

篇章分析

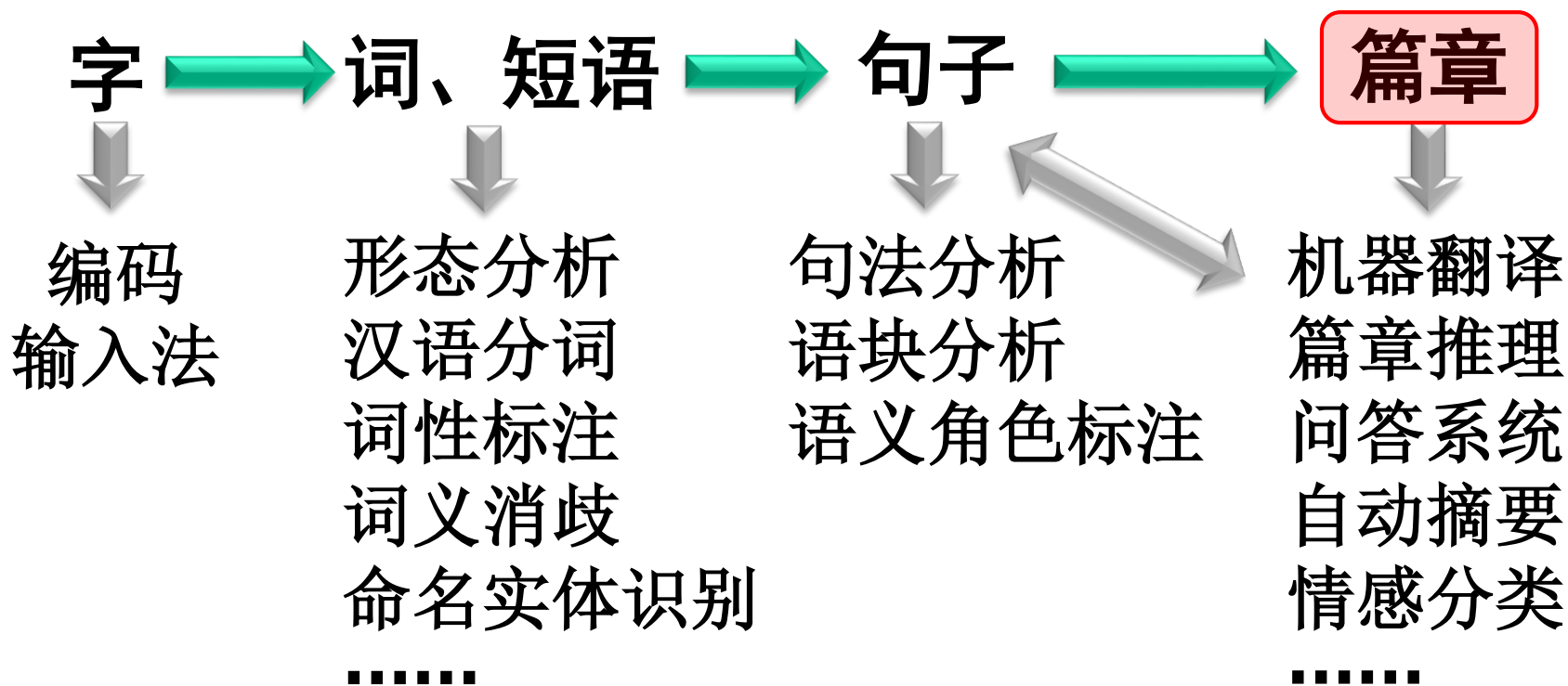
北京市海淀区中关村东路95号
邮编：100190



电话：+86-10-8254 4688
邮件：cqzong@nlpr.ia.ac.cn

1. 引言

◆ NLP处理单位



1. 引言

◆ 词汇链

(S1)数年前，**北海**还是北部湾一个默默无闻的小渔村，然而三五年时间**北海**已建成了**一个现代化都市的框架**，街上客流如潮，楼房拔地而起。

(S2)**北海**已成为中国**对外开放**中升起的一颗明星。

(S3)**北海市**的崛起，是近年来**广西**壮族自治区**对外开放**取得卓著成就的重要标志之一。

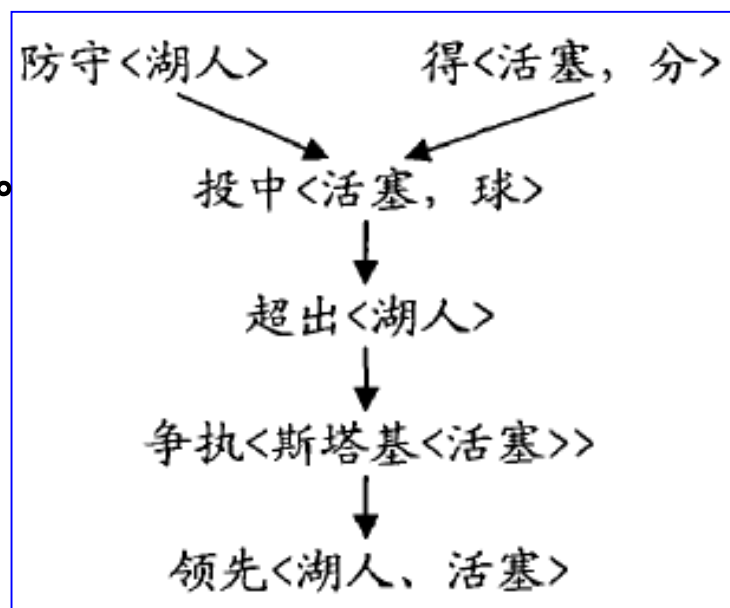
(S4)现在**广西**已初步形成了沿海开放城市、沿海经济开放区、**边境开放城镇**相结合，由沿海、沿边、沿江向腹地推进的多领域、多层次的**对外开放**总体格局。

(S5)统计资料显示，过去五年**广西****对外贸易**和利用**外资**规模迅速扩大，**进出口贸易额**累计达到一百亿美元，其中出口六十八点七亿美元，分别比“七五”时期（一九八六至一九九〇年）增长一点七八倍和一点四三倍；实际利用**外资**累计达到三十三点二四亿美元，……

1. 引言

◆ 事件链

- a. 第三节湖人加强了防守，拼得活塞只得9分。
- b. 本节打了8分钟后，活塞只投中一球。
- c. 湖人以16:3开始本节，一举以57:48超出。
- d. 其间活塞斯塔基和裁判还发生了争执。
- e. 三节过后，湖人以 61:54 领先活塞。



● 构建过程:

- 抽取实体词汇链
- 抽取词汇链上每个词最近的谓词-论元，构成事件链
- 判断相邻事件之间关系

1. 引言

◆话题链 (topic chain)

一组以名词回指(noun anaphora, **NA**)、代词回指(pronoun anaphora, **PA**)和零形回指(zero anaphora, **ZA**)形式的话题连接起来的小句或句子

- **回指(anaphora)**: 是指一个词或短语在语篇中用于(回)指代同一语篇中的另一个词或短语的概念(Quirk et. Al, 1985)。

例如:

中国科学院大学简称国科大, 其前身是中科院研究生院, 拥有浓郁的学术氛围和雄厚的科研实力, 国科大招收的第一届本科生于2018年夏季毕业。



1. 引言

◆话题链 (topic chain)

例：你们年纪还小，(S1)还要成家立业，(S2)不要虚度年华，
(S3)更不要成为社会讨厌的人。

话题链：你们—ZA—ZA—ZA

关联词：“还…更…”、“不要…不要…”

