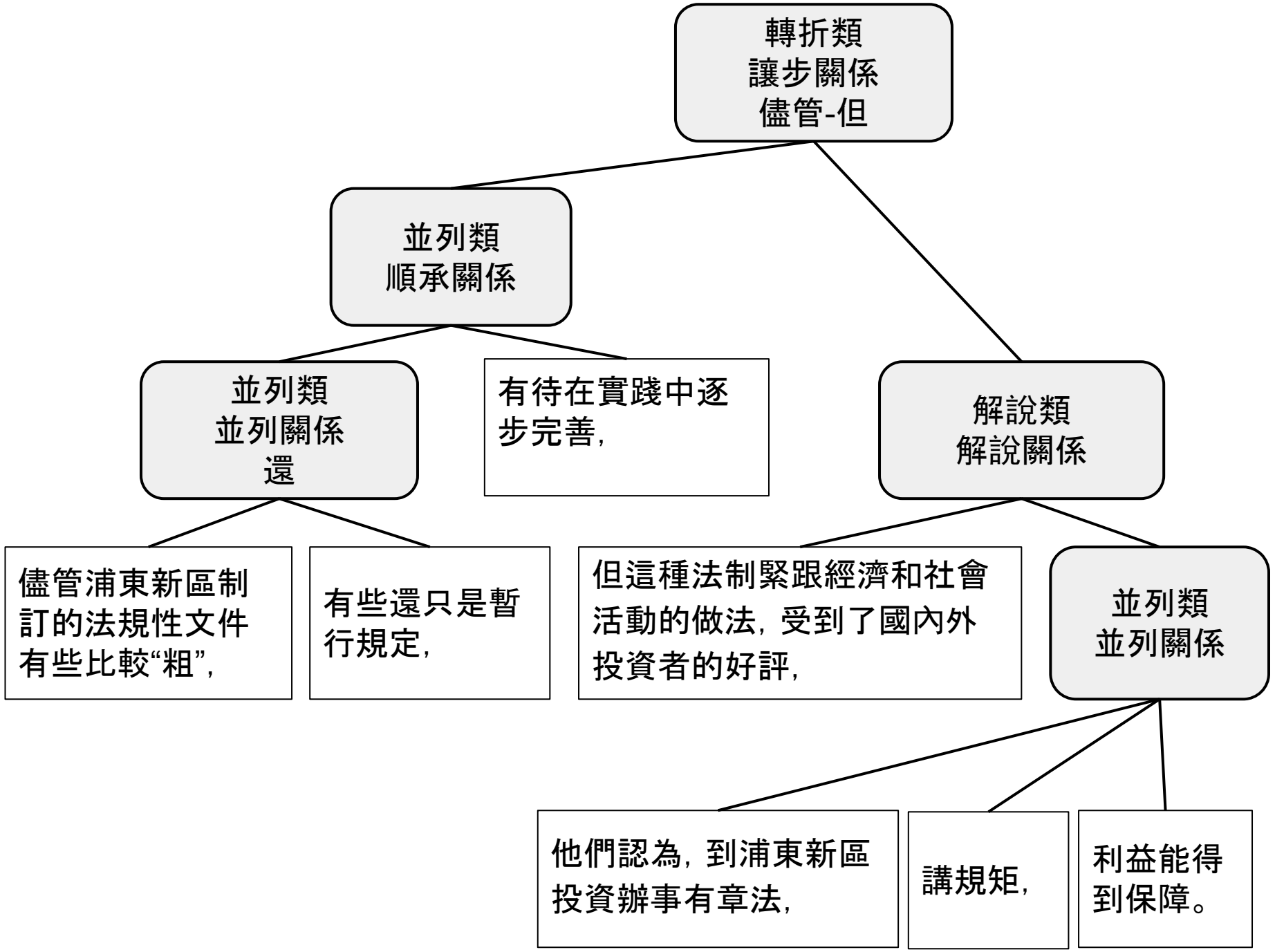
儘管浦東新區制訂的法規性文件有些比較“粗”，有些還只是暫行規定，有待在實踐中逐步完善，

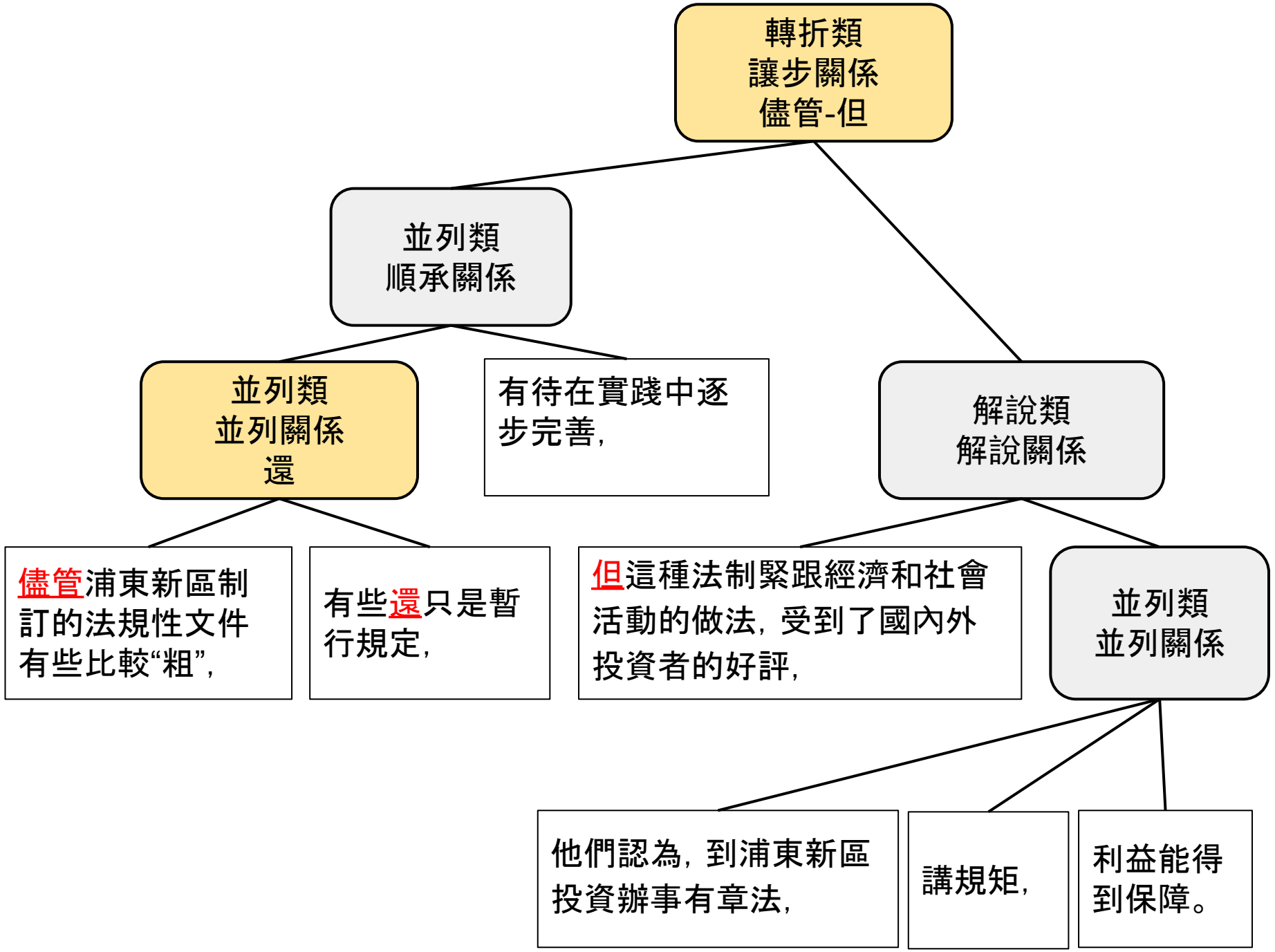
但這種法制緊跟經濟和社會活動的做法，受到了國內外投資者的好評，他們認為，到浦東新區投資辦事有章法，講規矩，利益能得到保障。

顯性關係

儘管浦東新區制訂的法規性文件有些比較“粗”，有些還只是暫行規定，有待在實踐中逐步完善，

但這種法制緊跟經濟和社會活動的做法，受到了國內外投資者的好評，他們認為，到浦東新區投資辦事有章法，講規矩，利益能得到保障。


轉折類
讓步關係
儘管-但
並列類
順承關係
並列類
並列關係
還
有待在實踐中逐步完善，
解說類
解說關係
儘管浦東新區制訂的法規性文件有些比較“粗”，
有些還只是暫行規定，
但這種法制緊跟經濟和社會活動的做法，受到了國內外投資者的好評，
並列類
並列關係
他們認為，到浦東新區投資辦事有章法，
講規矩，
利益能得到保障。

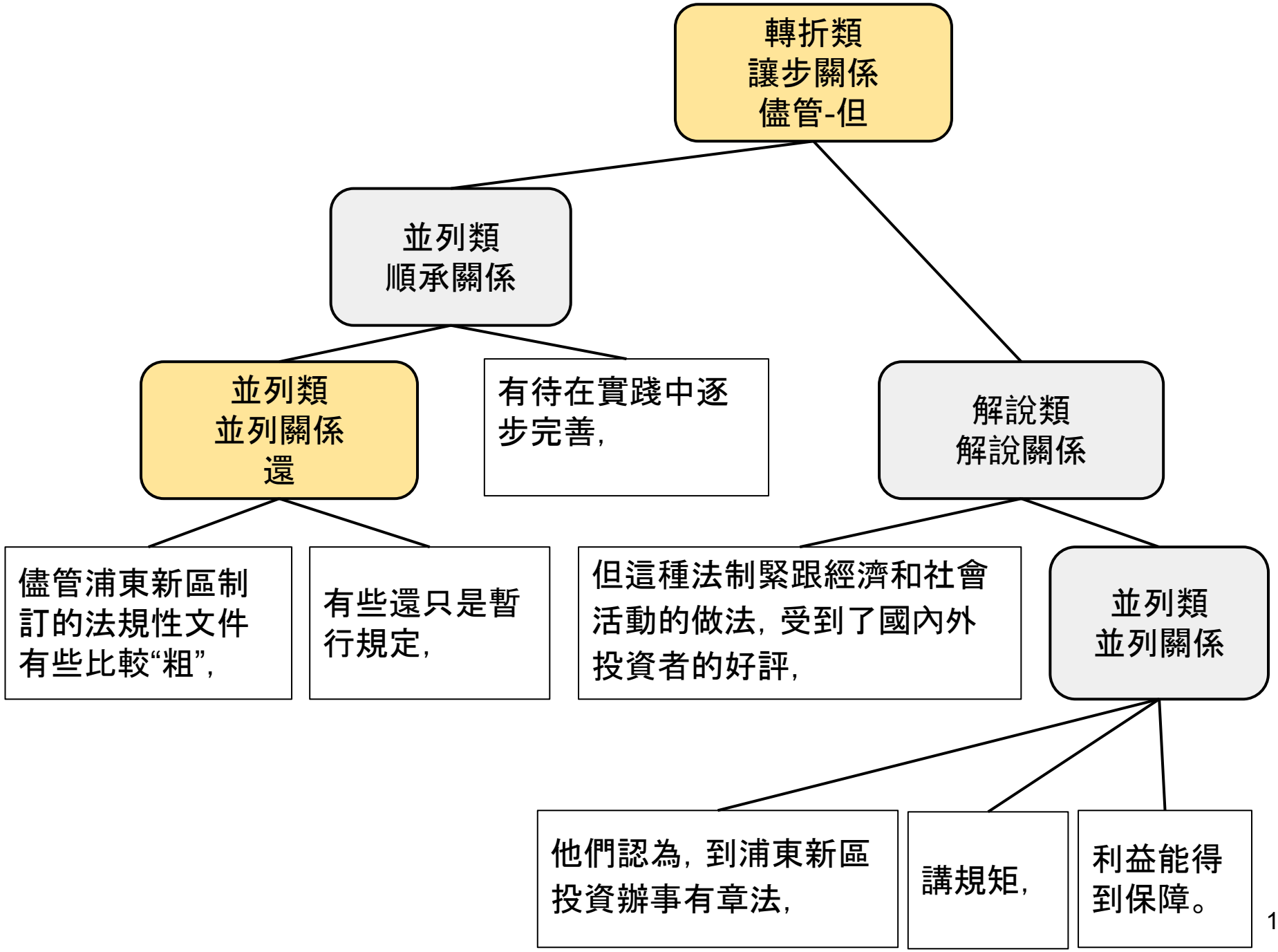

轉折類
讓步關係
儘管-但
並列類
順承關係
並列類
並列關係
還
有待在實踐中逐步完善，
解說類
解說關係
儘管浦東新區制訂的法規性文件有些比較“粗”，
有些還只是暫行規定，
但這種法制緊跟經濟和社會活動的做法，受到了國內外投資者的好評，
並列類
並列關係
他們認為，到浦東新區投資辦事有章法，
講規矩，
利益能得到保障。

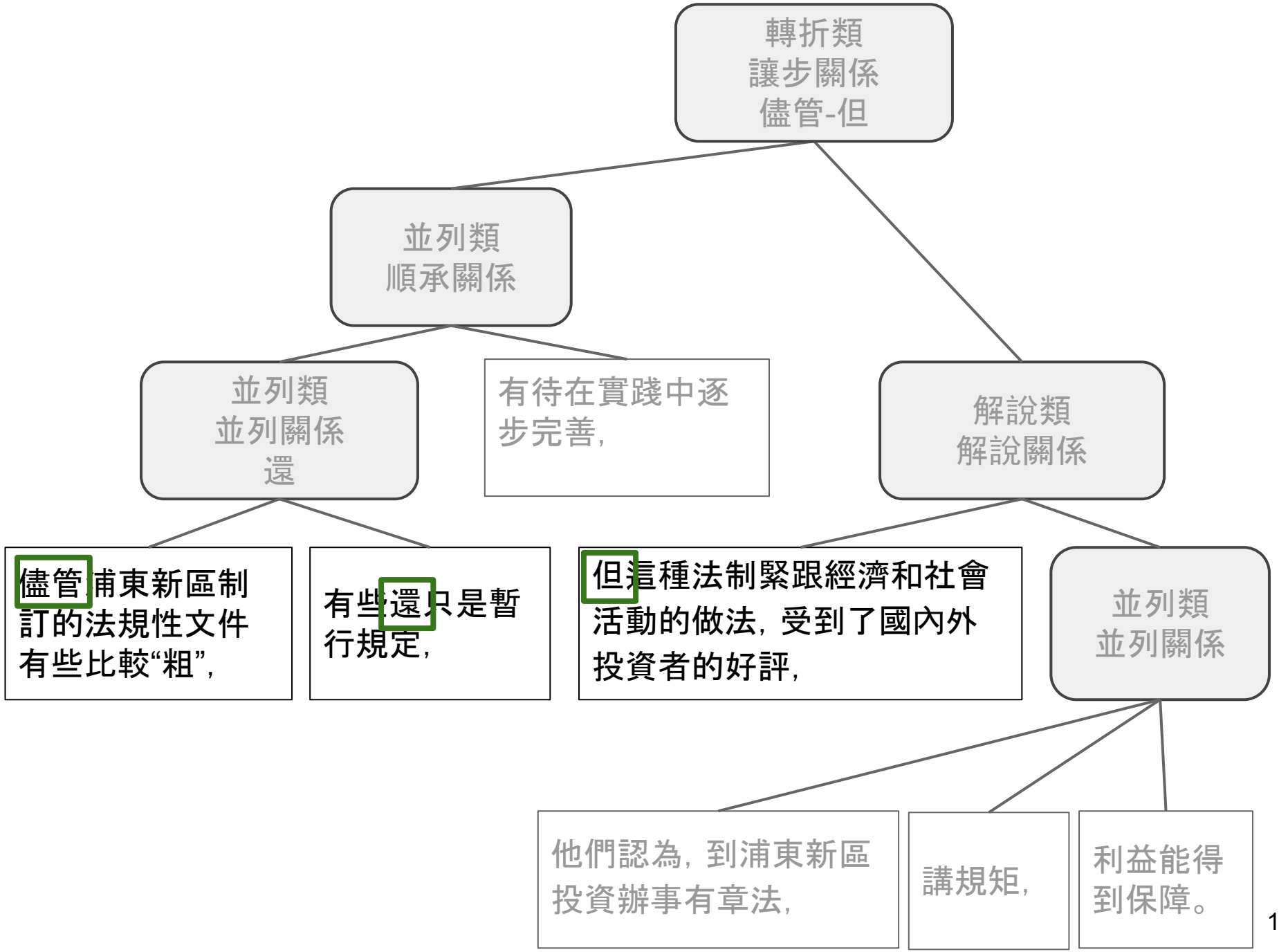
Issues for Chinese Discourse Analysis

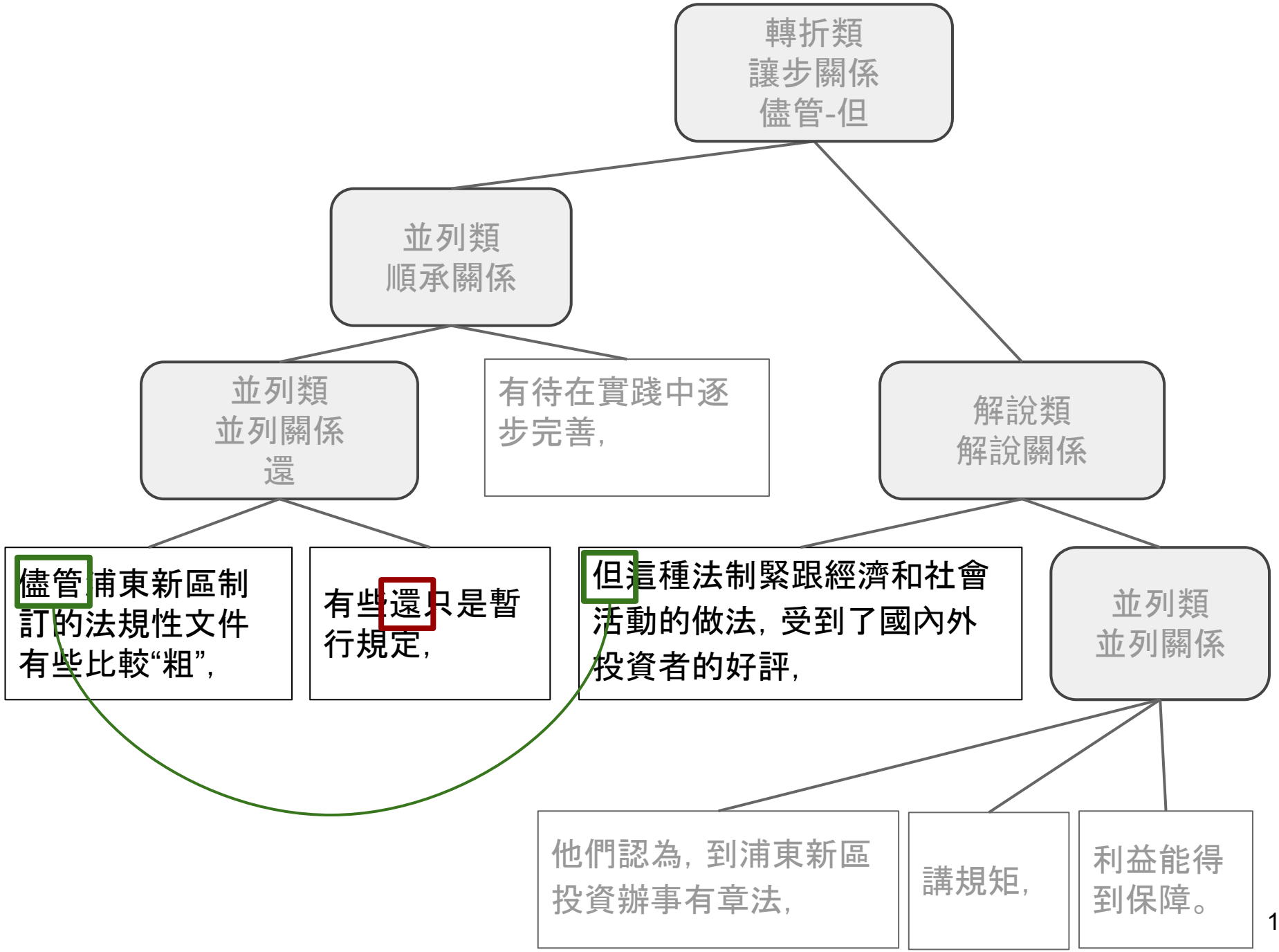
- More varieties for discourse connectives
- Ambiguous linking between connective components
- Unclear sentence boundaries

