北京大学学报(自然科学版) 第 53 卷 第 2 期 2017 年 3 月
Acta Scientiarum Naturalium Universitatis Pekinensis, Vol. 53, No. 2 (Mar. 2017)
doi: 10.13209/j.0479-8023.2017.038

基于中英文可比较语料的中文零指代消解

杨紫怡 贡正仙 孔芳† 周国栋

苏州大学计算机科学与技术学院，自然语言处理实验室，苏州 215006; † 通信作者, E-mail: kongfang@suda.edu.cn

摘要 针对中文篇章中的零指代问题，提出一种基于中英文可比较语料进行中文零指代识别和消解的方法，并提出英文对等句的概念。利用对等句，重新定义句子间隔，并引入双语词对齐特征。在基准平台基础上，从零指代项识别和零指代项消解两个方面进行研究。在 OntoNotes5.0 语料上的实验结果表明，与目前性能最好的系统相比，新提出的基于中英对等语料的中文零指代方法取得更好的性能。

关键词 中文零指代；双语；对等句；识别；消解

中图分类号 TP391

Exploit Comparable Corpus to Chinese Zero Pronoun Resolution

YANG Ziyi, GONG Zhengxian, KONG Fang†, ZHOU Guodong

Natural Language Processing Laboratory, School of Computer Science and Technology, Soochow University, Suzhou 215006;
† Corresponding author, E-mail: kongfang@suda.edu.cn

Abstract A bilingual approach based on a comparable corpus is proposed to better detect and to resolve Chinese zero pronouns. The concept of English equivalent sentence is defined firstly. Then the equivalent sentence is employed to redefine the distance between sentences and to extract bilingual word alignment features. In this way, both zero pronoun detection and resolution of the baseline system from bilingual perspective are improved. The experiments conducted on the OntoNotes5.0 corpus show that the proposed approach can significantly outperform the state-of-the-art system.

Key words Chinese zero pronoun; bilingual; equivalent sentence; detection; resolution

零指代(Chinese zero pronoun, ZP)是指代的一种。作为一种常见的语言现象，零指代在自然语言的多种表述中广泛存在，尤其在中文语言中非常普遍。一定的语境中，在不影响原有语义表达的情况下，省略了用于指代前文语言单位的语言成分，但其语义依然依赖于前文的语言单位，这个省略的句法成分就称为零指代，零指代回指的语言单位称为先行词(Antecedent)。对于省略的语言成分，读者根据上下文语境，可以轻易地理解其含义。但是，缺省的语言成分会对机器的理解造成很大的困扰。机器需要通过一定的方法，从上下文获取缺省的信息，从而进行下一步的理解，这个过程就是零指代消解。随着自然语言处理研究的不断深入，在机器翻译、篇章理解以及信息抽取等领域，零指代消解都起到相当重要的作用，因此日益被研究者关注。

尽管有关零指代在中文方面的研究已经取得一定的成果，但是中文零指代的性能仍然不尽如人意。汉语的复杂性带来的是比英语难度更高的句法分析和语义角色标注，大量存在的缺省和长句给中文零指代研究带来更大的挑战。目前主流的中文零指代方法几乎完全基于中文语料进行识别和消解，没有考虑句法结构更加严谨、句法分析更加精准的英文翻译对中文零指代产生的正面影响。

因此，本文提出一种基于中英文可比较语料的中文零指代消解方法，引入英文可比较语料来协助中文零指代的识别和消解，并提出英文对等句的概

国家自然科学基金(61333018, 61472264, 61305088)资助
收稿日期: 2016–07–21; 修回日期: 2016–10–03; 网络出版日期: 2016–11–28

念。利用英文对等句，将中文长句划分为若干子句(Clause)，以子句为单位来识别零指代项。重新定义句子间隔，引入双语词对齐特征，对零指代项进行消解。

本系统在 OntoNotes5.0[1]语料上进行实验，结果显示该方法的性能比现在性能最好的系统有显著提升。

1 相关工作

尽管零指代现象在中文中出现的频度很高(据 Kim[2]2000 年统计，中文中存在句法成分省略现象的比例高达 36%，而英文中句法结构比较严谨，省略现象少见，大概不超过 4%)，但中文零指代直到最近才成为研究热点。中文零指代的研究方法可分为有监督方法和无监督方法两类。