例：我无意中碰到了身边的一个什么东西，(S1)伸手一摸(O1)，
(S2)是他给我开的饭，两个干硬的馒头。

主话题链：我—ZA； 次话题链：Φ—ZA

实体链： 东西—Φ—饭—馒头

1. 引言

◆ RST修辞结构理论

泽巴里说，以美国为首的联军是应伊拉克要求在伊存在并对伊提供保护的，因为目前的伊临时政府不能控制安全局势，无力追剿恐怖分子和破坏分子。

① 泽巴里说，

② 以美国为首的联军是应伊拉克要求在伊存在并对伊提供保护的，

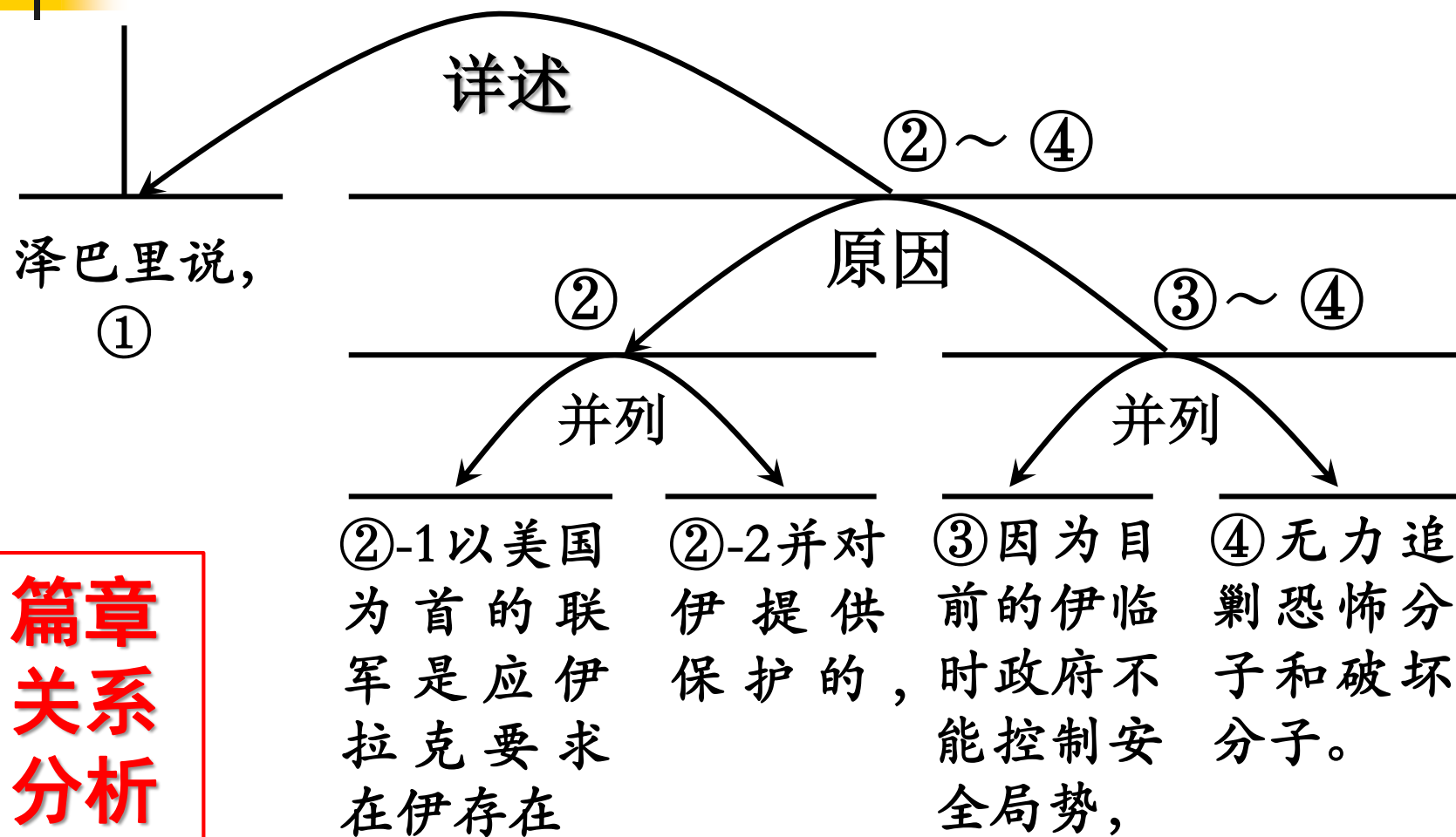
③ 因为目前的伊临时政府不能控制安全局势，

④ 无力追剿恐怖分子和破坏分子。

②-1 | ②-2



1. 引言



篇章
关系
分析



2. 篇章关系分析

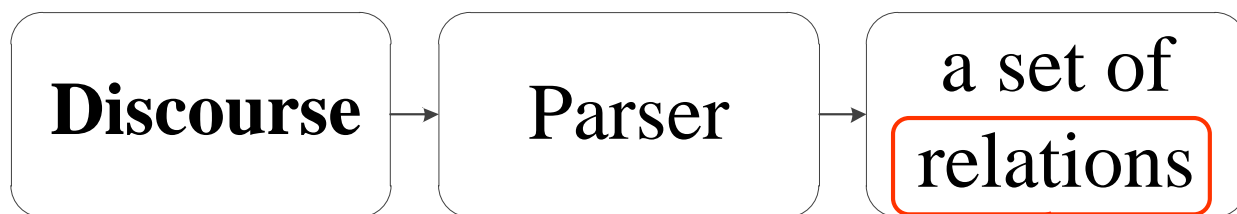
◆ 关于 CoNLL

The SIGNLL(ACL's Special Interest Group on Natural Language Learning) *Conference on Computational Natural Language Learning*

- Shared Task (2016): Multilingual Shallow Discourse Parsing (<http://www.cs.brandeis.edu/~clp/conll16st/>)

2. 篇章关系分析

◆ 浅层篇章关系分析



- Explicit or Non-explicit relations defined by the existence of discourse connectives
- A relation contains two discourse units (Arg1 & Arg2) and the relation sense between them

● Discourse Parser 的三大任务：

关联词识别、Arg抽取、Arg1与Arg2之间的篇章功能类型判断。

2. 篇章关系分析

中国建筑业对外开放呈现新格局

新华社北京二月十三日电

中国建筑市场近年来对外开放步伐进一步加快。

据初步统计，目前在中国境内承包工程的国外承包商已有一百三十七家，承包的工程达一百四十一项，其中最大规模的项目达二十七点七亿元；中外合资合作的建筑企业近二千家。

中国建筑业对外开放始于八十年代。

十几年来，已有美国、日本、法国、英国、德国、芬兰、意大利、新加坡、香港、台湾等十几个国家和地区的境外企业进入中国进行工程总承包或工程分包。

世界上最大的二百二十五家国际承包商中，有十几家已进入中国，其中不少公司与中国公司合资合作进行建设。

根据建设部的规定，凡属于国际金融组织贷款并由国际公开招标的工程全部由外国投资或赠款建设的工程，以及国内企业在技术上难以单独承包的中外合资建设工程，境外建筑企业在取得中国审批的外国企业承包工程资质证后，皆可进入中国境内承包建设项目。