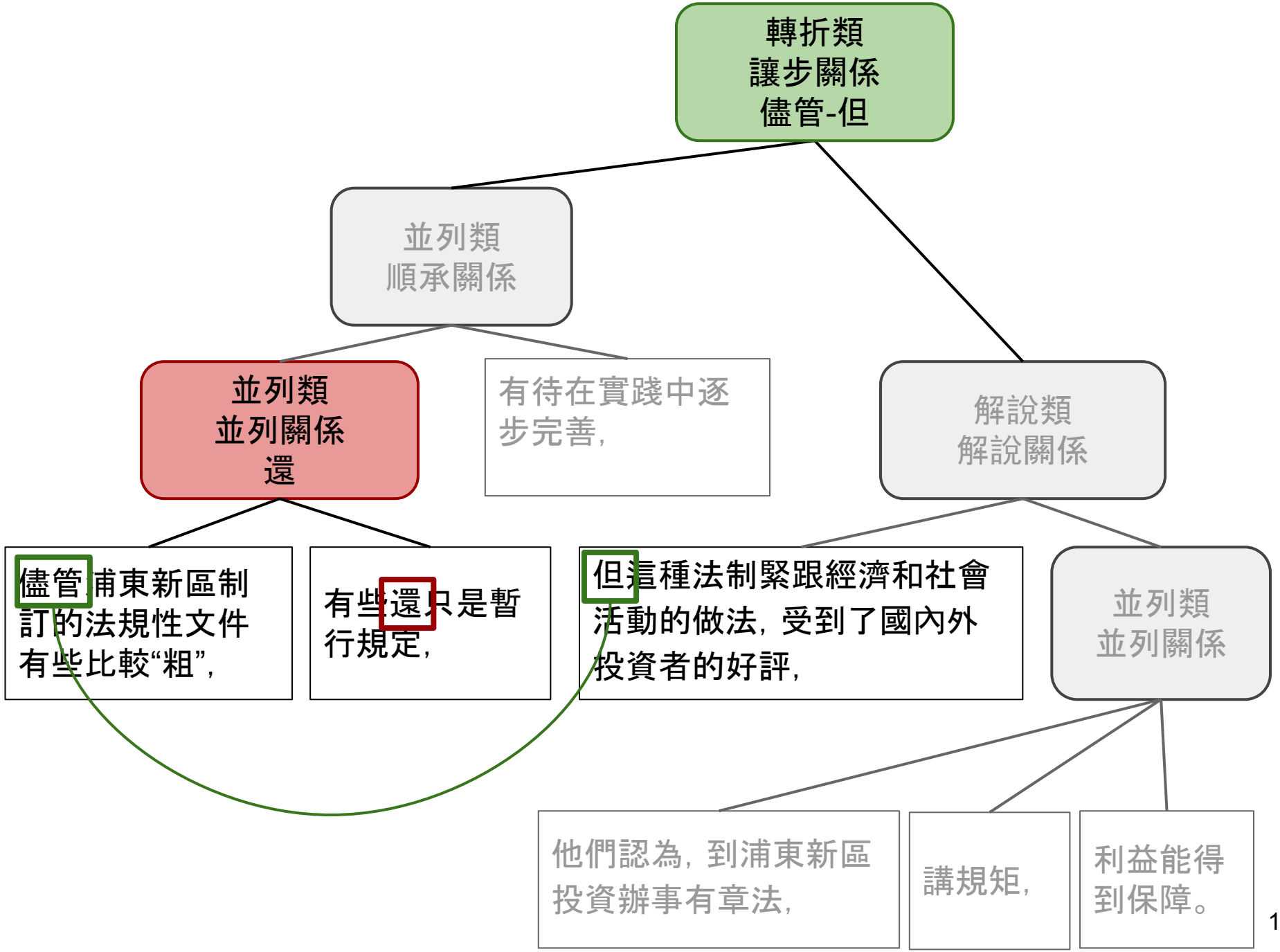
Issues for Chinese Discourse Analysis

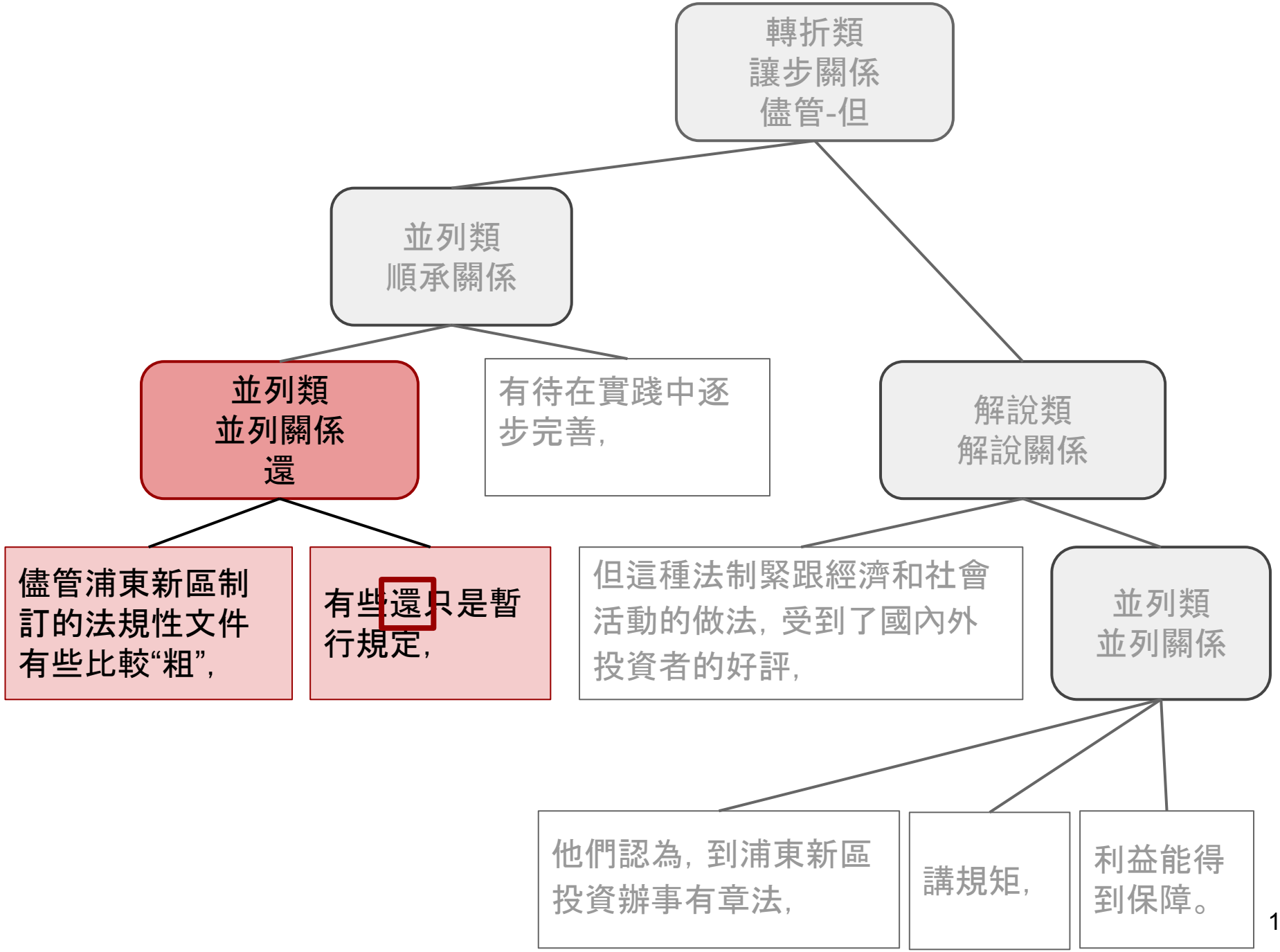
- Ambiguous linking between connective components

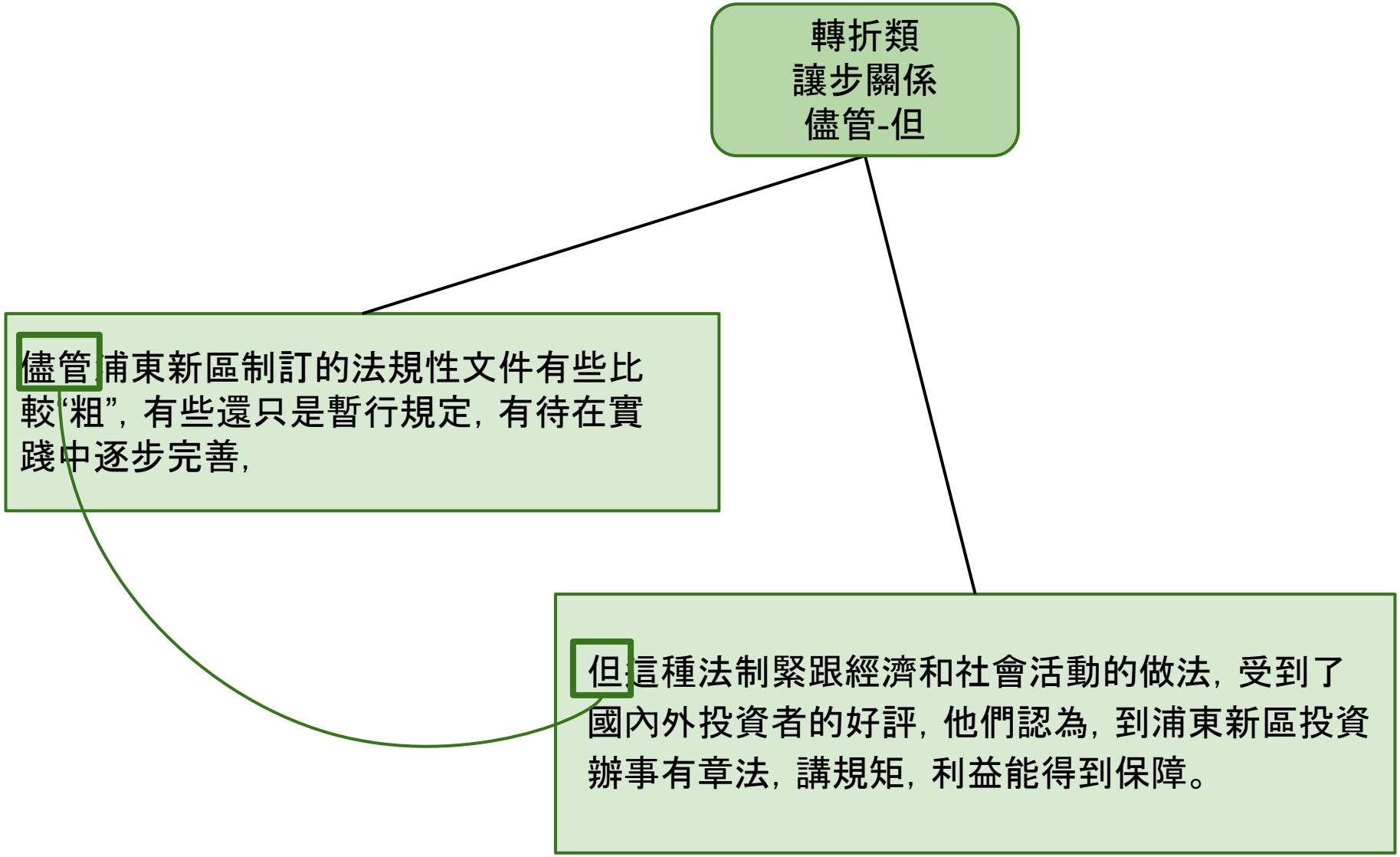

轉折類
讓步關係
儘管-但
並列類
順承關係
並列類
並列關係
還
有待在實踐中逐步完善，
解說類
解說關係
儘管浦東新區制訂的法規性文件有些比較“粗”，
有些還只是暫行規定，
但這種法制緊跟經濟和社會活動的做法，受到了國內外投資者的好評，
並列類
並列關係
他們認為，到浦東新區投資辦事有章法，
講規矩，
利益能得到保障。


轉折類
讓步關係
儘管-但
並列類
順承關係
並列類
並列關係
還
有待在實踐中逐步完善，
解說類
解說關係
儘管浦東新區制訂的法規性文件有些比較“粗”，
有些還只是暫行規定，
但這種法制緊跟經濟和社會活動的做法，受到了國內外投資者的好評，
並列類
並列關係
他們認為，到浦東新區投資辦事有章法，
講規矩，
利益能得到保障。


轉折類
讓步關係
儘管-但
並列類
順承關係
並列類
並列關係
還
有待在實踐中逐步完善，
解說類
解說關係
儘管浦東新區制訂的法規性文件有些比較“粗”，
有些還只是暫行規定，
但這種法制緊跟經濟和社會活動的做法，受到了國內外投資者的好評，
並列類
並列關係
他們認為，到浦東新區投資辦事有章法，
講規矩，
利益能得到保障。


轉折類
讓步關係
儘管-但
並列類
順承關係
並列類
並列關係
還
有待在實踐中逐步完善，
解說類
解說關係
儘管浦東新區制訂的法規性文件有些比較“粗”，
有些還只是暫行規定，
但這種法制緊跟經濟和社會活動的做法，受到了國內外投資者的好評，
並列類
並列關係
他們認為，到浦東新區投資辦事有章法，
講規矩，
利益能得到保障。


轉折類
讓步關係
儘管-但
並列類
順承關係
並列類
並列關係
還
有待在實踐中逐步完善，
解說類
解說關係
儘管浦東新區制訂的法規性文件有些比較“粗”，
有些還只是暫行規定，
但這種法制緊跟經濟和社會活動的做法，受到了國內外投資者的好評，
並列類
並列關係
他們認為，到浦東新區投資辦事有章法，
講規矩，
利益能得到保障。


轉折類
讓步關係
儘管-但
儘管浦東新區制訂的法規性文件有些比較“粗”，有些還只是暫行規定，有待在實踐中逐步完善，
但這種法制緊跟經濟和社會活動的做法，受到了國內外投資者的好評，他們認為，到浦東新區投資辦事有章法，講規矩，利益能得到保障。

1. Introduction
2. **Related Work**
3. Datasets
4. Methods & Experiments
5. Conclusion

Related Work for English Discourse Analysis

- Discourse corpora
- Discourse connective identification
- Relation type disambiguation
- Connective argument extraction
- Sentence level discourse parsing
- Document level discourse parsing

Related Work for Chinese Discourse Analysis

- Discourse corpora
  - HIT-CDTB (Zhang et al., 2014)
  - Chinese Discourse TreeBank 0.5 (Zhou and Xue, 2012, 2015)
  - DTBC (Zhou et al., 2014)
  - CDTB (Li, Feng et al., 2014)
- Discourse connective identification
  - T'sou et al. (1999, 2000)
  - Chan et al. (2000)
  - Hu et al. (2009)
  - Zhou et al. (2012)
  - Li, Carpuat et al. (2014)
  - Li et al. (2015)

Related Work for Chinese Discourse Analysis

- Linking resolution
  - Hu et al. (2011)
- Relation type disambiguation
  - Li, Carpuat et al. (2014)
  - Li et al. (2015)
- Discourse Structure Prediction
  - Huang and Chen (2012)

1. Introduction
2. Related Work
3. **Datasets**
4. Methods & Experiments
5. Conclusion

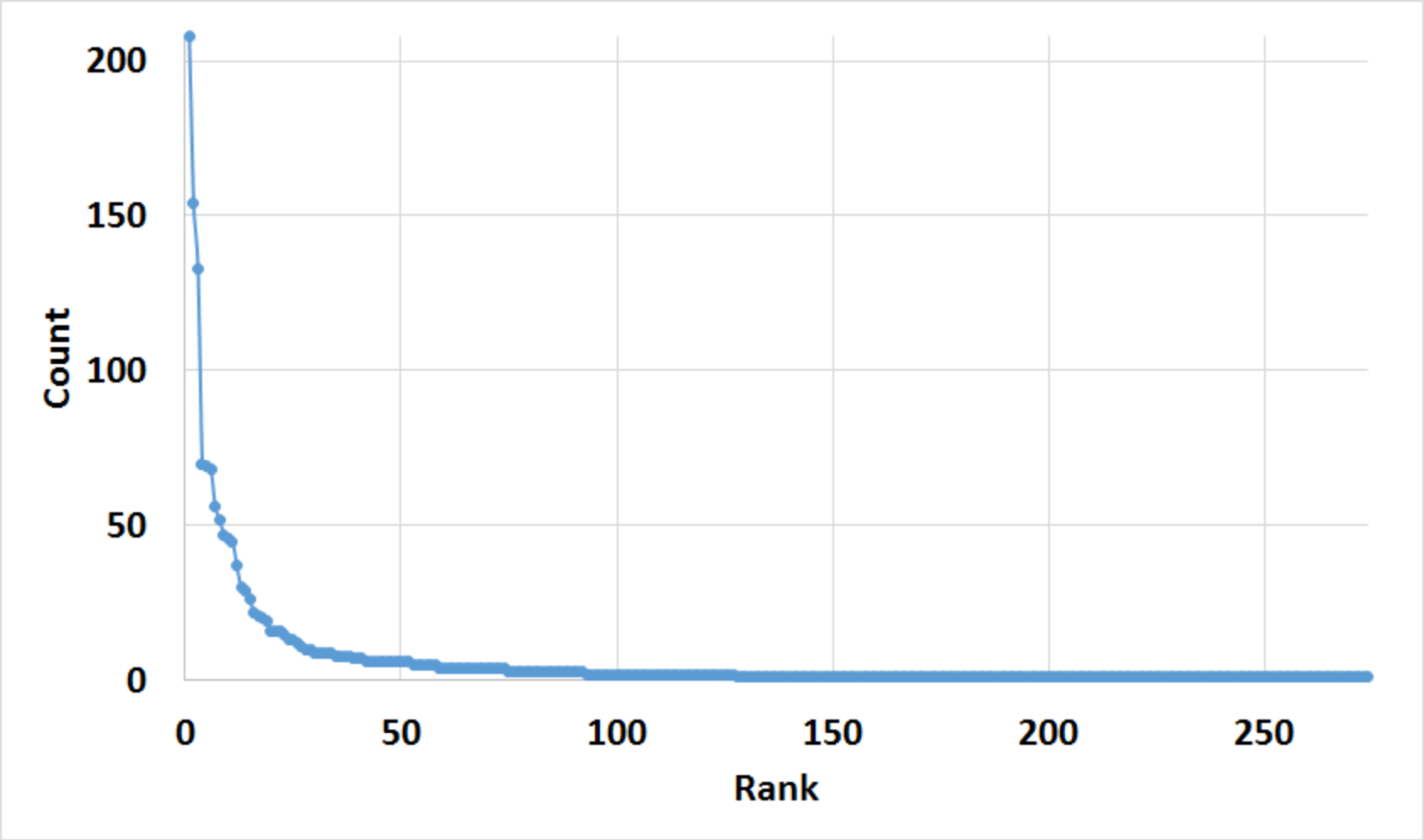
CDTB

- 500 articles selected from the Chinese Treebank (CTB) (Xue et al., 2005)
- 2,342 paragraphs

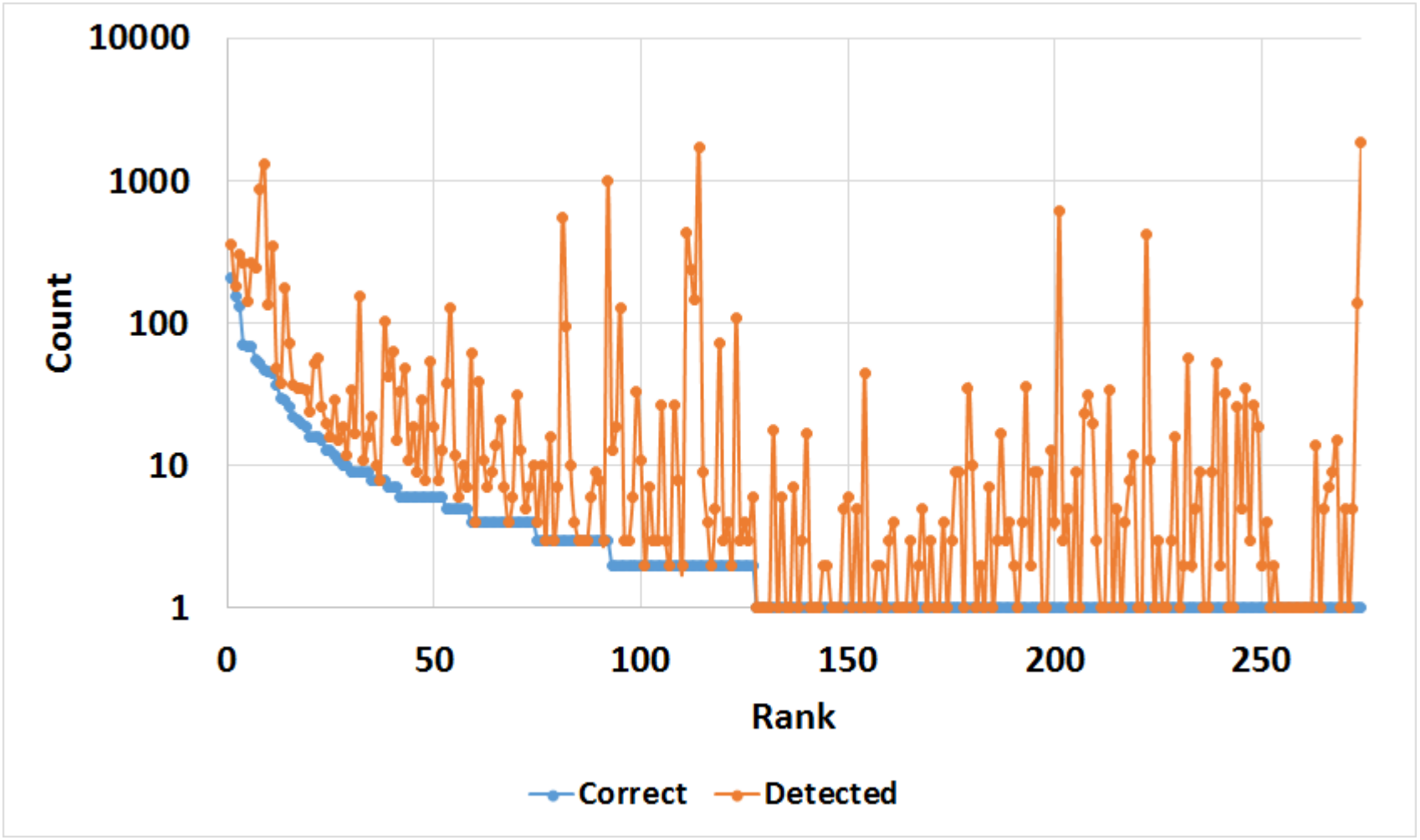
CDTB

- 227 classes of connective components
  - 2,131 connective component instances
- 274 classes of connectives
  - 1,813 connective instances