有监督方法是基于已有的标注语料训练生成模型对生文本进行预测的方法。Zhao 等[3]给出一个完整的基于机器学习的中文零指代消解方案，并提出一套有效的适于中文零指代任务的特征集合。但是，他们的工作主要关注零指代的消解子任务，对零指代项的识别仅给出一个保证高召回率的规则方法。他们的实验结果也表明，过低的零指代项识别准确率会严重影响后续消解的性能。Kong 等[4]给出一个中文零指代消解的完整框架，将中文零指代消解清晰地划分成零元素识别、零待消解项识别和零元素消解 3 个子任务，并采用基于树核函数的方法，分别给出每一个子任务适用的结构化特征集。但是，他们仅关注平台的统一性，只给出标准句法树上平台的性能，未给出完全自动状况下方法有效性的验证。Chen 等[5]首次给出完整的端到端的全自动状况下的中文零指代消解平台，并提出一组更有效的句法和上下文特征。

有监督方法无法避免对标注语料的依赖性。为了解决这一问题，Chen 等[6]给出一个无监督方法的生成式模型，并借助它进行中文零指代消解。基于这一工作，Chen 等[7]进一步在生成式模型中基于概率将零待消解项识别和消解任务进行联合学习，取得目前最好的性能。

不论采用有监督还是无监督方法，目前中文零指代消解的研究都集中在单语模式下。已经有研究表明，零指代的存在是影响中英文翻译性能的关键要素之一[8]。因此，本文尝试在双语可比较语料上进行中文端的零指代消解研究。

2 中文零指代基准平台

在中文句子中，零指代需同时满足两个必要条件：1) 在句法结构上不是一个完整的句子，在句法成分上存在缺省；2) 句中缺省的成分指代前文中某一个特定的真实实体。只有同时满足这两个条件的缺省，才能被判定为零指代项。所以，一般情况下，传统的中文零指代消解包含零指代项识别和零指代项消解两个部分。中文零指代项识别，即确定在中文篇章中哪些位置存在句法成分缺省并回指了前文中某个语言单位；中文零指代项消解，即确定已识别出的零指代项所回指的特定真实实体，即先行词。本文构建的中文零指代基准平台也由这两个模块构成。

2.1 零指代项识别

在本文构建的基准平台中，零指代项识别模块包含两个步骤：零指代候选集生成和零指代项判定。在中文中，零指代项主要以 3 种形式存在：1) PRO + VP + NP; 2) NP + VP + PRO; 3) PRO + PP + NP。根据这 3 种形式，我们在进行零指代候选集生成时，预定义以下 3 个规则来抽取潜在的零指代项，生成零指代项候选集合。

规则 1 当前节点是 VP 节点，同时该 VP 节点的父节点不是 VP 节点时，若它的左兄弟节点不是 NP 节点，或者该 VP 节点没有左兄弟节点，则该 VP 节点左边相邻位置是零指代项候选。

规则 2 当前节点是 VP 节点，同时该 VP 节点的父节点不是 VP 节点时，若是子节点中存在及物动词且没有右兄弟节点，则该 VP 节点后边相邻位置是零指代项候选；若该 VP 节点的子节点是一个不及物动词且后兄弟节点不存在，则该 VP 节点的右边相邻位置不是一个零指代项候选。

规则 3 当前节点是 PP 节点，同时该 PP 节点的父节点不是一个 PP 节点时，若它的左兄弟节点不存在，则该 PP 节点的左边相邻位置是一个零指代项候选。

零指代项候选集合生成后，需要对在候选集中出现的每一项进行判定，确定其是否为真正的零指代项。在零指代项的判定过程中，我们采用 Zhao 等[3]提出的部分特征，并额外地扩展一部分特征，使用分类器进行零指代候选集的判定。表 1 为本模块中零指代项判定所用到的特征。其中 Z 是一个零指代项候选，我们规定 Wl 和 Wr 分别表示 Z 的