一九九五年九月建设部和外经贸部联合发布的《关于设立外商投资建筑业企业的若干规定》，使中国的建筑市场从允许境外企业到中国承包工程进入到允许境外企业到中国办合资建筑企业。

(完)

1

Type: Implicit

Arg1: 中国建筑市场近年来对外开放步伐进一步加快

Arg2: 目前在中国境内承包工程的国外承包商已有一百三十七家，承包的工程达一百四十一项，其中最大规模的项目达二十七点七亿元；中外合资合作的建筑企业近二千家

Sense: Expansion

2

Type: Implicit

Arg1: 目前在中国境内承包工程的国外承包商已有一百三十七家

Arg2: 承包的工程达一百四十一项

Sense: Conjunction

3

Type: Explicit

Connective: 其中

Arg1: 承包的工程达一百四十一项

Arg2: 最大规模的项目达二十七点七亿元

Sense: Expansion

4

Type: Implicit

Arg1: 目前在中国境内承包工程的国外承包商已有一百三十七家，承包的工程达一百四十一项，其中最大规模的项目达二十七点七亿元

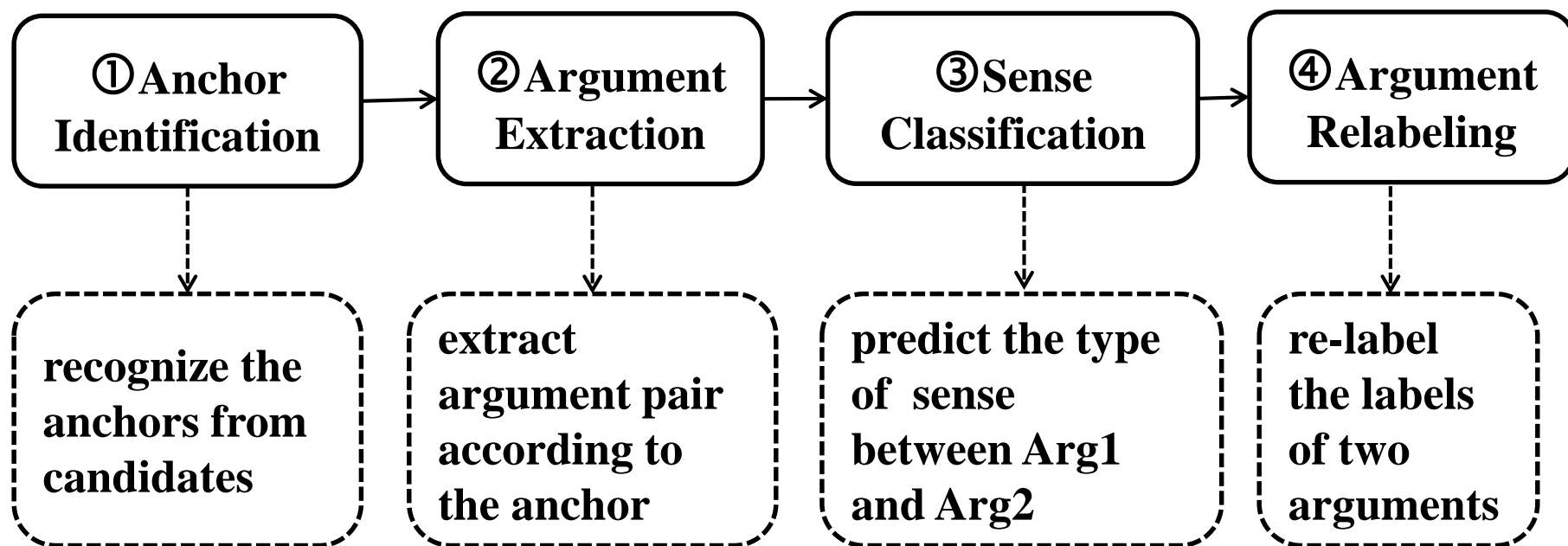
Arg2: 中外合资合作的建筑企业近二千家

Sense: Conjunction

详见第4章课件，
P57-60。

2. 篇章关系分析

◆ 基本分析框架



2. 篇章关系分析

① Anchor Identification

➤ **Explicit: connectives (关联词)**

➤ **Non-explicit: punctuations**

✓ 句中 (middle of sentence, MOS)

✓ 句末 (end of sentence, EOS)