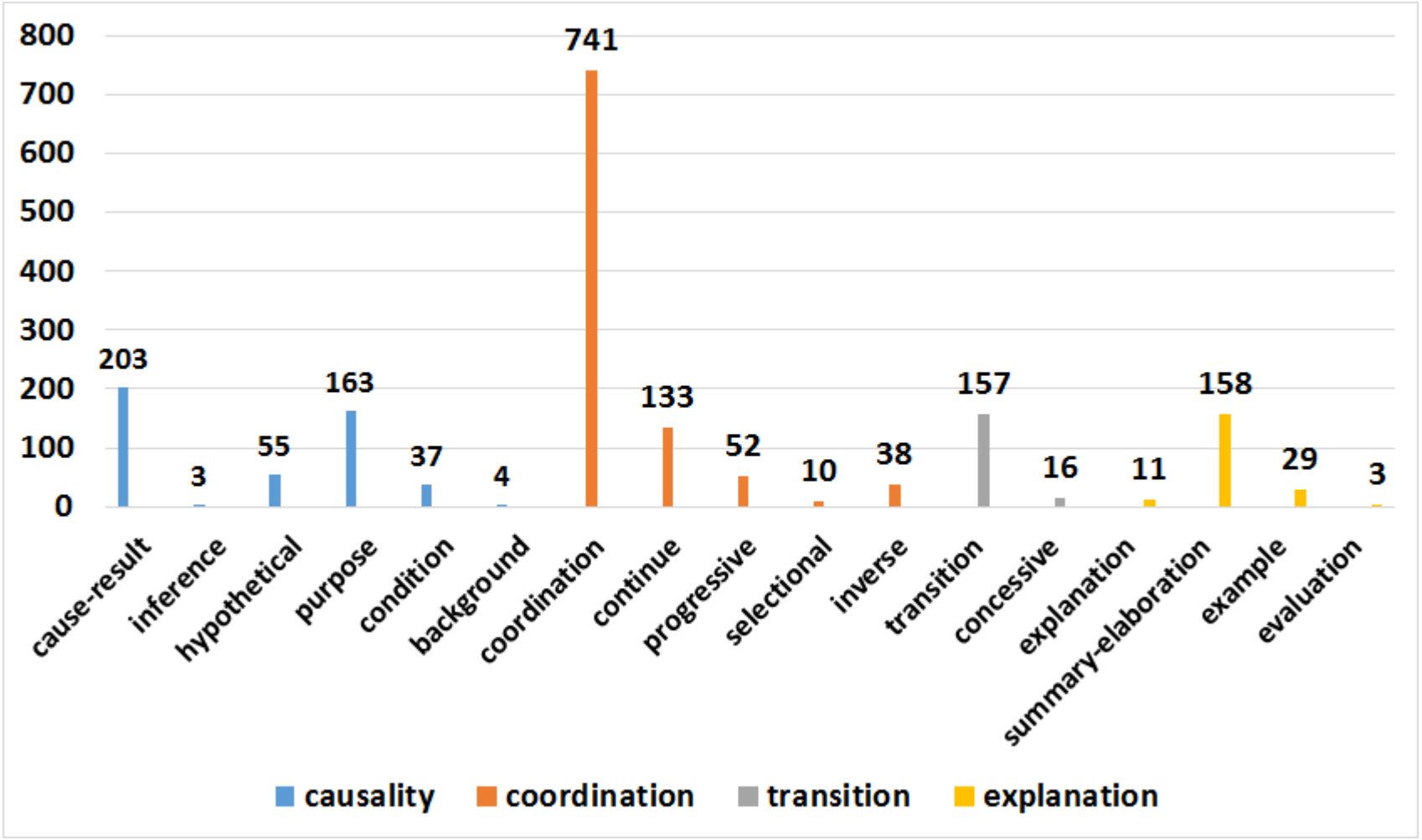
Frequencies of Connectives



Frequencies of Detected Connective Candidates



Relation Type Distribution for Explicit Relations



NTU PN-Gram Corpus (Yu et al., 2012)

- POS-tagged Chinese texts selected from the ClueWeb09 dataset (Callan et al., 2009)
- A subset extracted by Huang et al. (2014)
  - Each sentence must have exactly two clauses.
  - Each sentence must have exactly one instance of connective.
  - Each of the two clauses must not have more than 20 Chinese characters.
- 21,217,147 unique sentences
- 326,996,602 tokens

Word Embeddings

- Fixed-dimensional vectors for words
- Similar representations for words in similar contexts

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. In Proceedings of Workshop at ICLR, 2013.

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global vectors for word representation." Proceedings of the Empirical Methods in Natural Language Processing (EMNLP 2014) 12 (2014): 1532-1543.

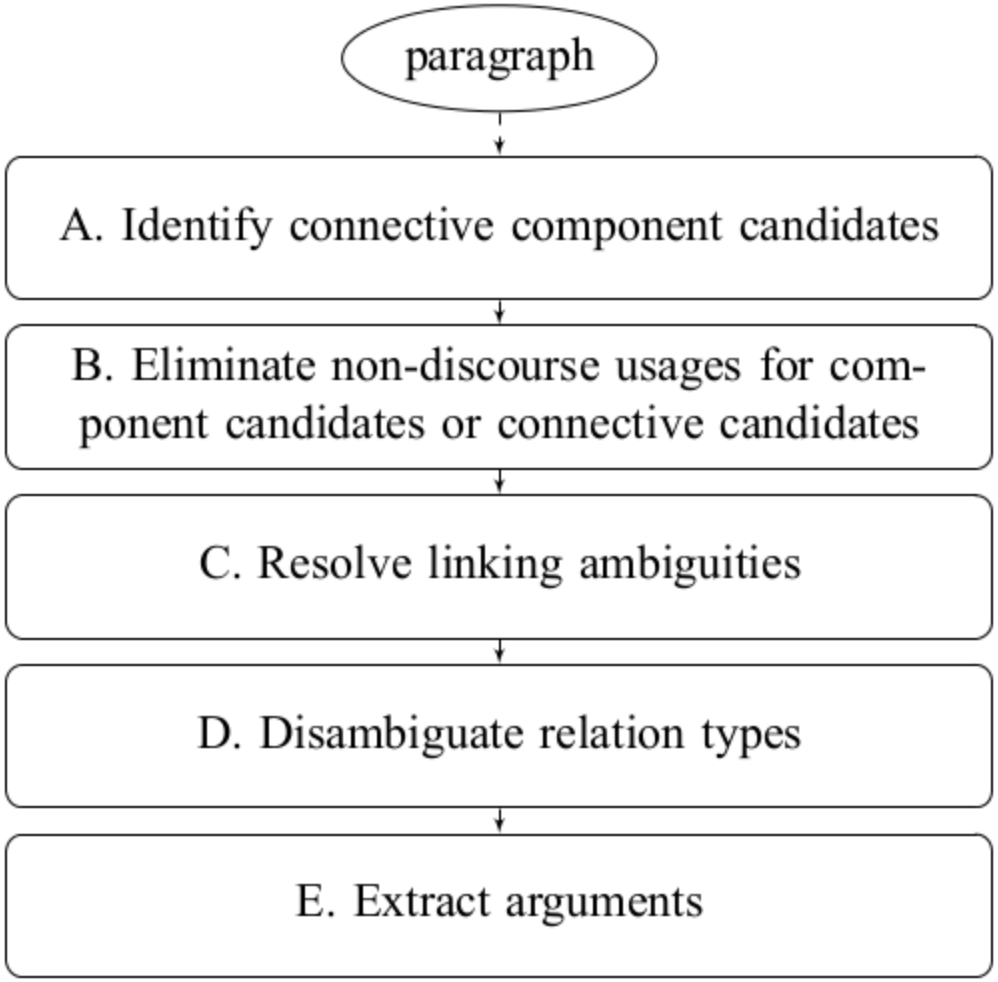
Linking Directions of Connective Components

- 一方面　　Forward/Couple
- 又　　Backward/Couple
- 也　　Backward/Couple
- 以及　　Backward

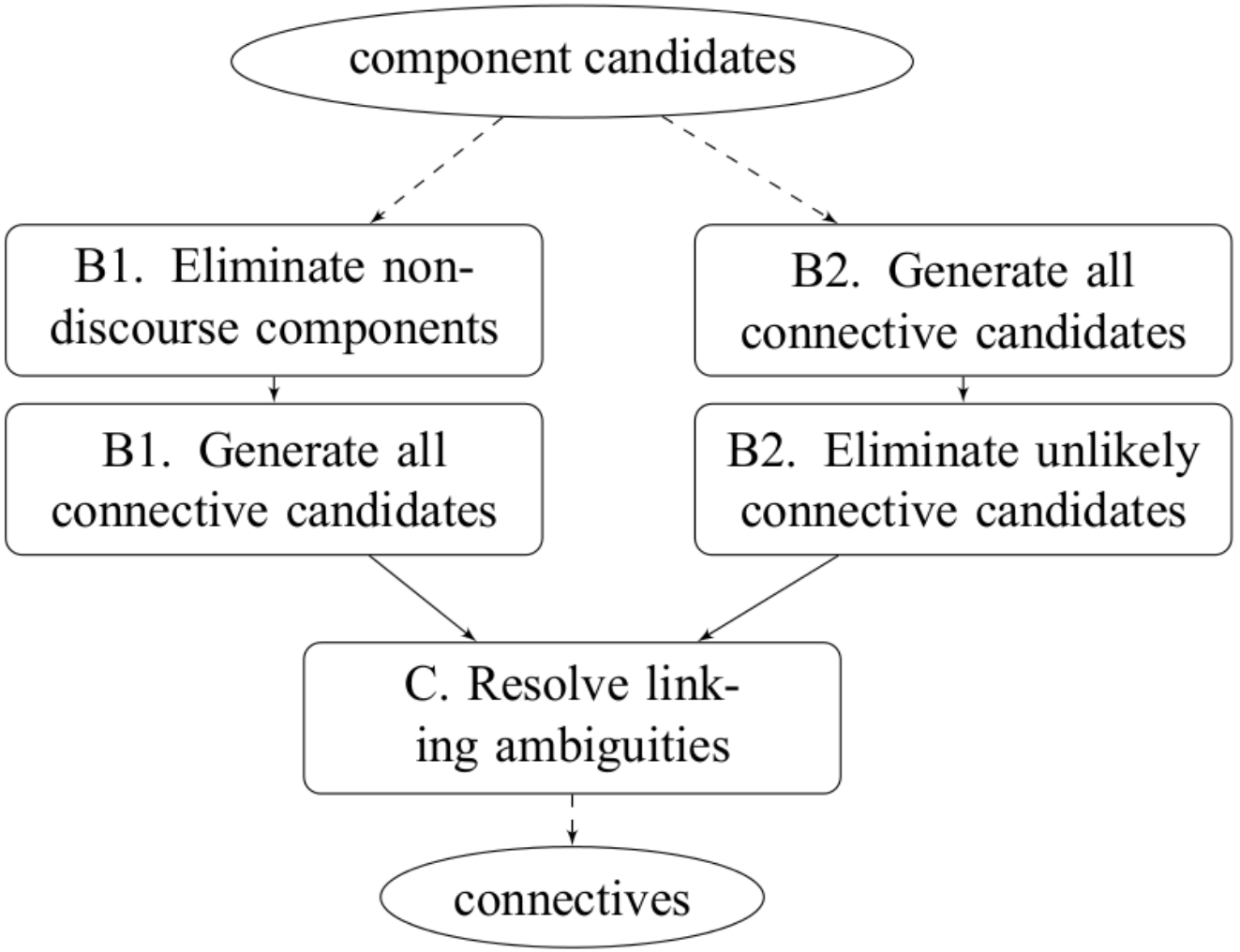
Hen-Hsen Huang, Tai-Wei Chang, Huan-Yuan Chen and Hsin-Hsi Chen (2014). “Interpretation of Chinese Discourse Connectives for Explicit Discourse Relation Recognition.” Proceedings of the 25th International Conference on Computational Linguistics, 23-29 August 2014, Dublin, Ireland, 632–643.

1. Introduction
2. Related Work
3. Datasets
4. **Methods & Experiments**
5. Conclusion

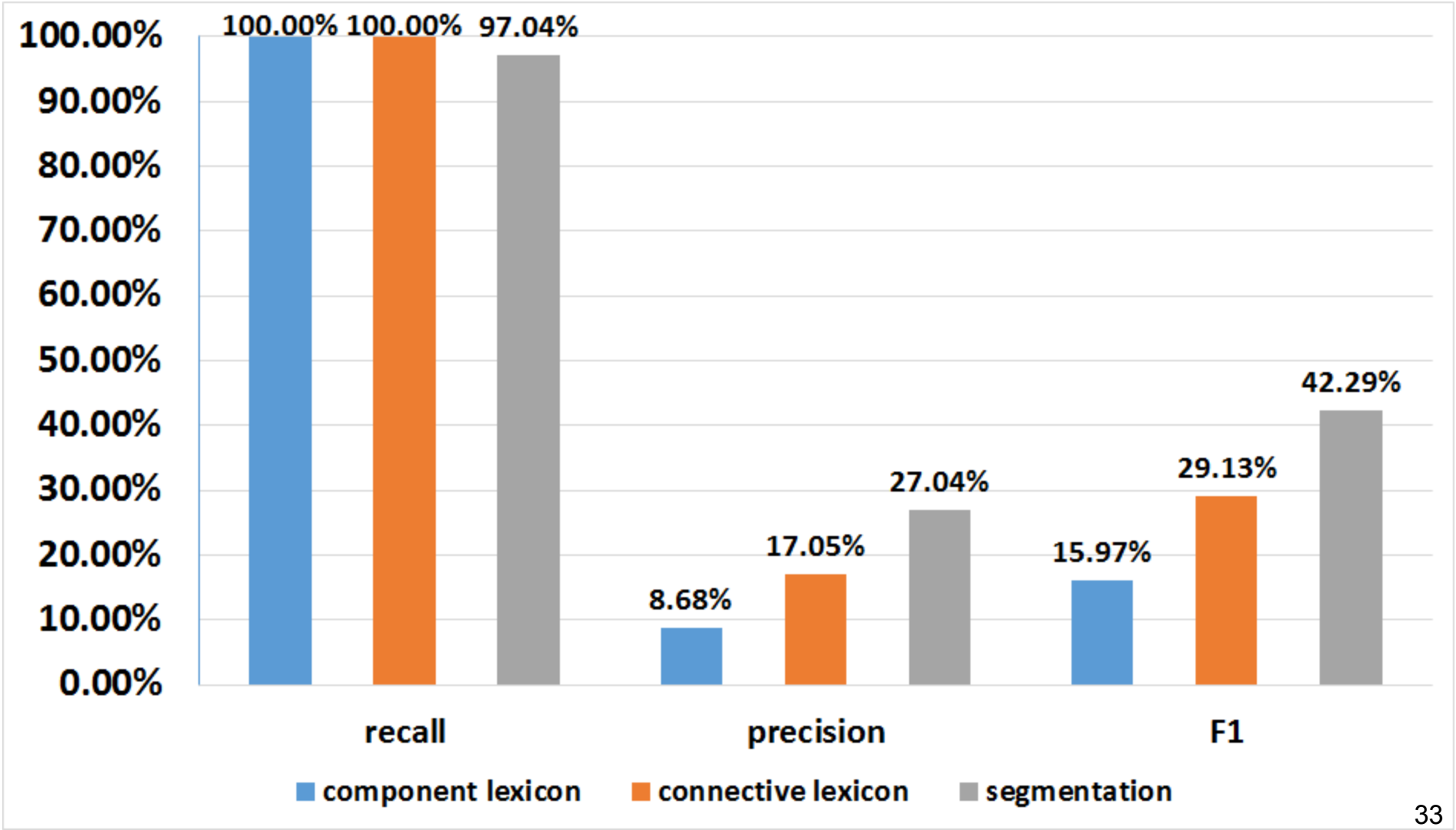
System Overview



Alternative Approaches



Component Candidate Extraction



Match Complete Tokens

- 當前/ 經濟/ 的/ 關鍵/ 不/ 是/ 爭取/ 更/ 高/ 的/ 增長/ 速度/ ,/ 而是/ 提高/ 效益/ 。

- 雙方/ 表示/ 希望/ 在/ 和平/ 計劃/ 的/ 基礎/ 上/ 解決/ 問題/ 。

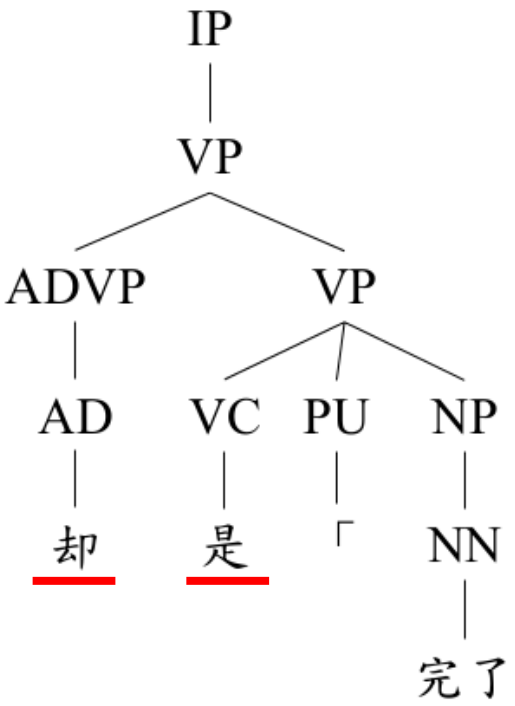
Discourse Usage Disambiguation

- Logistic Regression classifier
  - On component level
  - On connective level

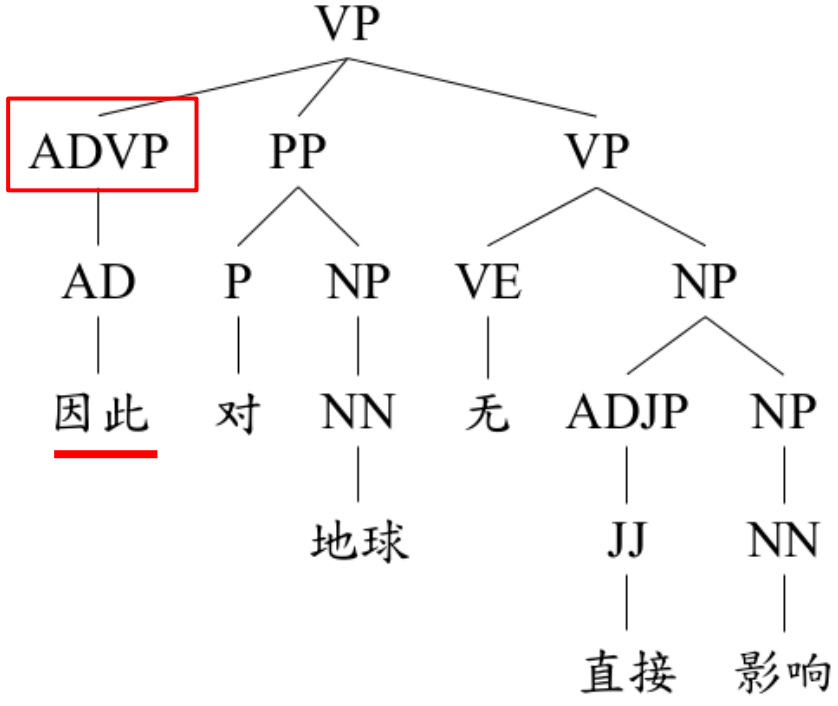
Features for Components

- P&N (Pitler and Nenkova, 2009)
  - self-category, parent, left-sibling, right-sibling
- POS
  - left token, right token, itself
- NUM
  - linking ambiguity, left distance, right distance
- SKIPGRAM
  - left token, right token, itself