将所有分号、逗号、冒号、句号、破折号、省略号、问号和感叹号作为候选

尽管她的动作潇洒自如，但难度无法与罗莉相比，只获得 9.875 分，夺得银牌。

中国建筑市场近年来对外开放步伐进一步加快。据初步统计，目前在中国境内……

2. 篇章关系分析

➤ 用于猫点词识别的特征

Features		Explicit	Non-explicit	
			MOS	EOS
词法特征 Lexical	candidate itself	✓	✓	✓
	number of the candidate words	✓		
	POS of the candidate	✓		
	POS of the previous word	✓		
	POS of the first and last word in the context clauses		✓	
	POS of the first/last three words in context sentences			✓
	embeddings of the next three words	✓		
	embeddings of first and last word in the context clauses		✓	
	embeddings of the first and last three words in the context sentences			✓
	the previous and next word	✓		
	location in the sentence (start, middle, end)	✓		
	the previous and next punctuation	✓		
	whether the previous or next character is punctuation	✓		

2. 篇章关系分析

Features		Explicit	Non-explicit	
			MOS	EOS
句法特征 Syntactic	the parent of candidate's node (the lowest node in the syntax tree that completely covers the candidate)	✓	✓	
	the left and right siblings of candidate's node	✓	✓	
	the left and right clause's node		✓	
	the production rules of candidate	✓		
	the path from the candidate's node to root	✓		
	the path from candidate's node to right clause's node		✓	
	whether the leftmost sibling of candidate's parent is PP		✓	
	whether the left/right sub-tree contains VP or IP	✓	✓	
	The number of IP in siblings of the candidate's parent		✓	
	whether the right sub-tree contains AD or CS if the leftmost sibling is IP		✓	

➤ 分类器：ME-based/ SVM/ CRFs Classifiers



2. 篇章关系分析

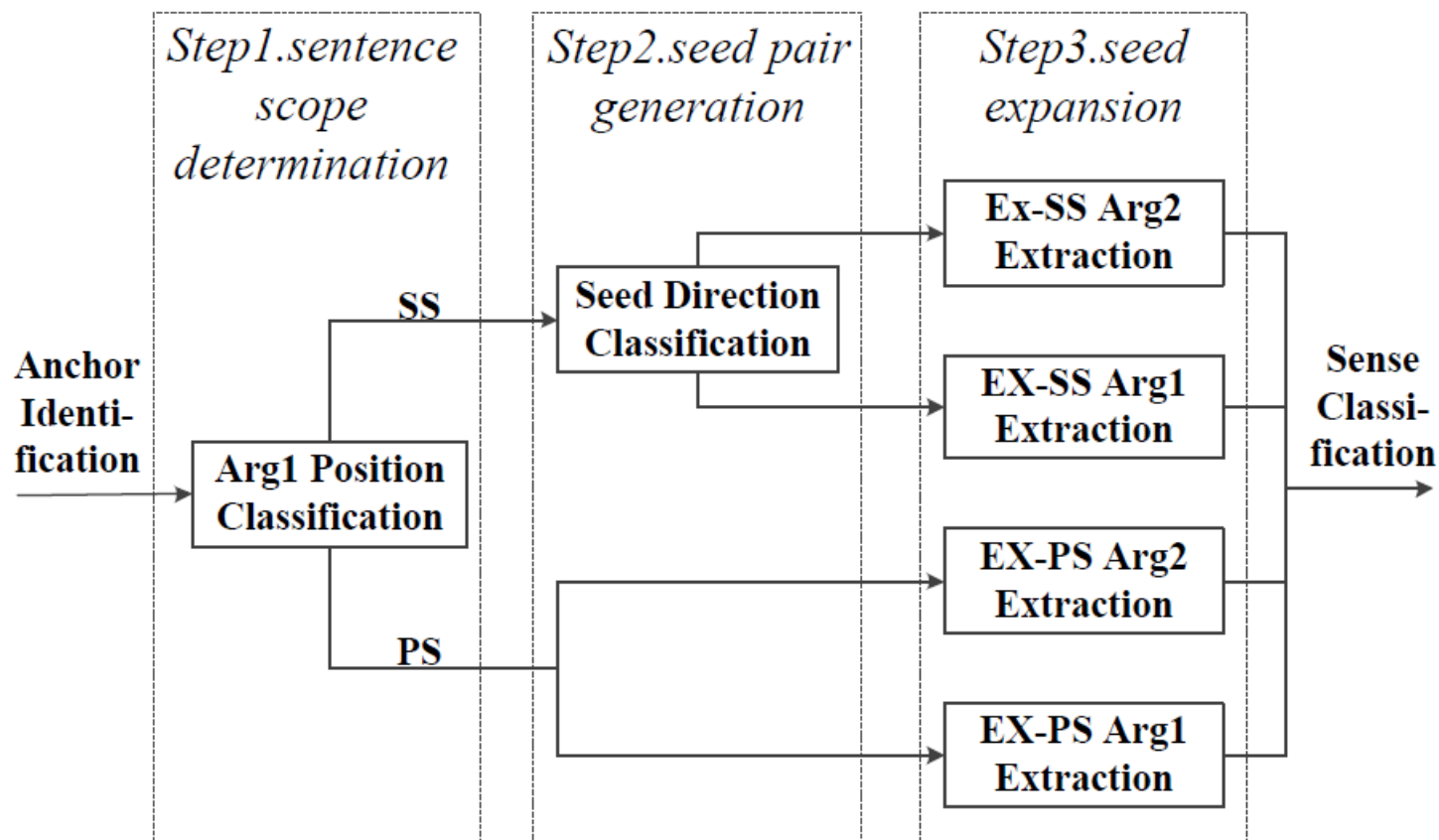
②Argument Extraction

➤ Some observations

- ✓ In most cases, Arg1 and Arg2 are in the same sentence or two adjacent sentences respectively;
- ✓ An argument consists of one or several consecutive clauses;
- ✓ Explicit Arg2 is located in the same sentence as its connective anchor;
- ✓ In most cases, the span of Arg1 and the span of Arg2 are adjacent. There is no clause between them.

2. 篇章关系分析

➤ Seed of Argument-Pair expansion





2. 篇章关系分析

Step1: Sentence scope determination

Is Arg1 in the SS(same sentence) or PS(previous sentence) of Arg2?

- **Explicit:** determined by Arg1 position classification
- **Non-explicit:** determined by anchor(MOS or EOS)

2. 篇章关系分析

举例

● Explicit: SS

{ 【今年 7 月，在里海划分问题上俄以退为进，对其原先坚持的“只分边缘、中间共管”的立场作了重大调整，接受哈萨克斯坦提出的“按中心线划分海底、水域共享”的原则，并不顾伊朗、土库曼斯坦等其它里海沿岸国家的反对，同哈就划分里海北部地区达成协议。】 }

● Non-explicit:

{ 【目前在中国境内承包工程的国外承包商已有一百三十七家，承包的工程达一百四十一项，……】 }

2. 篇章关系分析

Step2: Seed pair generation

➤ Explicit:

① If it is SS, another seed will be determined by seed direction classifier



② If it is PS, the clause with the connective and the last clause in the previous sentence are a seed pair

➤ Non-explicit:

the left and right clauses

Features
connective itself
POS of the connective
whether there are co-occurrence of nouns, verbs and quantifiers between current clause and previous/next clause
the parent of previous/next clause's node the punctuation between previous/next clause and current clause
the relationship of previous and current clause's node (left, right, middle, contain, none)