P&N

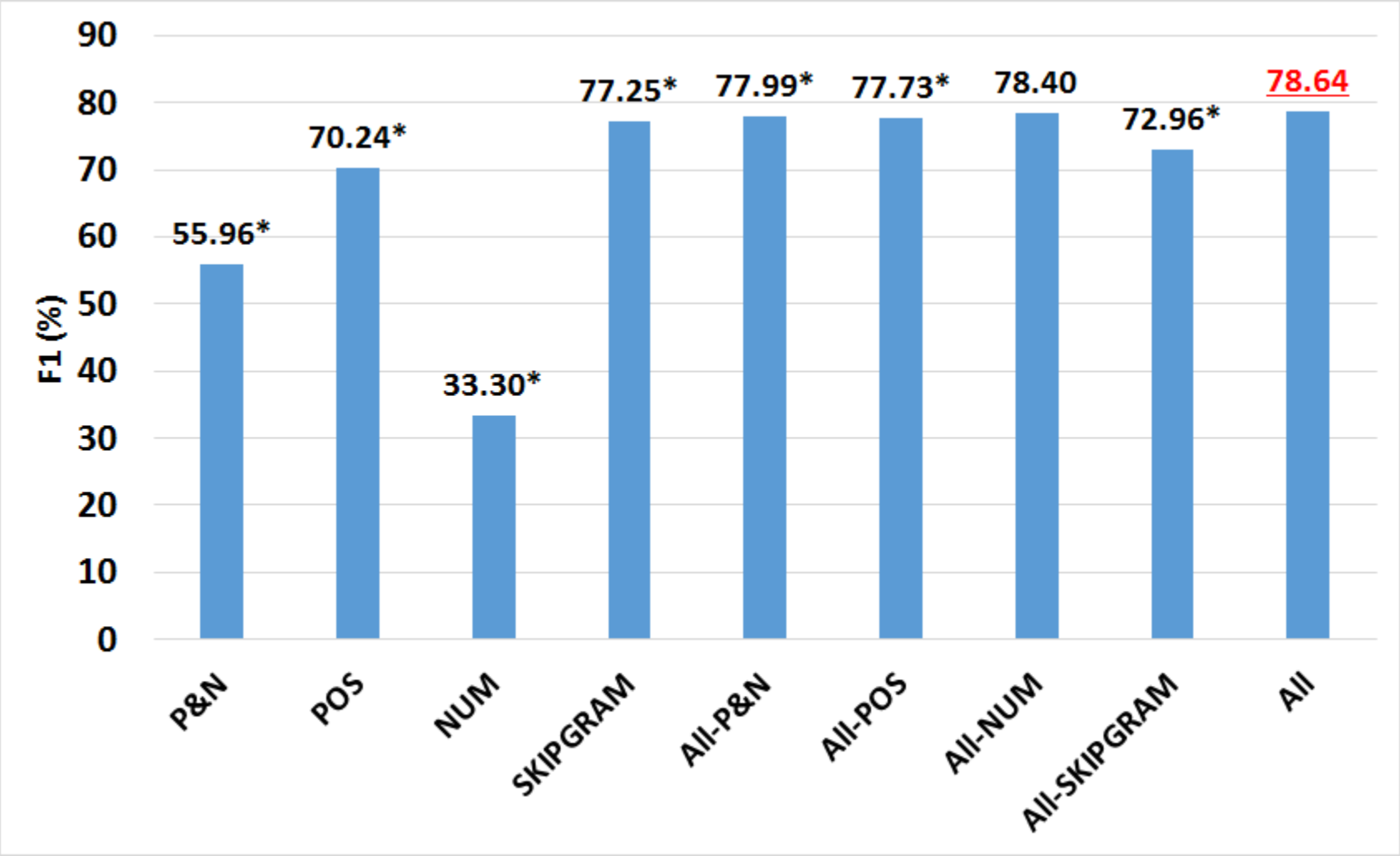


A sub-parsing tree for 却是.



A sub-parsing tree for 因此.

Component Identification with Different Features (B1)



* statistically significant with Wilcoxon signed-rank test at confidence level 0.05

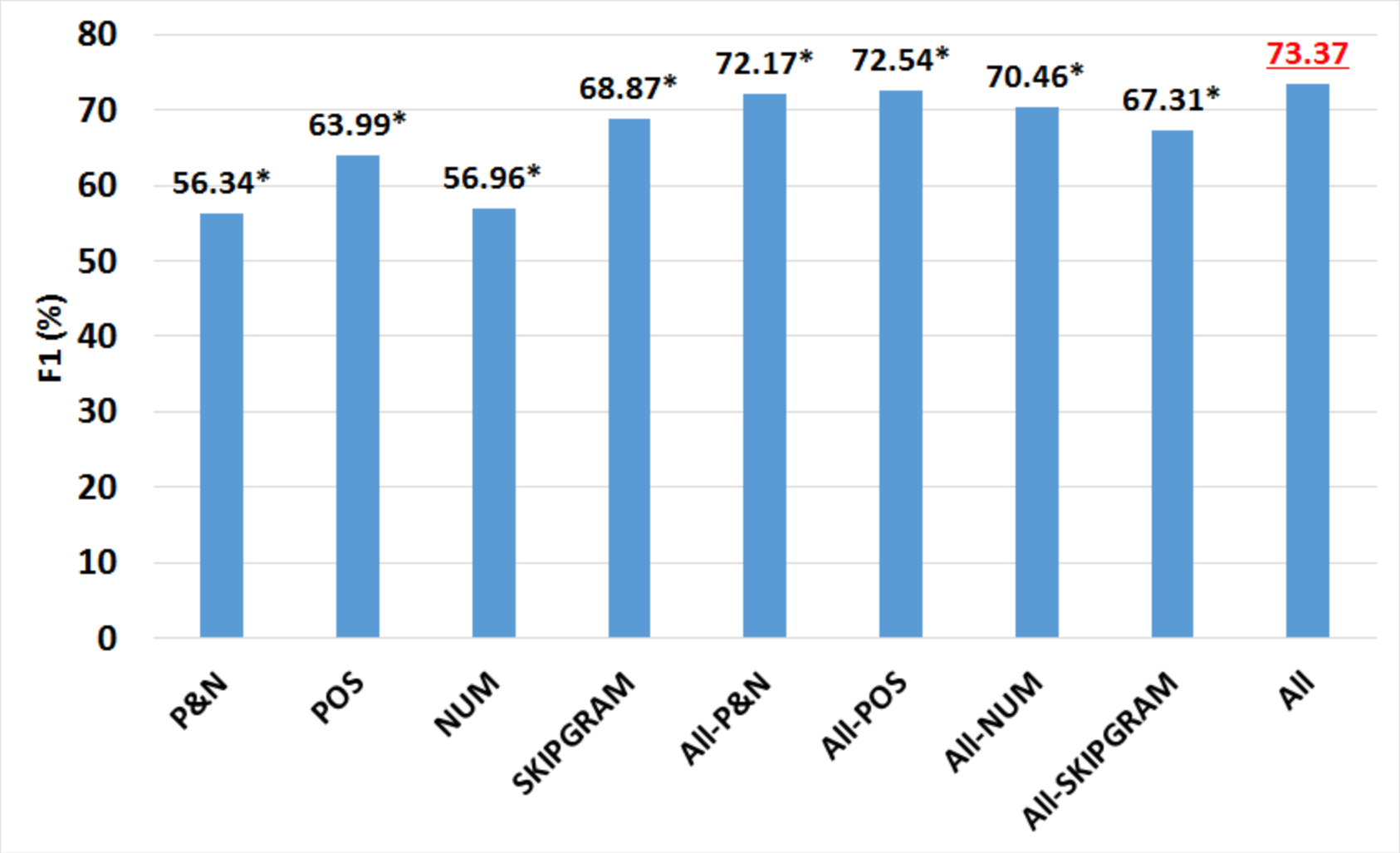
Features for Connectives

- P&N
- POS
- NUM
- SKIPGRAM

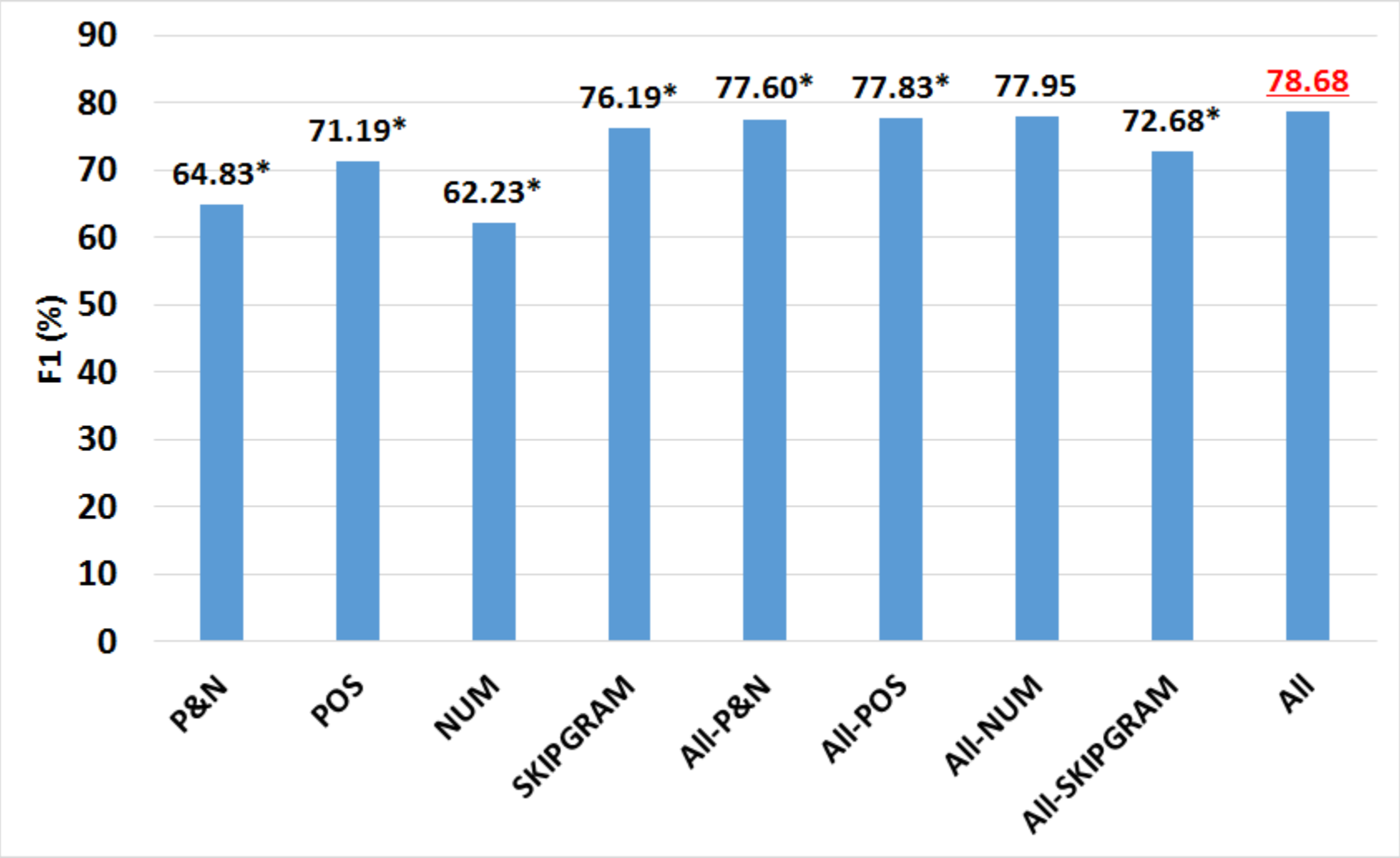
Features for Connectives

- NUM
  - #overlapped candidates
  - #crossed candidates
  - distance between leftmost and rightmost tokens
  - geometric mean of distances between neighboring connective components
  - left distance
  - right distance
  - minimum distance

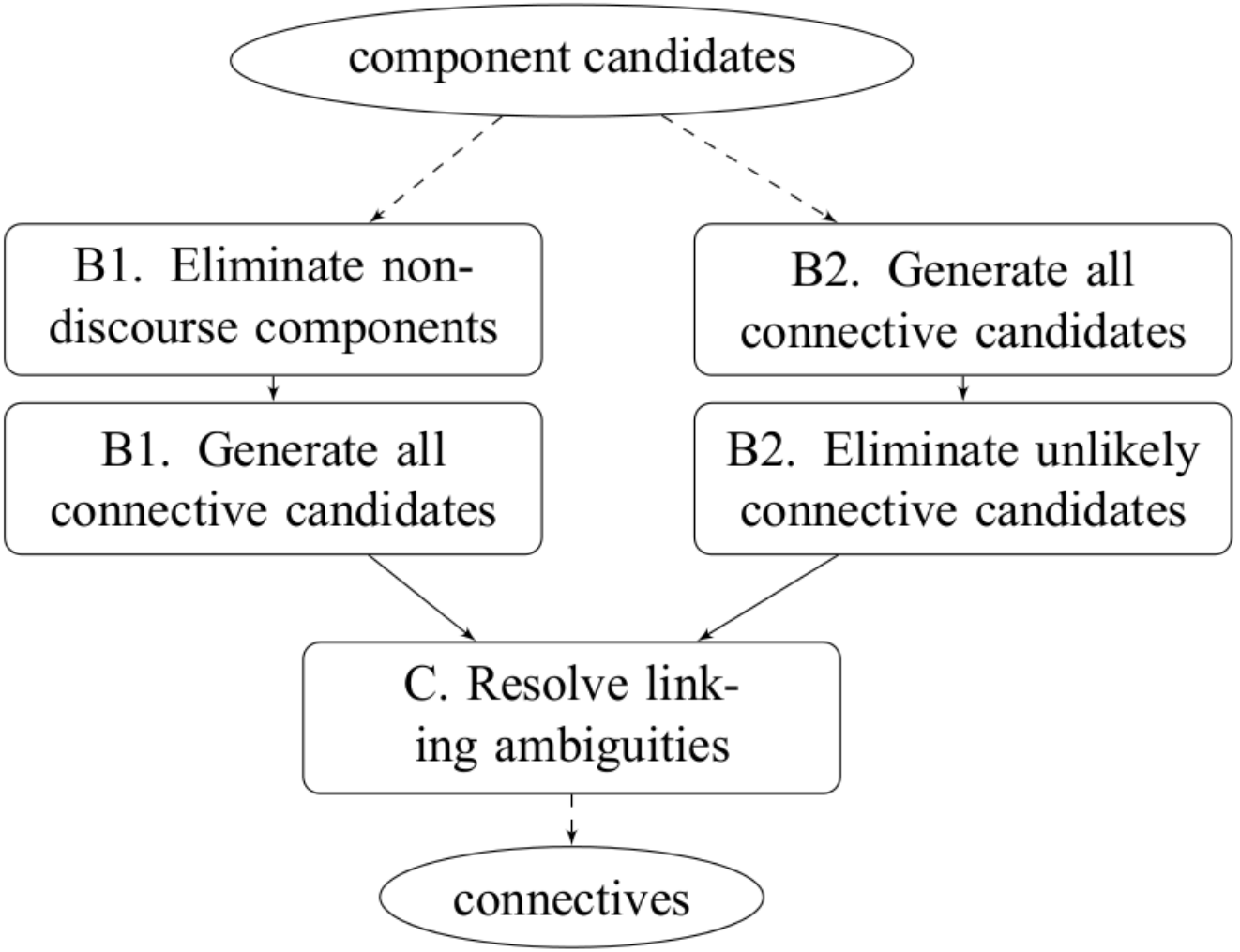
Connective Identification with Different Features (B2)



Component Identification with Different Features (B2)



Linking Disambiguation



Algorithm 1

Algorithm 1 Linking Resolution Algorithm

Input: C: A set of connective candidates
Output: A: A set of accepted connectives

$A \leftarrow \{\}$ ▷ Initialize A
while C is not empty **do**
    Compute linking ambiguities for all connective components exist in C
    if there exists a connective component cc_i that has linking ambiguity $= 1$ **then**
        let c_i be the unique connective candidate the component involves
    else
        Rank all connective candidates in C
        let c_i be the connective candidate that has the highest priority
    end if
    $C \leftarrow C - \{c_i\}$
    $A \leftarrow A \cup \{c_i\}$
    Remove all connective candidates $c_j \in C$ that overlap with c_i
end while
return A

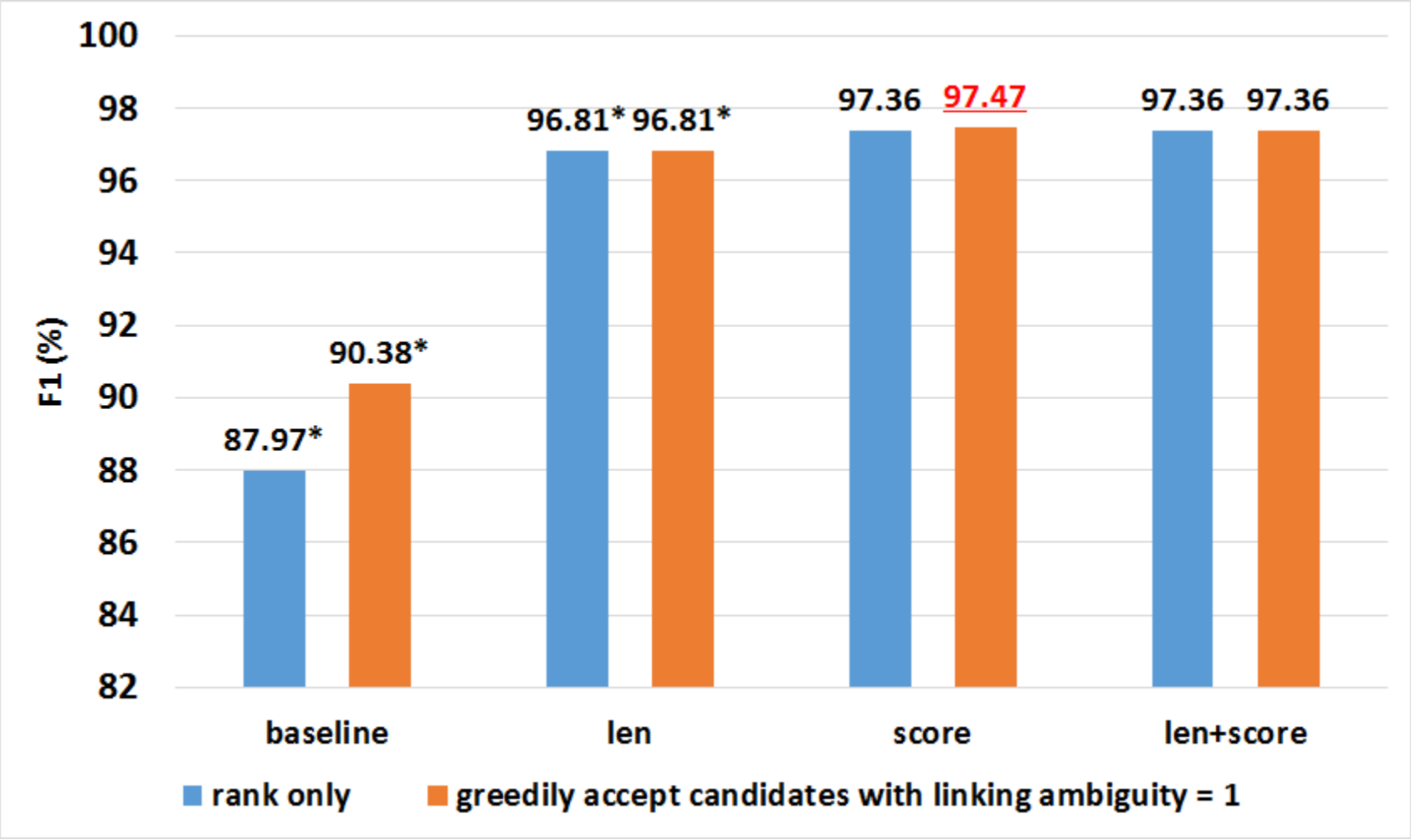
Algorithm 2: Ranking Only

Algorithm 2 Linking Resolution Algorithm by Ranking Only

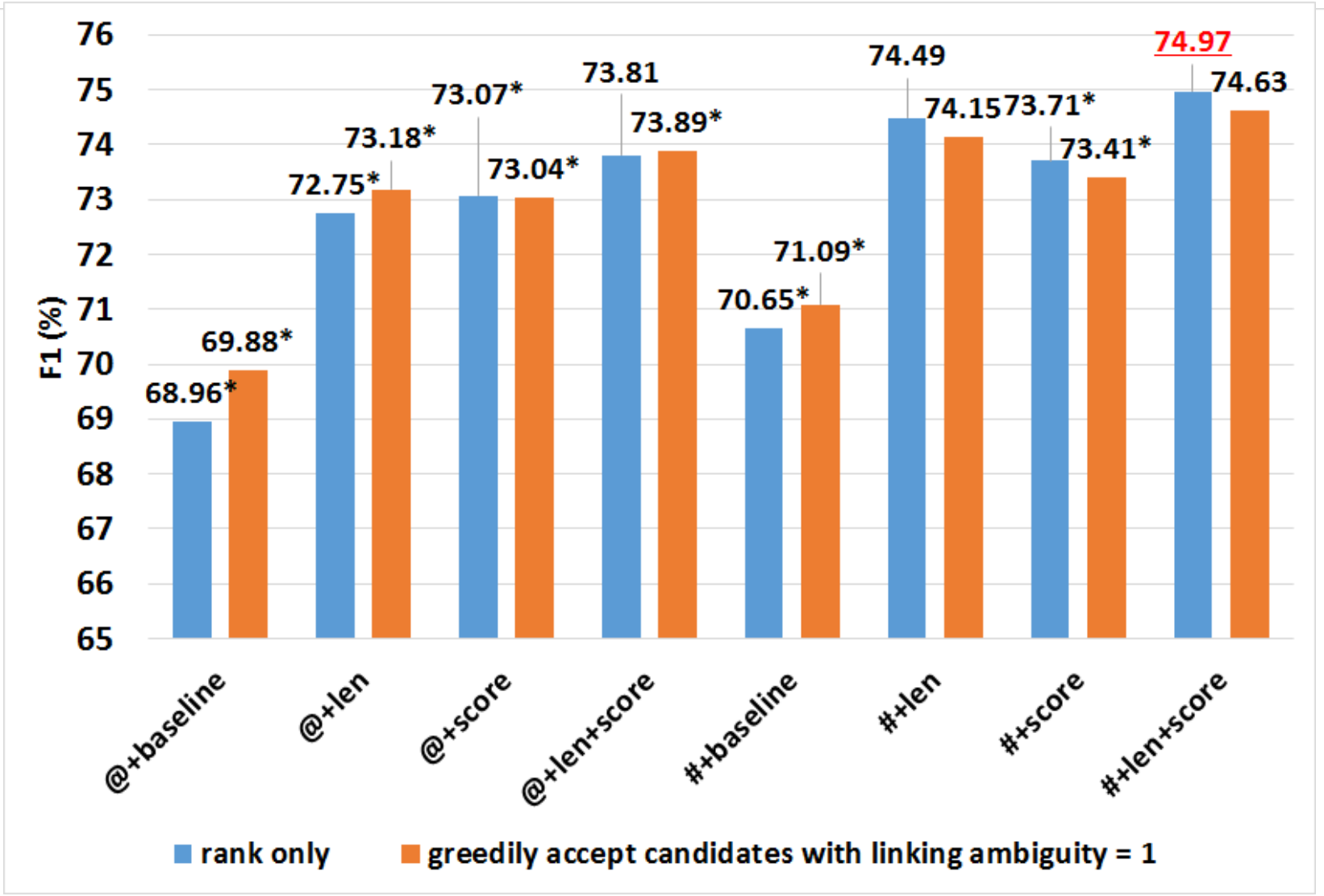
Input: C: A set of connective candidates
Output: A: A set of accepted connectives

1: $A \leftarrow \{\}$ ▷ Initialize A
2: Rank all connective candidates in C
3: **while** C is not empty **do**
4: let c_i be the connective candidate that has the highest priority
5: $C \leftarrow C - \{c_i\}$
6: $A \leftarrow A \cup \{c_i\}$
7: Remove all connective candidates $c_j \in C$ that overlap with c_i
8: **end while**
9: **return** A

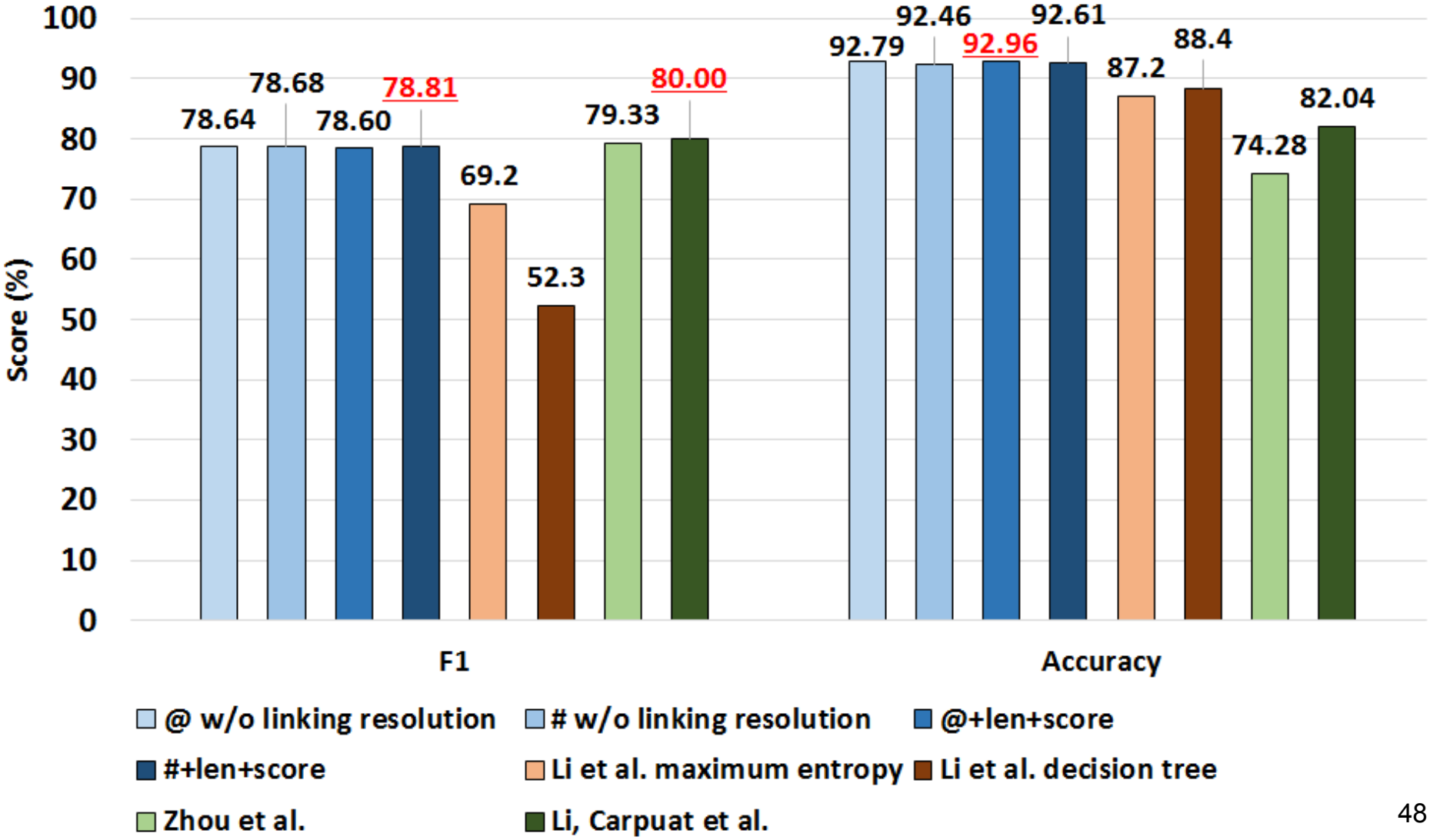
Linking Disambiguation for Known Components



Linking Disambiguation within the Pipeline System



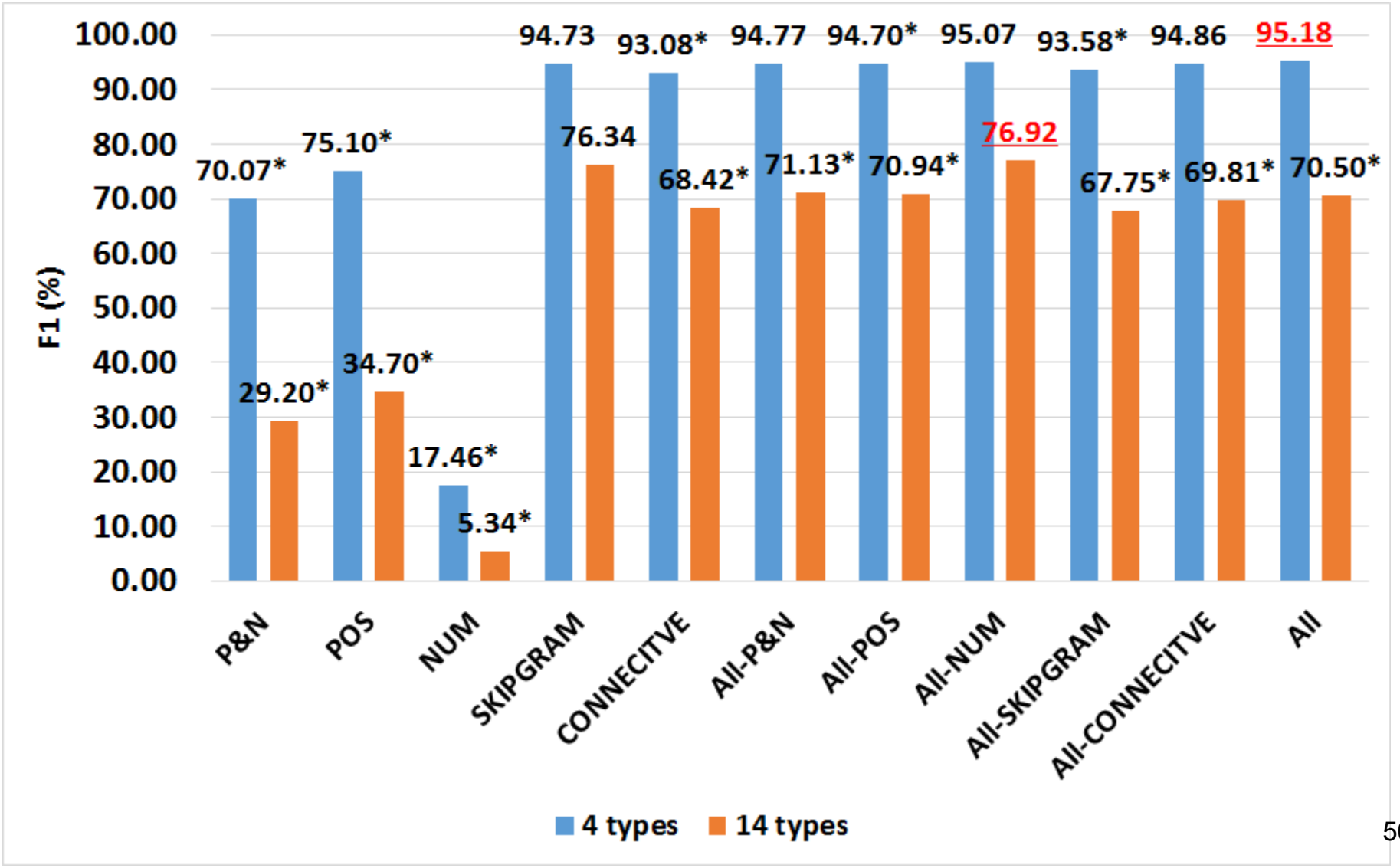
Component Identification Compared with Related Work



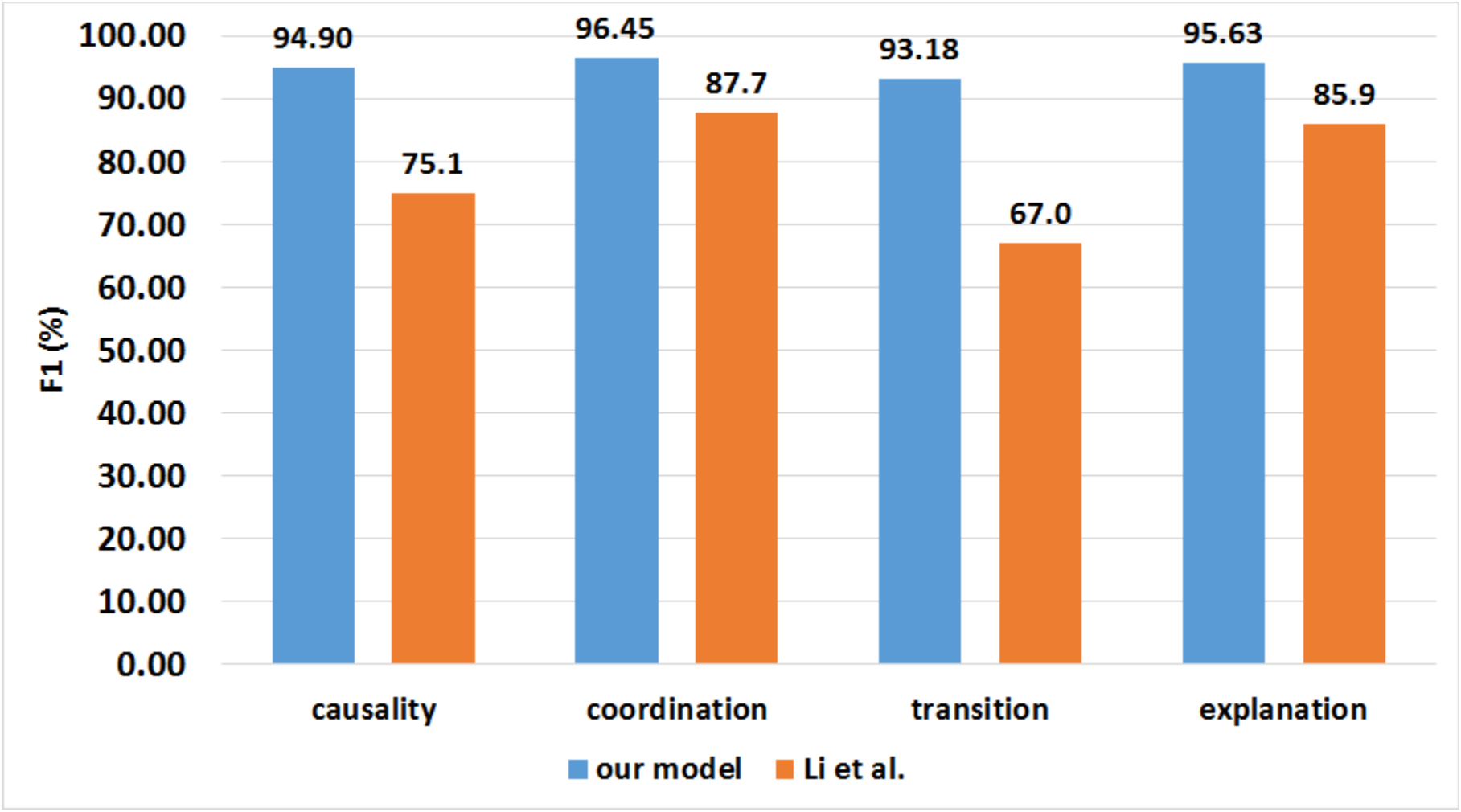
Relation Type Disambiguation

- Same features for connectives
- +connective itself as features
- Eliminate 2nd-level types that have less than 10 instances (inference, background, and evaluation)
  - 1,803 instances left for 2nd-level evaluation

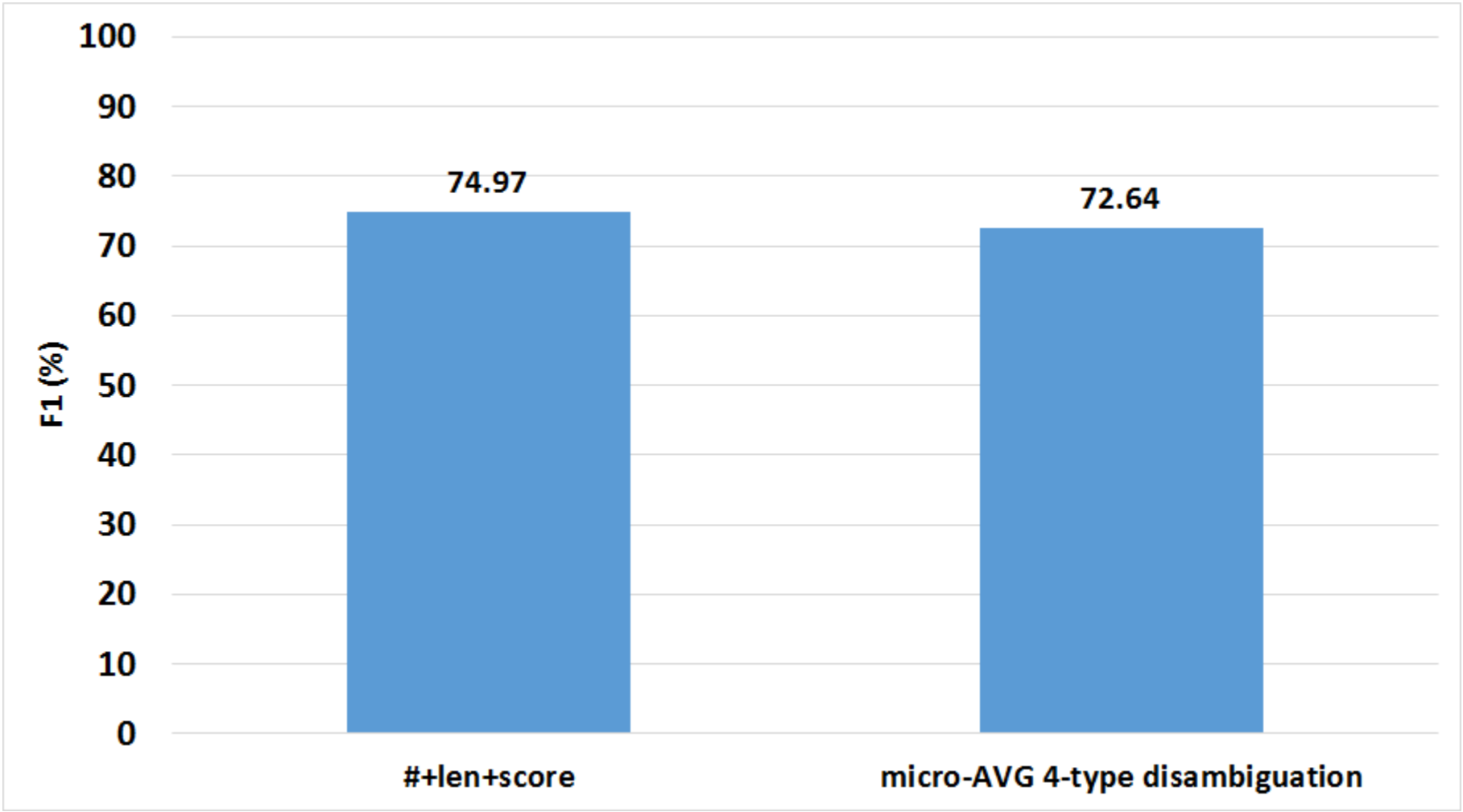
Relation Type Disambiguation for Connectives (macro-AVG)