2. 篇章关系分析

举例

● Explicit:

{ 【今年 7 月，在里海划分问题上俄以退为进，对其原先坚持的“只分边缘、中间共管”的立场作了重大调整，接受哈萨克斯坦提出的“按中心线划分海底、水域共享”的原则，并不顾伊朗、土库曼斯坦等其它里海沿岸国家的反对，同哈就划分里海北部地区达成协议。】 }

● Non-explicit:

{ 【目前在中国境内承包工程的国外承包商已有一百三十七家，承包的工程达一百四十一项，……】 }



2. 篇章关系分析

Step3: Seed expansion

- Toward a fixed direction (forward or backward)
- Expand clause-by-clause
- Select the longest candidate predicted OK



2. 篇章关系分析

➤ Features for seed expansion

	Features
Lexical	anchor itself
	punctuations between the previous candidate and the current clause
	POS of the first and last word of the previous candidate and the current clause
	whether there are co-occurrence of nouns between the previous candidate and the current clause
	whether there are co-occurrence of verbs between the previous candidate and the current clause
Syntactic	the parent of anchor's node
	the current clause's node and its left and right siblings
	the current candidate's node and its parent
	the path from previous candidate's node to current clause's node
	the path from previous seed clause's node to current clause's node
	the relationship of current clause and the seed clause (left, right, middle, contain, none)
	the relationship of current clause and the previous candidate
Others	whether the current clause is the start/end of sentence
	the relative length of current clause and seed clause (short, middle, long)
	the relative length of current clause and previous candidate

2. 篇章关系分析

举例

● Explicit:

{ **【** 今年 7 月，| 在里海划分问题上俄以退为进，| 对其原先坚持的“只分边缘、中间共管”的立场作了重大调整，| **C4** 接受哈萨克斯坦提出的“按中心线划分海底、水域共享”的原则，| 并不顾伊朗、土库曼斯坦等其它里海沿岸国家 **C5** 的反对，| 同哈就划分里海北部地区达成协议。**】** }