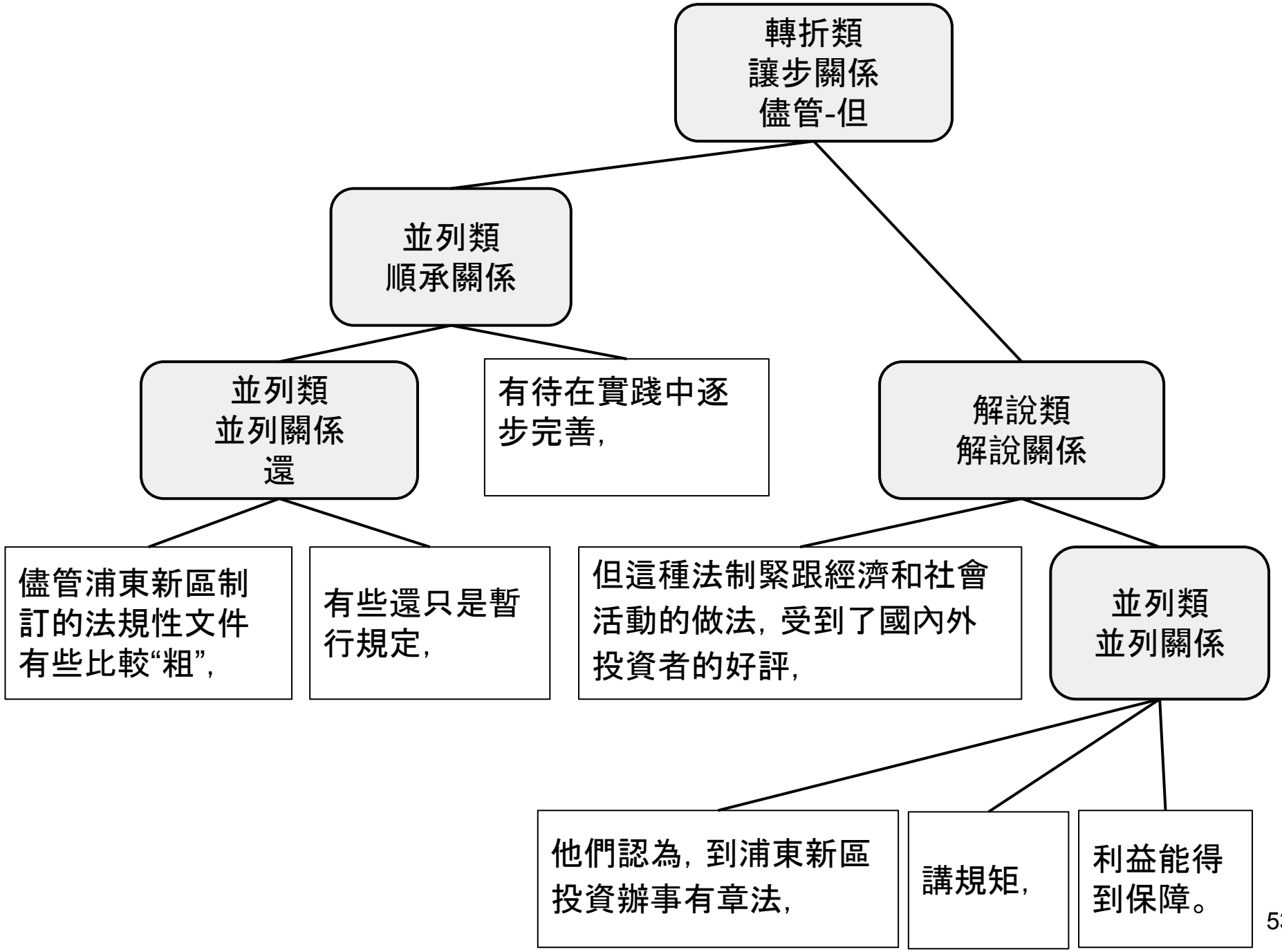


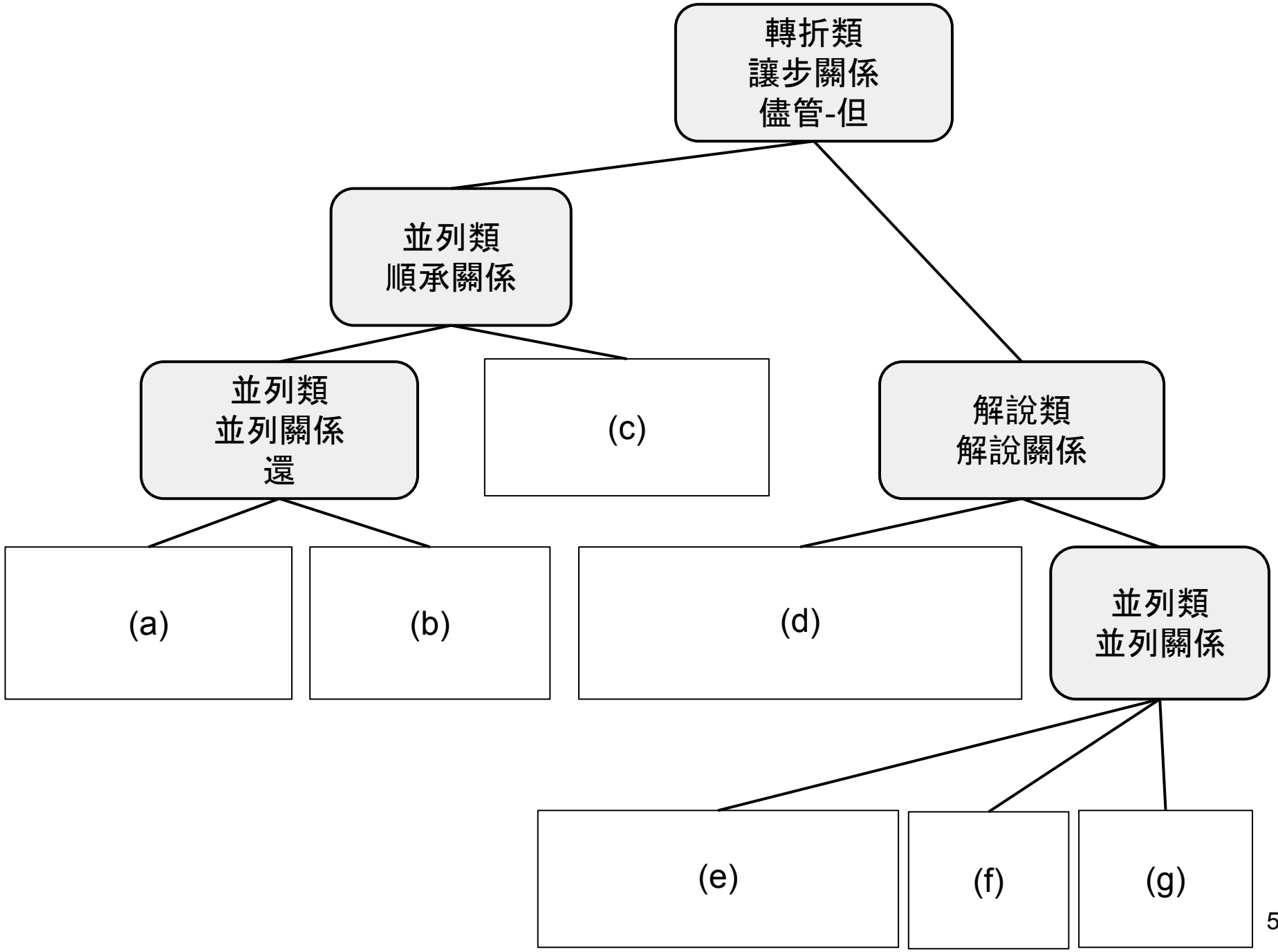
Comparison: Relation Type Disambiguation for Components



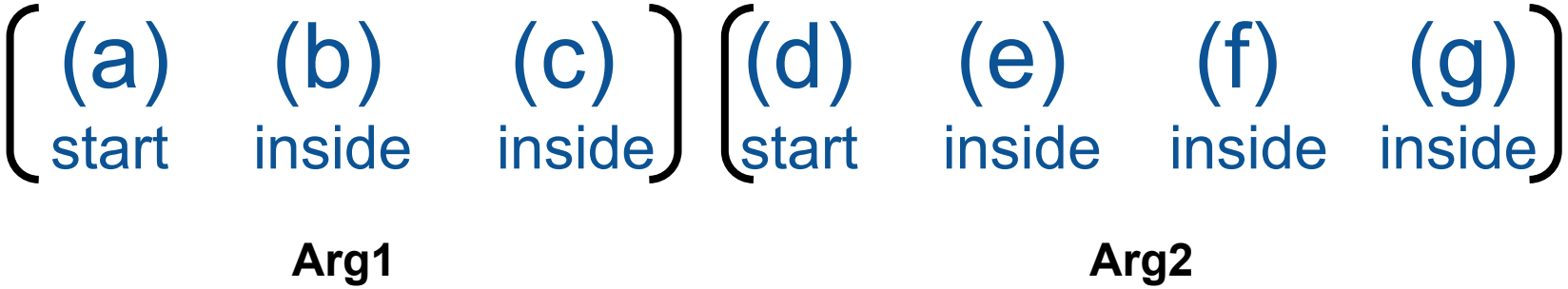
Connective Identification along with Relation Type



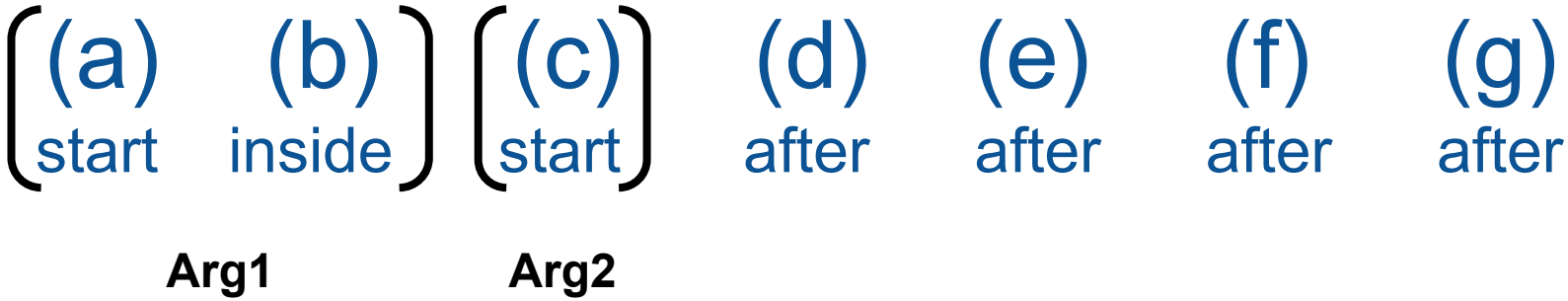

轉折類
讓步關係
儘管-但
並列類
順承關係
並列類
並列關係
還
有待在實踐中逐步完善，
解說類
解說關係
儘管浦東新區制訂的法規性文件有些比較“粗”，
有些還只是暫行規定，
但這種法制緊跟經濟和社會活動的做法，受到了國內外投資者的好評，
並列類
並列關係
他們認為，到浦東新區投資辦事有章法，
講規矩，
利益能得到保障。


轉折類
讓步關係
儘管-但
並列類
順承關係
並列類
並列關係
還
(c)
解說類
解說關係
(a)
(b)
(d)
並列類
並列關係
(e)
(f)
(g)

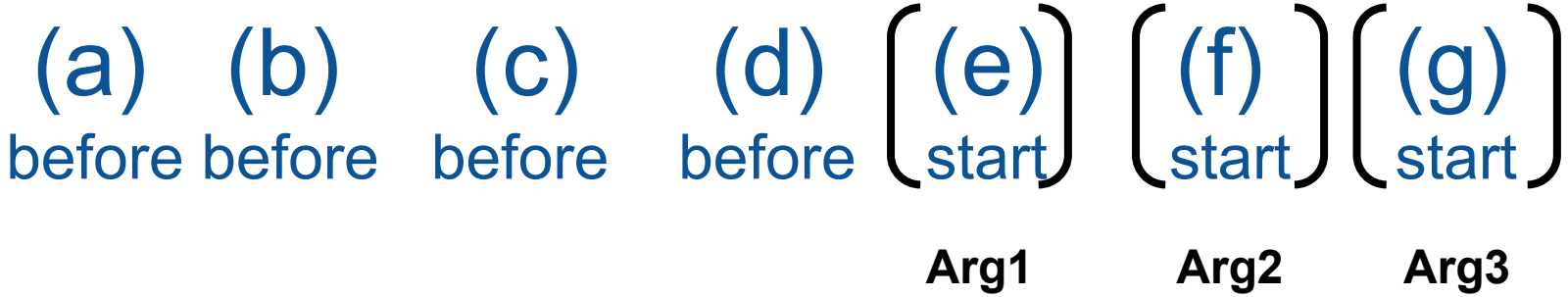
讓步關係



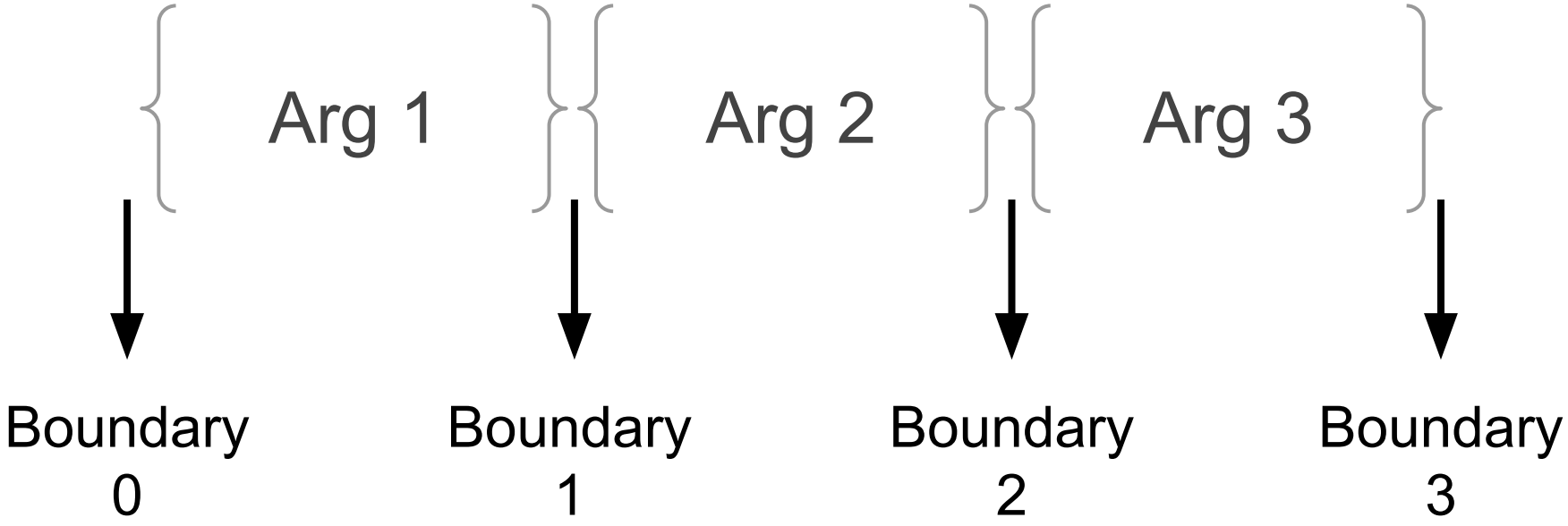
順承關係



並列關係



Argument Boundary Identification



Divide the Text into Segments

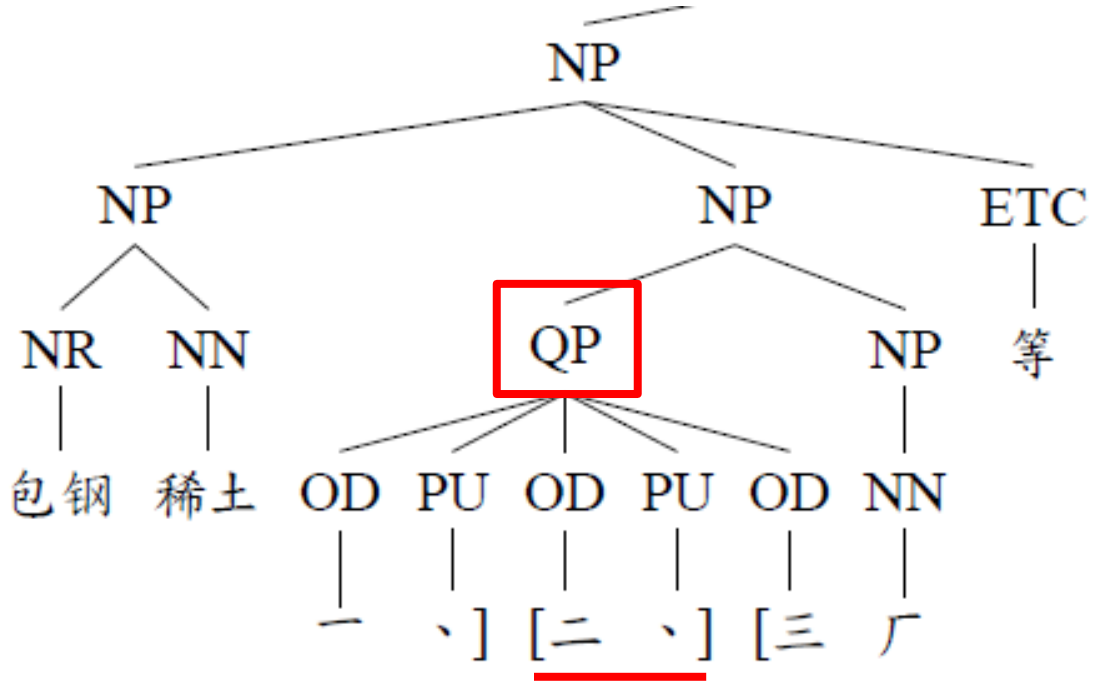
但這種法制緊跟經濟和社會活動的做法，
受到了國內外投資者的好評，
他們認為，
到浦東新區投資辦事有章法，
講規矩，
利益能得到保障。

CRF: Features for a Segment

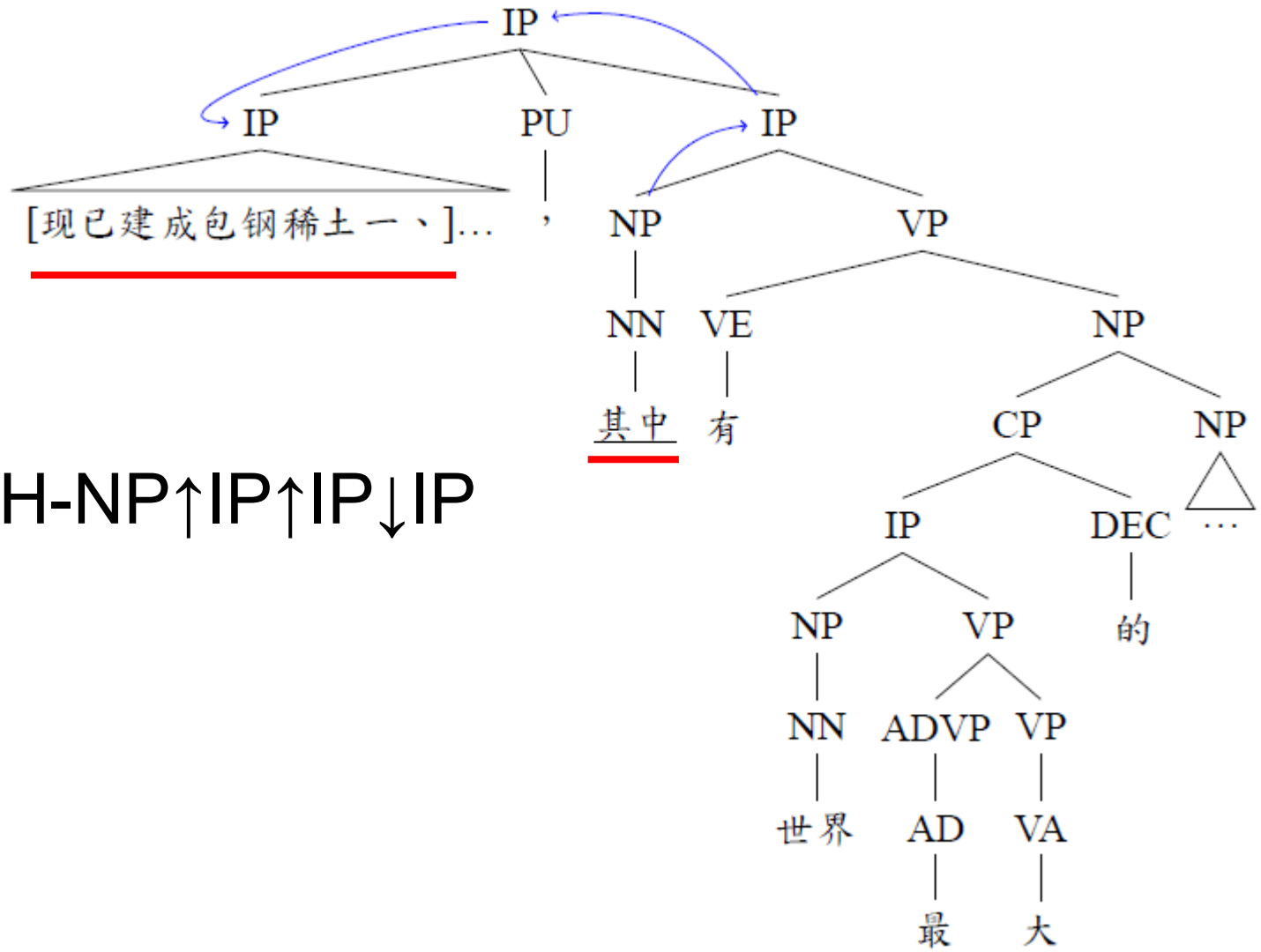
- CONTEXT
- PATH
- POS
- SUBJ
- ENDCHAR
- LINK
- CONNECTIVE
  - connective string, #components
- COMPONENT
  - if a component exists & the component string
  - component positions: begin, end, only
  - if the current segment is before/after all components & the distance to the segment with a component

CONTEXT

- self-category + parent + left-sibling + right-sibling
  - e.g., CONTEXT-QP-NP-Null-NP