C1 **C2** **C3**

C6

Clause1, Clause2, Clause3, Clause4, || Clause5, Clause6

{C4||C5}

{C3,C4||C5}

{C2,C3,C4||C5}

{C1,C2,C3,C4||C5}

{C4||C5,C6}

{C3,C4||C5,C6}

{C2,C3,C4||C5,C6}

{C1,C2,C3,C4||C5,C6}

2. 篇章关系分析

● Explicit + Implicit 混合情况：

{ **【** 尽管 她的动作潇洒自如, | 但 难度无法与罗莉相比, | 只获得 9.875分, | 夺得银牌。 **】** }

Diagram illustrating clause relationships in the sentence:

- C1** points to 尽管
- C2** points to 但
- C3** points to 只获得 9.875分, | 夺得银牌。

C4

Clause1, Clause2, || Clause3, Clause4

Diagram illustrating clause relationships in the sentence:

- C4** points to Clause1
- A red arrow points from Clause1 to Clause2
- A blue arrow points from Clause3 to Clause4

{C2||C3}

{C2||C3,C4}

{C1,C2||C3,C4}



2. 篇章关系分析

③Sense Classification

Features		Explicit	Non-explicit
Lexical	connective itself	✓	
	POS of the connective	✓	
	embedding of the connective	✓	
	the previous and next punctuation of the connective	✓	
Syntactic	the parent of connective's node	✓	
	the left and right siblings of connective's node	✓	
	the Arg1's node and Arg2's node	✓	
	the parent of Arg1's node and Arg2's node	✓	
	the relationship of Arg1's node and Arg2's node	✓	
	the production rules of Arg1		✓
	the production rules of Arg2		✓
	the production rules of Arg1 and Arg2		✓



2. 篇章关系分析

④Argument Relabeling

	Features
Lexical	anchor itself
	POS of previous and next word of the anchor
	location of the anchor in the sentence
	whether there are co-occurrence of nouns between Arg1 and Arg2
	whether there are co-occurrence of verbs between Arg1 and Arg2
	whether there are co-occurrence of quantifiers between Arg1 and Arg2
Syntactic	the parent of connective's node
	the left and right siblings of connective's node
	the Arg1's node and Arg2's node
	the parent of Arg1's node and Arg2's node
	the relationship of Arg1's node and Arg2's node
Others	the relative length of Arg1 and Arg2
	the relation sense



2. 篇章关系分析

Argument标签重新标记的目的是进一步确定Arg1和Arg2孰前孰后。Arg1和Arg2的前后关系是根据语义决定的，而不是位置。例如在因果关系中，Arg1表示原因，Arg2表示结果。如：

我生病了，所以今天没来上班。

“我生病了”为Arg1，“今天没来上班”为Arg2。

如果句子变成：我今天没来上班，因为我生病了，那么“我今天没来上班”为Arg2，“我生病了”为Arg1。前面的三个步骤中只是为了方便暂时将前一个命名为Arg1，后一个命名为Arg2，但这并不影响分类结果，因为两个Argument提取的特征都是对称的。



2. 篇章关系分析

- Data sets for the shared task of CoNLL'2016

Sets	# Documents	# Sentences	# Relations
Train (CDTB)	455	6,332	10,240
Dev	24	349	383
Test	30	348	455
Blind	64	1,140	2,101

2. 篇章关系分析

● Results

*More than **20%** sharp decrease of F_1 in explicit parser on the blind set. This is mainly due to the error propagation of discourse connective identification.*

	Task	Dev	Test	Blind
Explicit	Conn	0.8356	0.7263	0.5627
	Arg1	0.5479	0.5587	0.3853
	Arg2	0.6849	0.6816	0.4444
	Both	0.4521	0.4916	0.2650
	Sense	0.7534	0.6480	0.4811
	Parser	0.4521	0.4859	0.2446
Non-Explicit	Conn	—	—	—
	Arg1	0.6282	0.6266	0.5526
	Arg2	0.6798	0.6762	0.6017
	Both	0.5341	0.5379	0.4457
	Sense	0.5068	0.4987	0.4082
	Parser	0.3982	0.3869	0.2712
All	Conn	0.8356	0.7263	0.5627
	Arg1	0.6261	0.6328	0.5439
	Arg2	0.6932	0.6921	0.5843
	Both	0.5317	0.5418	0.4178
	Sense	0.5640	0.5333	0.4326
	Parser	0.4120	0.4089	0.2690

2. 篇章关系分析

- The error of connective identification

- The flexible parallel connectives (组合关联词)

另外，女单 12 号种子、德国选手哈克也在第一轮比赛中被淘汰，男单 16 号种子、捷克选手科达赛前因伤退出了比赛。

- The connectives in the middle of the sentence

刘华清说，中泰两国人民有传统的友谊，两国的关系也十分友好。😊

杰克逊和鲍威尔参加的男子 110 米栏和跳远这次也被列入非大将赛项目。😞

2. 篇章关系分析

*The seed-expansion method can get acceptable argument. The F_1 of Arg1 and Arg2 individually is about **10%** higher than jointly.*