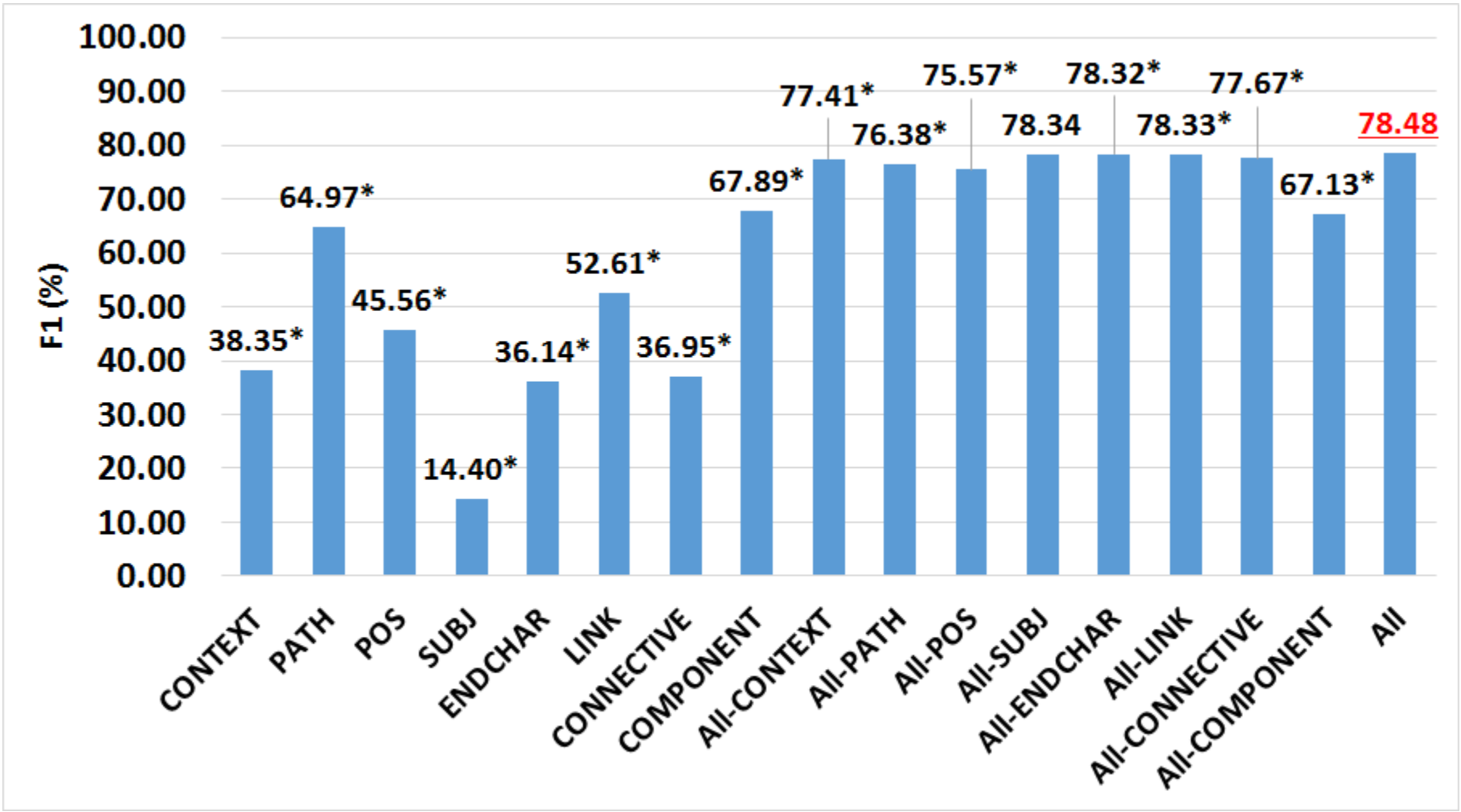


PATH

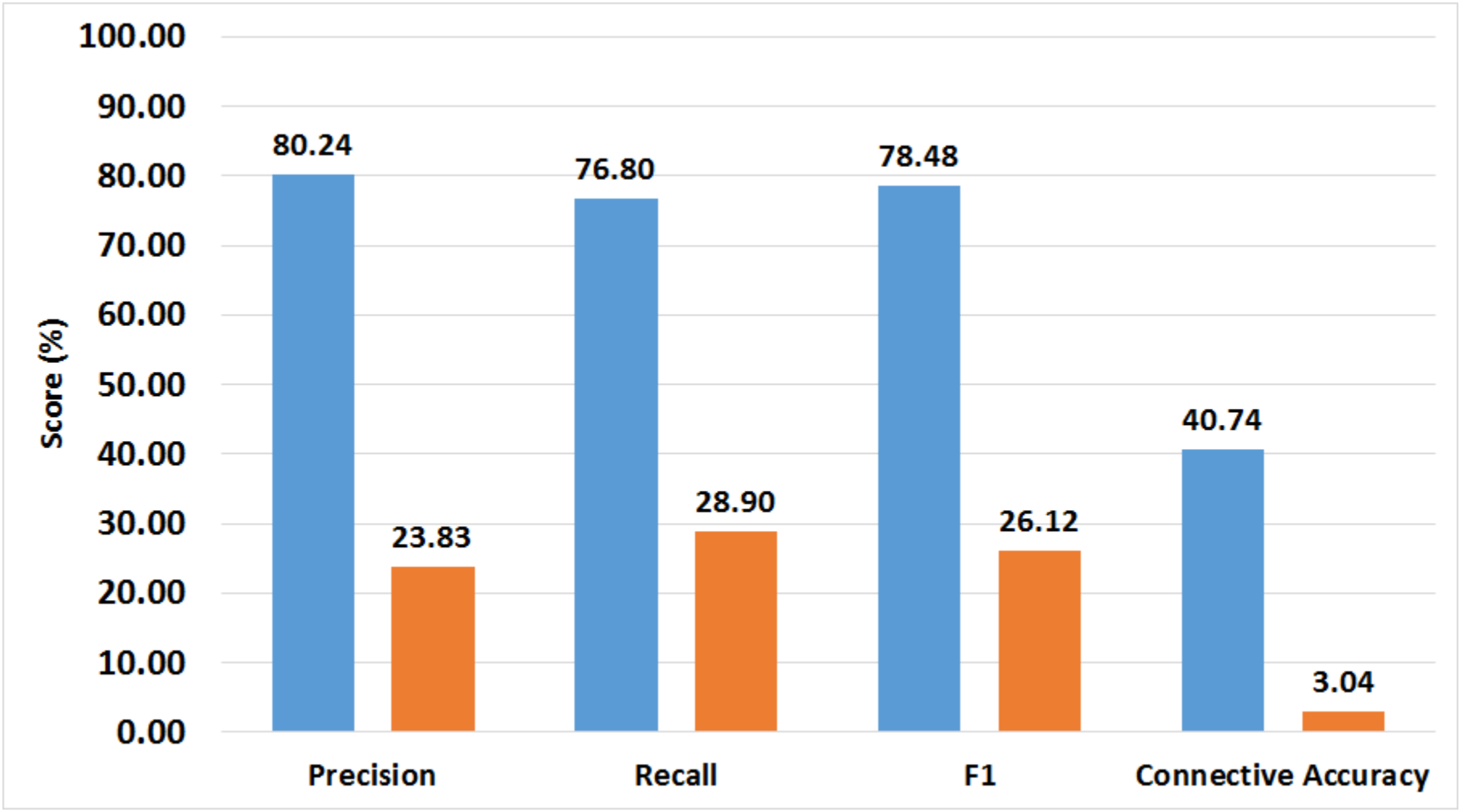


- e.g., PATH-NP↑IP↑IP↓IP

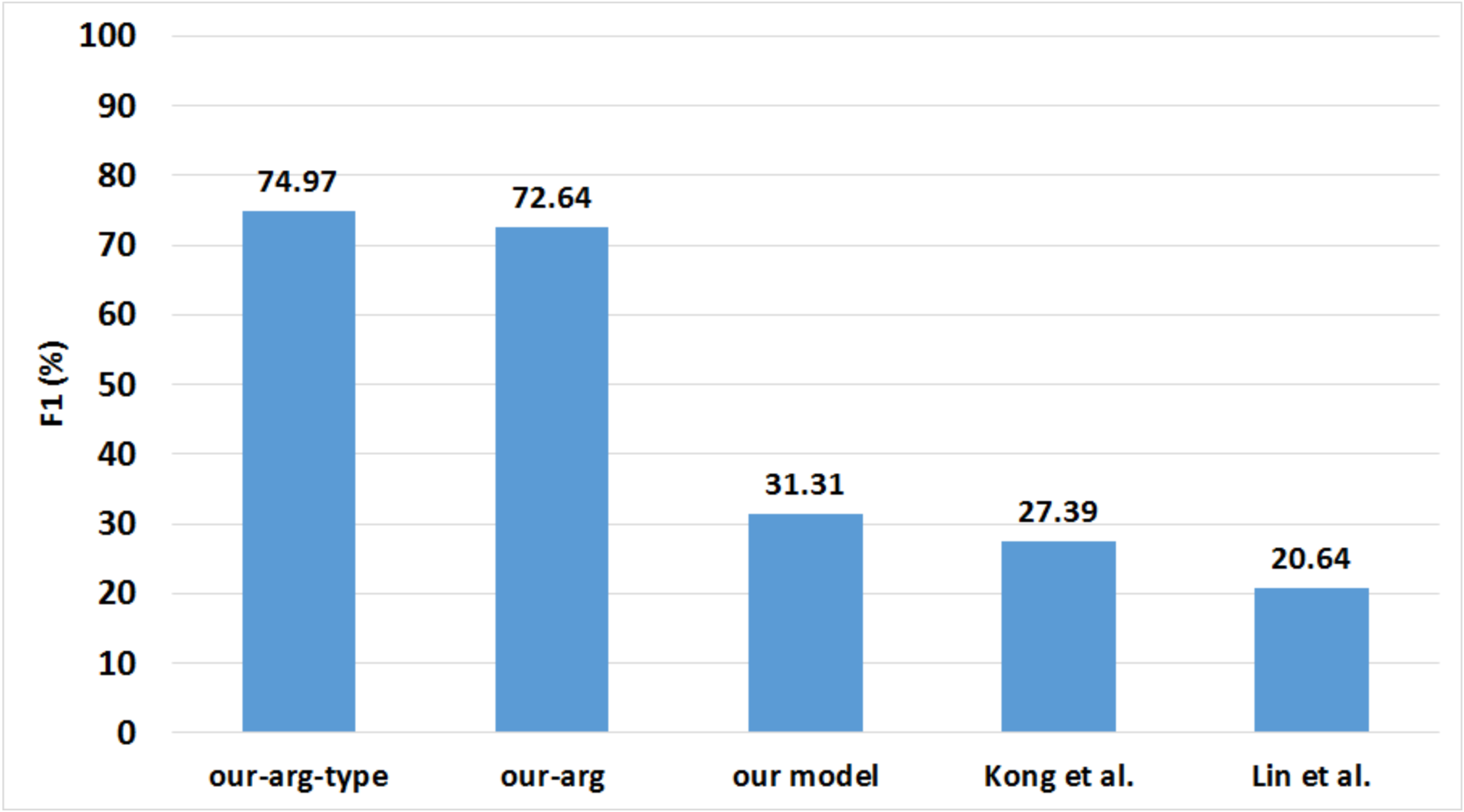
Boundary Identification with Different Features



Boundary Identification Compared with the Baseline



Final Results



1. Introduction
2. Related Work
3. Datasets
4. Methods & Experiments
5. **Conclusion**

Conclusion

- Developed an end-to-end system to identify explicit relations
  - word embeddings as features
  - linking resolution when spurious components exist
  - Chinese connective argument extraction

Future Work

- Integration of Discourse Usage and Linking Disambiguation
- Utilizing Connective Arguments for Disambiguation
- Identifying Implicit Relations for Discourse Parsing

Thanks!

Appendix

他們認為，到浦東新區
投資辦事有章法，
講規矩，
利益能得到保障。

並列關係

他們認為，到浦東新區投資辦事有章法，

講規矩，

利益能得到保障。

但這種法制緊跟經濟和社會活動的做法，受到了國內外投資者的好評，

他們認為，到浦東新區投資辦事有章法，講規矩，利益能得到保障。

解說關係

但這種法制緊跟經濟和社會活動的做法，受到了國內外投資者的好評，

他們認為，到浦東新區投資辦事有章法，講規矩，利益能得到保障。

隱性關係

但這種法制緊跟經濟和社會活動的做法，受到了國內外投資者的好評，

他們認為，到浦東新區投資辦事有章法，講規矩，利益能得到保障。

Goals

- Identify discourse connectives
- Disambiguate linking among connective components
- Disambiguate relation types
- Extract connective arguments

Related Work for English Discourse Analysis

- Discourse corpora
  - RST-DT (Carlson et al., 2001)
  - PDTB (Prasad et al., 2008)
- Discourse connective identification
  - Pitler and Nenkova (2009)
  - Wellner (2009)
  - Faiz and Mercer (2013)
  - Johannsen and Søgaard (2013)
- Relation type disambiguation
  - Pitler and Nenkova (2009)
  - Wellner (2009)

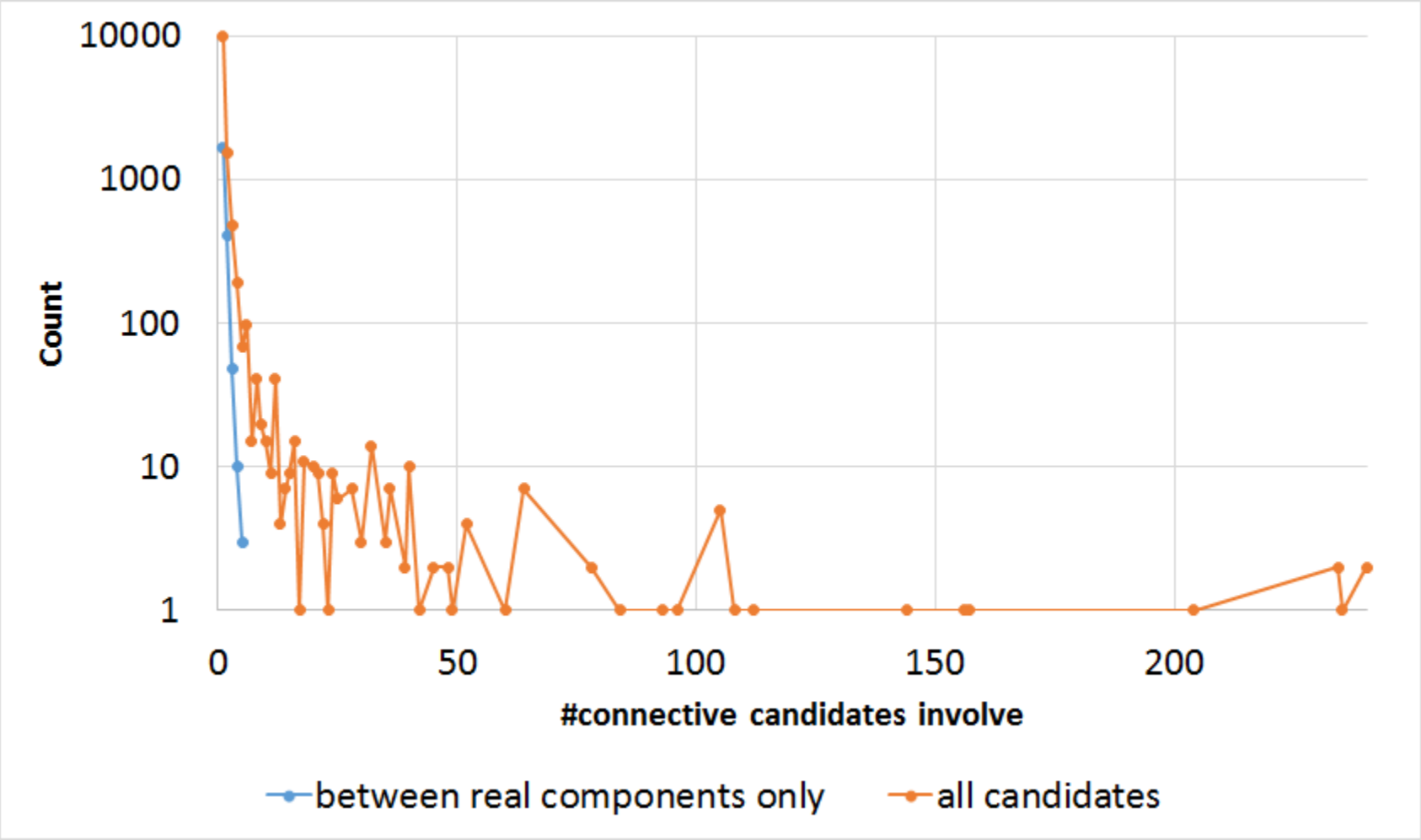
Related Work for English Discourse Analysis

- Connective argument extraction
  - Dines et al. (2005)
  - Wellner and Pustejovsky (2007)
  - Elwell and Baldridge (2008)
  - Wellner (2009)
  - Ghosh et al. (2011, 2012)
  - Kong et al. (2014)
  - Lin et al. (2014)

Related Work for English Discourse Analysis

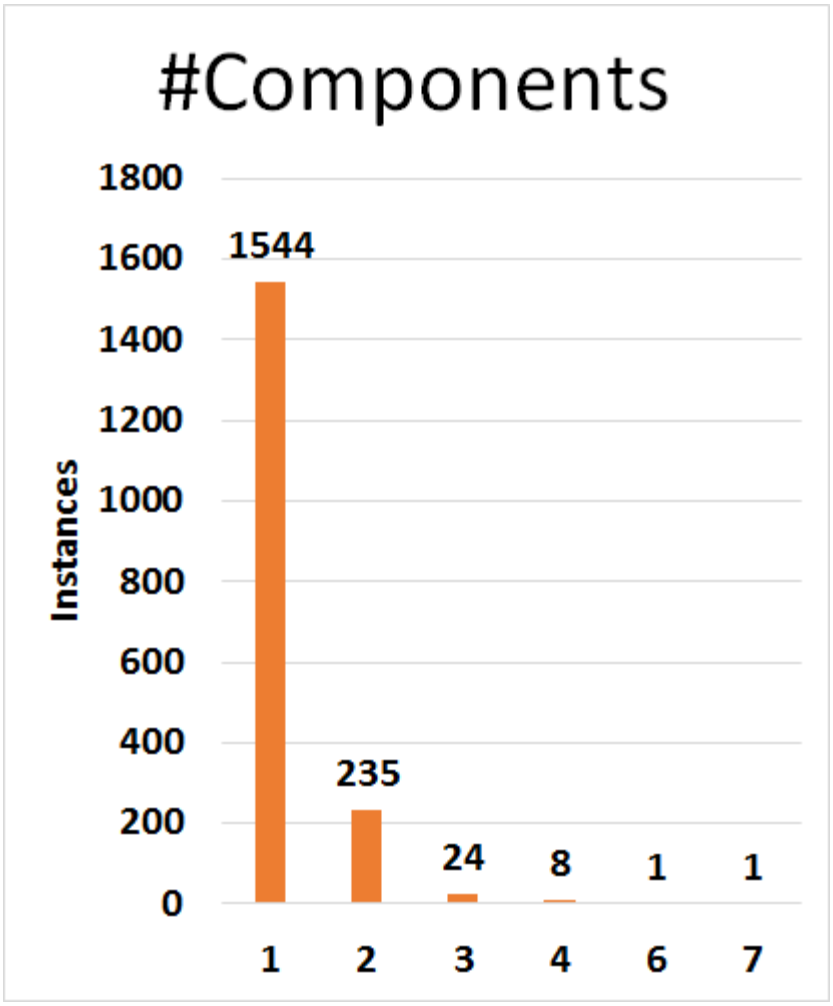
- Sentence level discourse parsing
  - Soricut and Marcu (2003)
  - Sporleder and Lapata (2005)
  - Fisher and Roark (2007)
  - Joty et al. (2012)
- Document level discourse parsing
  - Hernault et al. (2010)
  - Feng and Hirst (2012)
  - Li, Li et al. (2014)