	Task	Dev	Test	Blind
Explicit	Conn	0.8356	0.7263	0.5627
	Arg1	0.5479	0.5587	0.3853
	Arg2	0.6849	0.6816	0.4444
	Both	0.4521	0.4916	0.2650
	Sense	0.7534	0.6480	0.4811
Non-Explicit	Parser	0.4521	0.4859	0.2446
	Conn	—	—	—
	Arg1	0.6282	0.6266	0.5526
	Arg2	0.6798	0.6762	0.6017
	Both	0.5341	0.5379	0.4457
All	Sense	0.5068	0.4987	0.4082
	Parser	0.3982	0.3869	0.2712
	Conn	0.8356	0.7263	0.5627
	Arg1	0.6261	0.6328	0.5439
	Arg2	0.6932	0.6921	0.5843
	Both	0.5317	0.5418	0.4178
	Sense	0.5640	0.5333	0.4326
	Parser	0.4120	0.4089	0.2690



2. 篇章关系分析

➤ The error of argument extraction

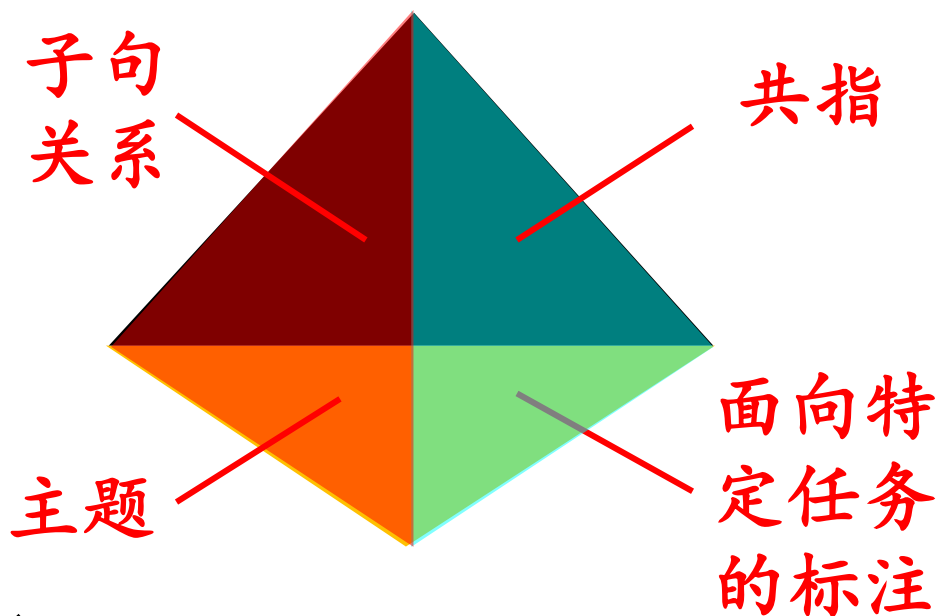
由于 备战 广岛 亚运会，我 国 大部分 田径 好手
没有 报名 参加 此次 比赛，国家队 只有 2 1 名
选手 在 短跑、接力、跳高、跳远、铅球 等 项
目 中 露面，辽宁省 女子 中长跑队 也 没有 前来。

请参阅:

X. Kang et al. An End-to-End Chinese Discourse Parser with Adaptation to Explicit and Non-explicit Relation Recognition. *Proc. CoNLL: Shared Task*. Berlin, August 11-12, 2016

2. 篇章关系分析

◆ **基本观点：** 基于通用语料获取基础语法，面向具体应用建立任务模型的多视角、多层次汉语语篇标注体系和分析系统。





3. 延伸阅读

Xiaomian Kang *et al.* A Survey of Discourse Representations for Chinese Discourse Annotation. *ACM TALLIP*, 2019.1, 18(3)

Fang Kong and Guodong Zhou. A CDT-Styled End-to-End Chinese Discourse Parser. *ACM TALLIP*, 2017.9, 16(4)

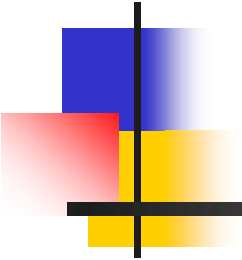
Kong Fang *et al.* A constituent-based approach to argument labeling with joint inference in discourse parsing. *Proc. EMNLP*, Oct. 25-29, 2014, Qatar. Pages 68-77

宋洋, 王厚峰. 共指消解研究方法综述, 中文信息学报, 2015(1): 1-12

宋洋, 王厚峰. 基于马尔可夫逻辑网络的中文零指代消解, 计算机研究与发展, 52(9): 2114-2122, 2015

宋柔, 汉语篇章广义话题结构研究, 北京语言大学语言信息处理研究所研究报告, 2012

孔芳等, 指代消解综述, 计算机工程, 36(8), 2010



Thanks

谢谢!