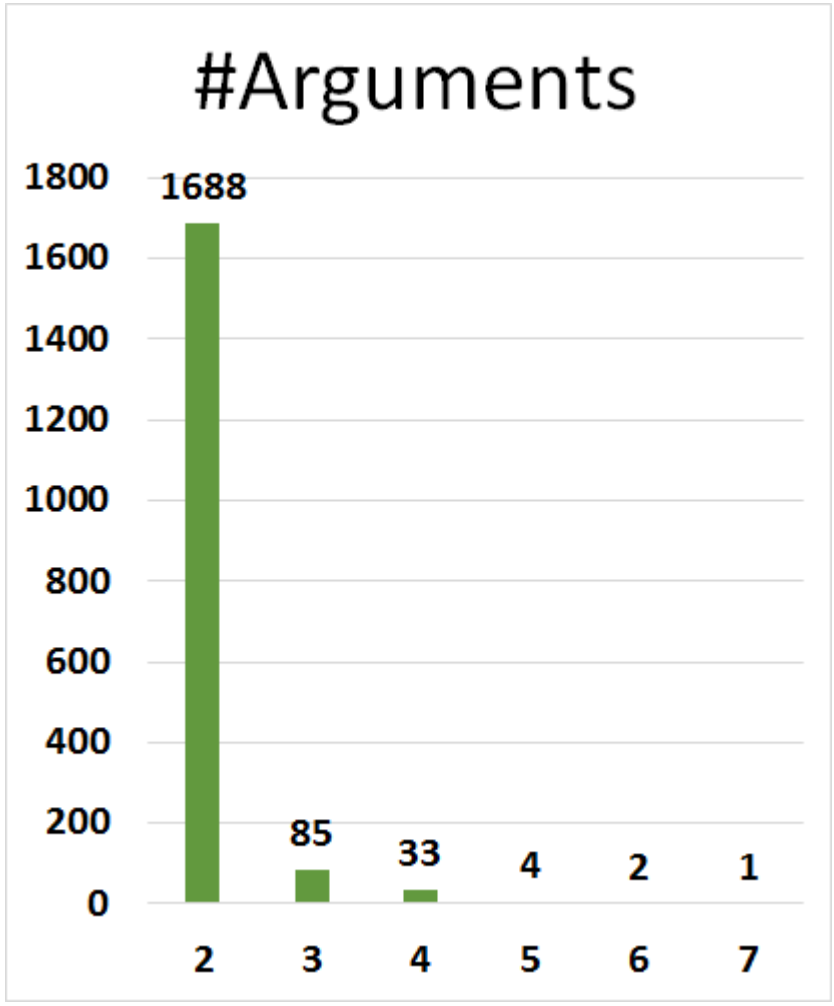
Linking Ambiguity for Components



Relation Type Ambiguity for Connectives

#Top-level Relation Types	1	2	3	4	5
#Connective Classes	258	15	1	-	-
#Instances	1388	355	70	-	-
#2nd-level Relation Types	1	2	3	4	5
#Connective Classes	243	24	5	1	1
#Instances	1030	379	126	208	70

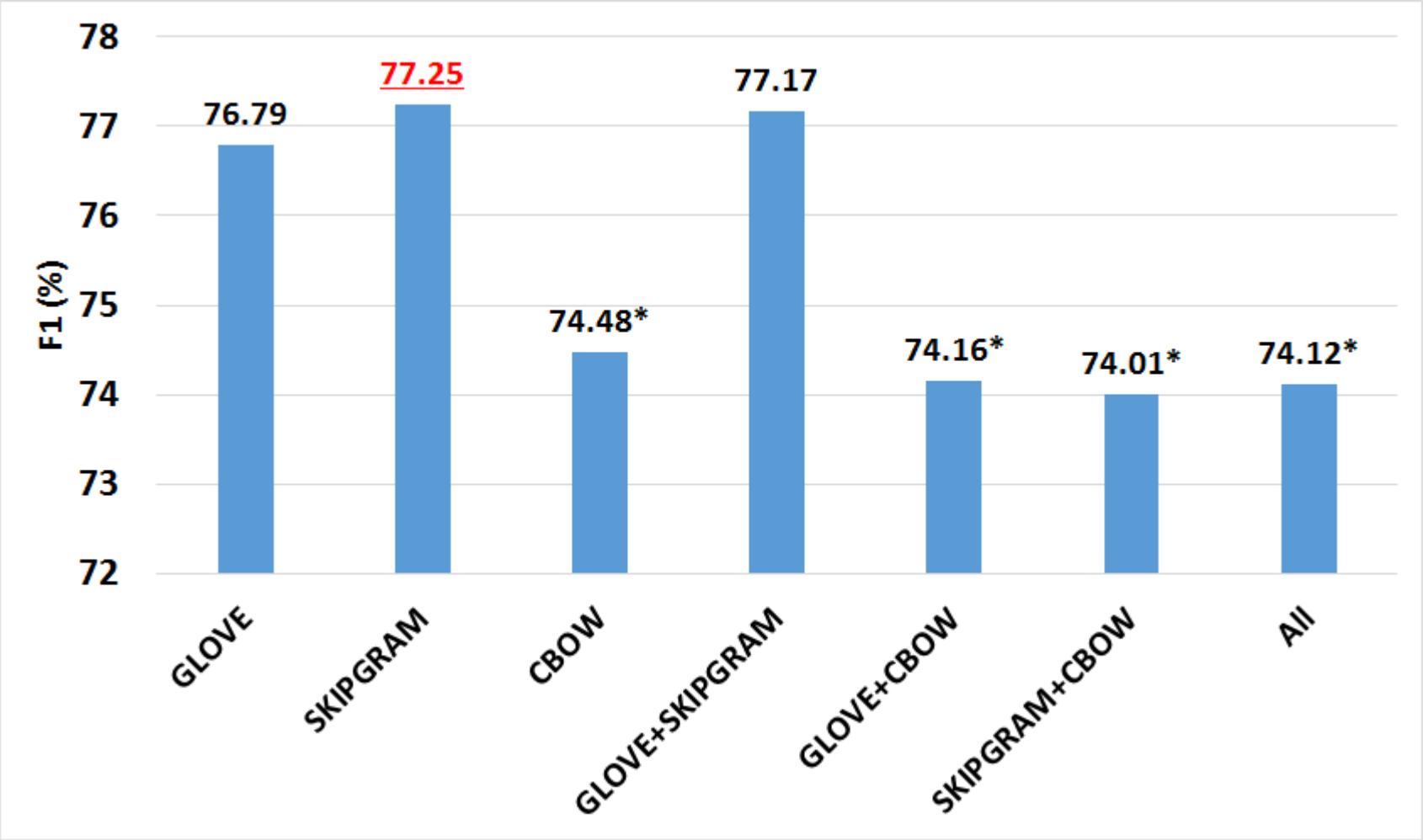

#Components
Instances
1800
1600
1400
1200
1000
800
600
400
200
0
1544
235
24
8
1
1
1
2
3
4
6
7


#Arguments
1800
1600
1400
1200
1000
800
600
400
200
0
1688
85
33
4
2
1
2
3
4
5
6
7

Component Candidate Extraction

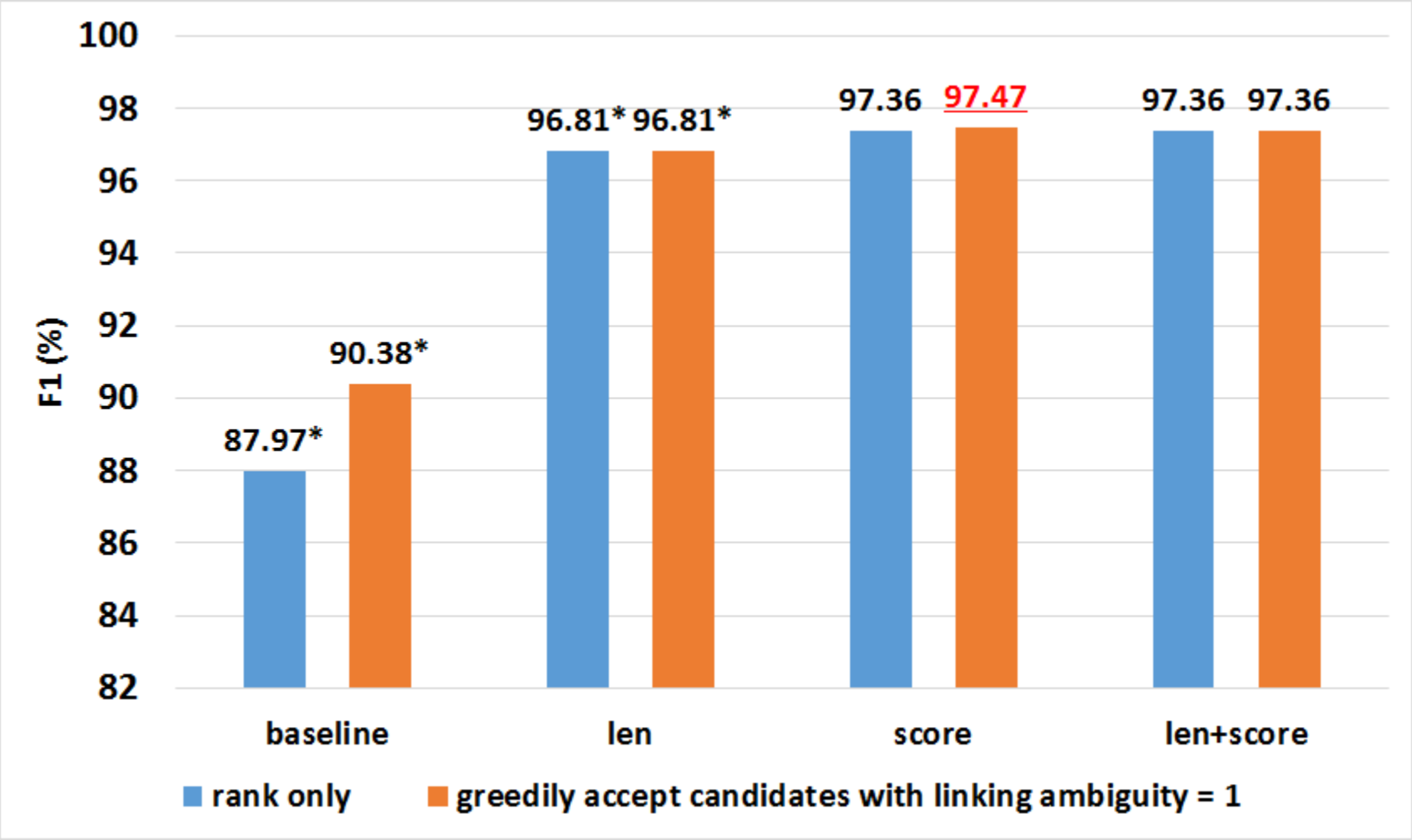
- Use connective component lexicon
  - 24,539 candidates / 2,131 correct components
- Use connective lexicon
  - 12,498 candidates / 2,131 correct components
- Match complete tokens
  - 7,649 component candidates,
    recovering 2,068 of 2,131 correct components
  - 7,976 connective candidates formed,
    recovering 1,755 of 1,813 correct connectives

Component Identification with Different Word Embeddings

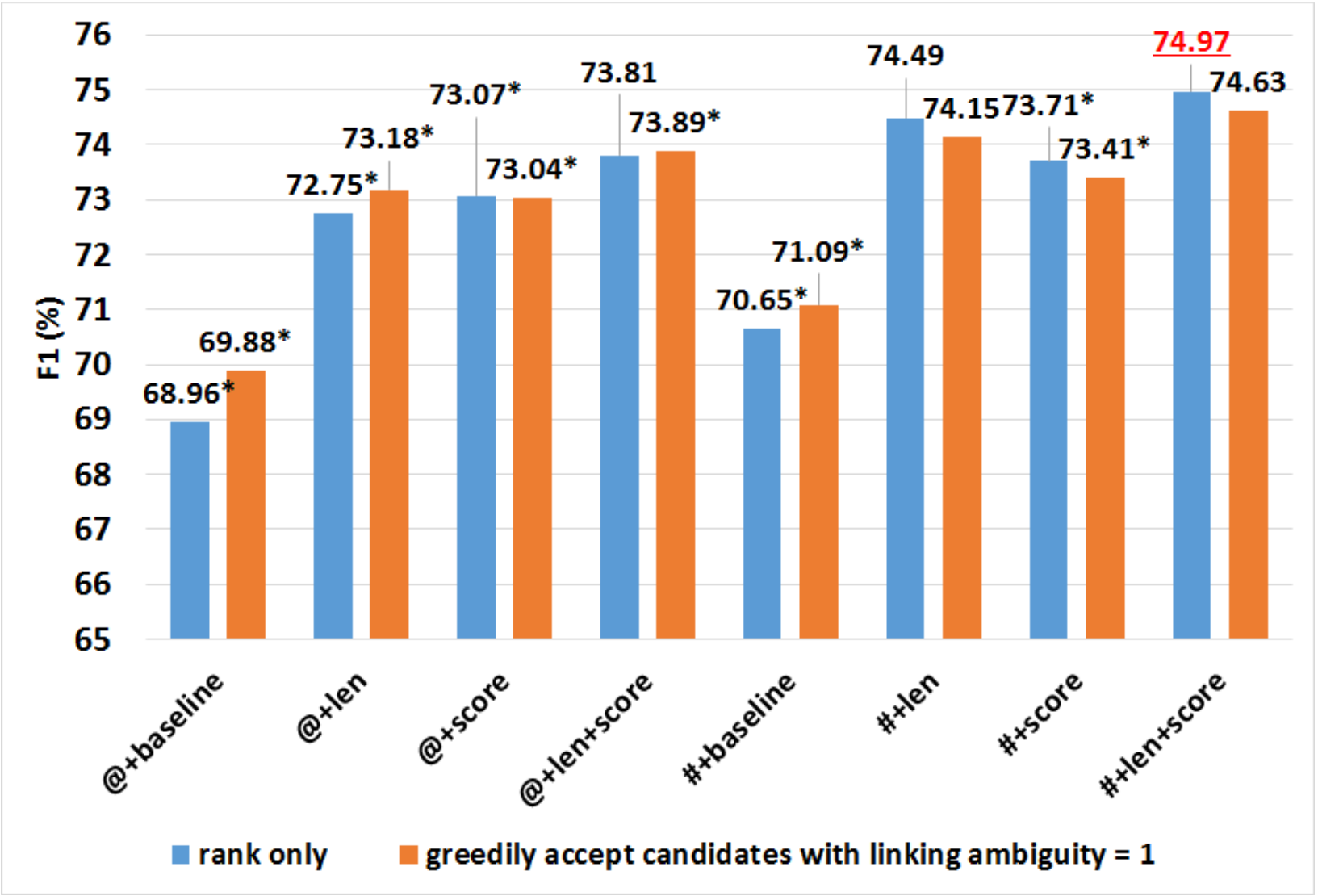


* statistically significant with Wilcoxon signed-rank test at confidence level 0.05

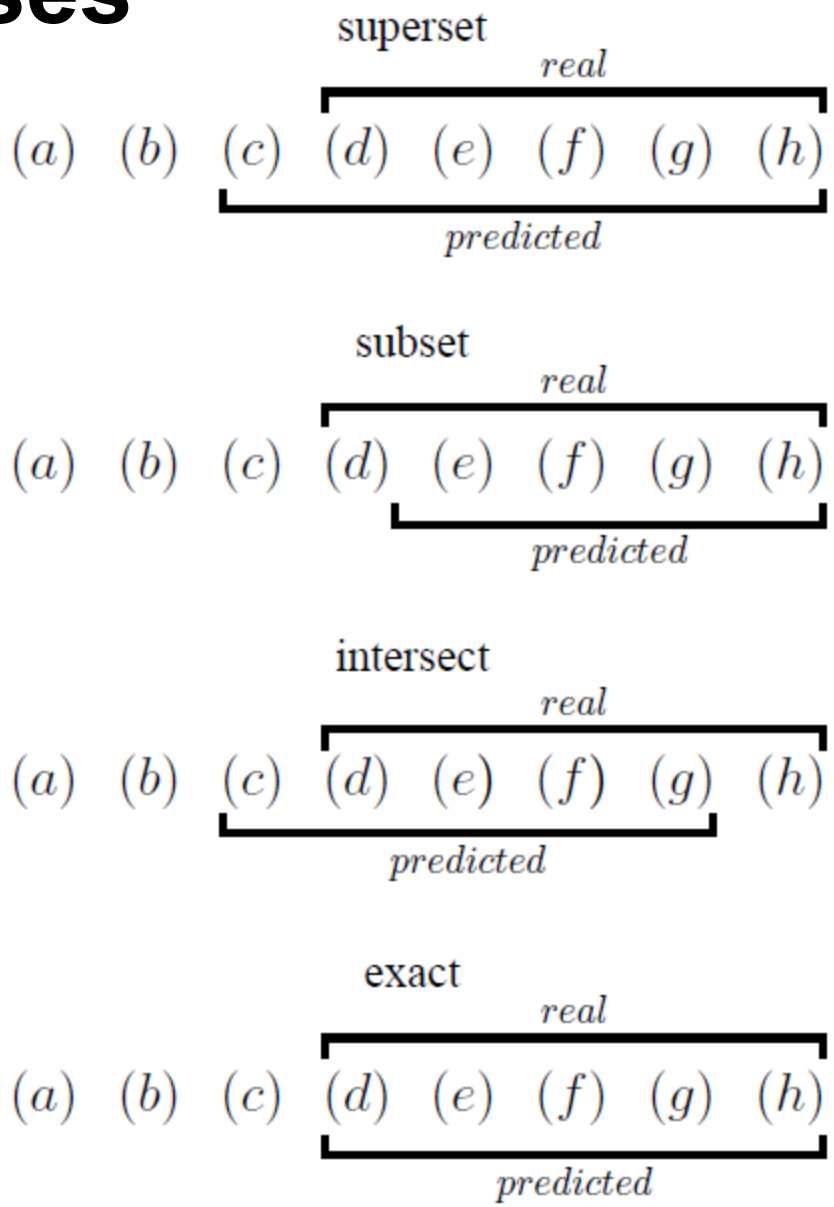
Linking Disambiguation for Known Components



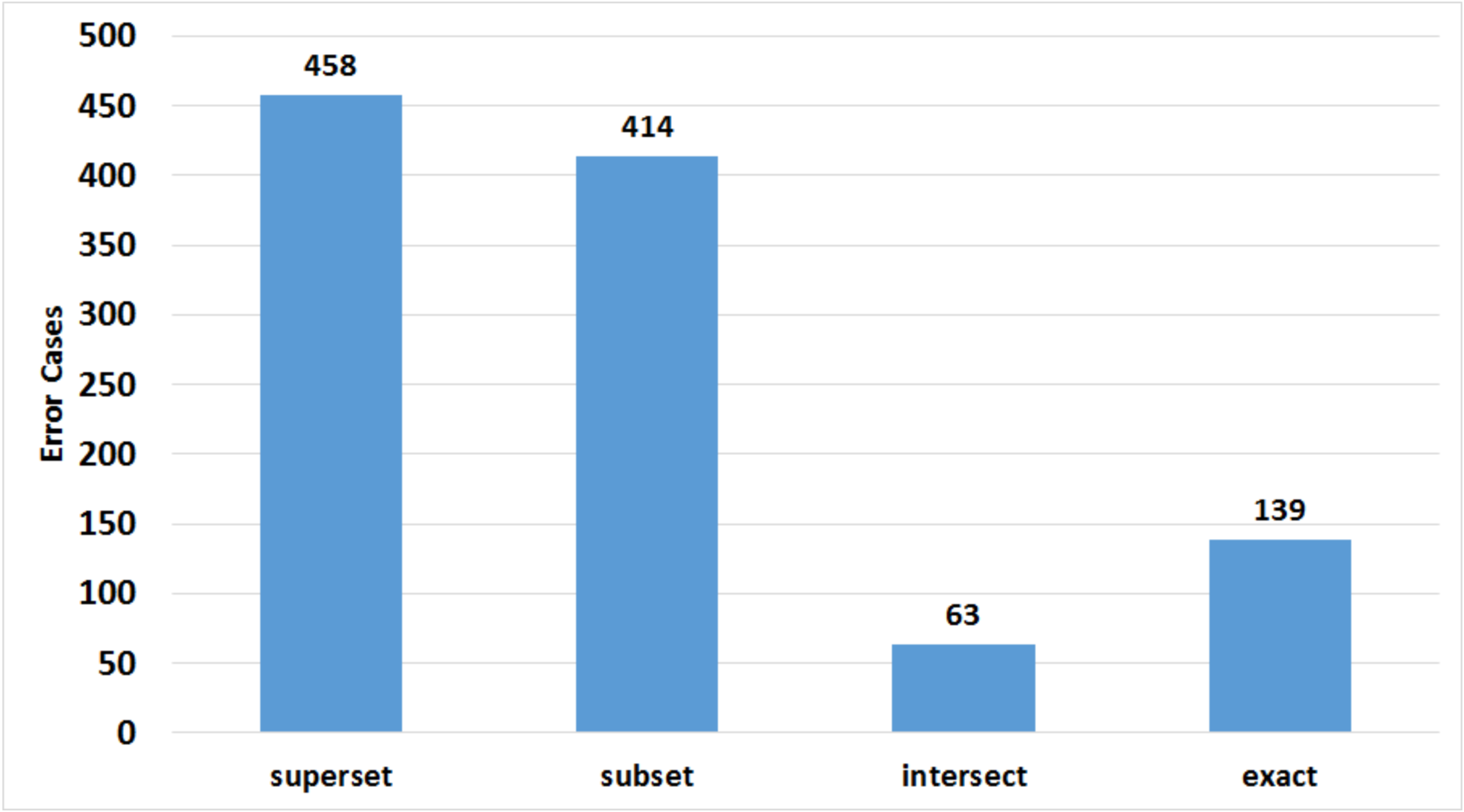
Linking Disambiguation within the Pipeline System



Error Cases


superset
real
(a) (b) (c) (d) (e) (f) (g) (h)
predicted
subset
real
(a) (b) (c) (d) (e) (f) (g) (h)
predicted
intersect
real
(a) (b) (c) (d) (e) (f) (g) (h)
predicted
exact
real
(a) (b) (c) (d) (e) (f) (g) (h)
predicted

Error Analysis



Error Analysis