表 1　中文零指代项识别特征

Table 1　Chinese zero pronoun detection features

序号	特征	说明
F1	First_Gap	若 Z 是句子第一个缺省则 T, 否则 F
F2	Pl_Is_NP	若 Z 是句子第一个缺省则 NA, 否则如果 Pl 是一个 NP 节点则 T, 否则 F
F3	Pr_Is_VP	若 Z 是句子第一个缺省则 NA, 否则如果 Pr 是一个 VP 节点则 T, 否则 F
F4	Pl_Is_NP&Pr_Is_VP	若 Z 是句子第一个缺省则 NA, 否则若 Pl 是一个 NP 节点且 Pr 是一个 VP 节点, 则 T, 否则 F
F5	P_Is_VP	若 Z 是句子第一个缺省则 NA, 否则如果 P 是 VP 节点则 T, 否则 F
F6	IP_VP	Wr 节点到 C 的节点路径上, 若存在一个 VP 节点且它的父节点是一个 IP 节点则 T, 否则 F
F7	Has_Ancestor_NP	如果 V 有一个 NP 祖先节点则 T, 否则 F
F8	Has_Ancestor_VP	如果 V 有一个 VP 祖先节点则 T, 否则 F
F9	Has_Ancestor_CP	如果 V 有一个 CP 节点则 T, 否则 F
F10	Left_Comma	如果 Z 是句子的第一个缺省则 NA, 否则如果 Wl 是一个逗号则 T, 否则 F
F11	LastWord	Z 的左边相邻的分词
F12	NetxtWord	Z 的右边相邻的分词
F13	LastWordPos	Z 的左边相邻的分词词性
F14	NextWordPos	Z 的右边相邻的分词词性
F15	LastWordPos_NextWordPos	Z 的左边相邻的分词词性+Z 的右边相邻的分词词性

左右单词节点。P 是 Wl 节点与 Wr 节点的最低公共祖先节点的句法树节点。Pl 是 P 节点的孩子节点，且 Pl 是 Wl 的祖先节点。类似地，Pr 是 P 节点的子节点，且 Pr 是 Wr 的祖先节点。如果 Z 是句子的第一个缺省，那么 Wl，P，Pl 和 Pr 都将默认为 NA；如果 Z 不是句子的第一个缺省，则规定最高的顶点 C 作为 P 节点，否则，C 做为整棵句法树中最高的根顶点。

2.2　零指代项消解

本文中基准平台的另一个模块为零指代项消解模块。零指代项消解的任务就是为识别出来的零指代项寻找合适的先行词。在本模块中，我们使用语料中提供的分句、分词、词性标注、句法分析、命名实体识别和语义类别等预处理信息。在先行词候选集的生成过程中，我们将语料中所有的最长名词性短语和充当修饰性词的名词短语提取出来并加入候选集合。例如(NP (NN 建筑)(NN 公司))里最长的名词性短语是“建筑公司”，充当修饰性词语的名词短语是“建筑”，所以我们提取“建筑公司”和“建筑”，并将它们加入先行词候选集。

在模块的构建过程中，样例的生成是一个核心的环节。训练时，对于每一个零指代项，首先根据训练语料中标注的指代链信息来判断它是否存在于某一指代链。若不存在任何指代链，则认为是非零指代待消解项，无需配对形成训练实例；倘若该零指代项属于某一指代链，则将当前的零指代项从后向前依次与其前面出现的名词性短语进行配对，直至遇到同一条指代链上的另一个名词性短语为止。位于同一指代链的名词性短语配对定义为正例，其余为负例。测试时，每一个识别出的零指代项依次向前与其前面的名词性短语进行配对，抽取向量特征交给训练模型进行分类。若被判断为正例，则说明两者之间构成指代关系，零指代项指向该名词性短语，配对结束。若被判断为负例，则说明不存在指代关系，继续向前寻找其先行词，直至文本开始部分。在判断先行词候选集中的候选项与零指代项是否有指代关系的过程中，我们的基准平台使用的特征如表 2 所示。

3　基于双语的中文零指代消解

英文的词法、句法结构比中文更严谨、更规范。从中英文双语对照的角度，对 OntoNotes5.0 语料统计分析后，我们发现以下规律。

1) 中文里包含零指代项的小句，在英文中可翻译成独立小句或从句。不论形式上如何变化，具有指代关系的零指代项一般在英文翻译中都会出现显

表 2 零指代项消解特征

Table 2 Zero pronoun resolution features

序号	特征	说明
1	Dist_Sentence	若 Z 与 A 在同一个句子中为 0, 相差一个句子则为 1, 依此类推
2	Dist_Segment	若 Z 与 A 在同一个分句中为 0, 相差一个分句则为 1, 依此类推
3	Sibling_NP_VP	若 Z 与 A 在不同的句子中为 F, 否则若都是根节点的子节点且是兄弟节点则 T, 否则 F
4	Closet_NP	若 A 是距离 Z 最近的候选先行词则 T, 否则 F
5	A_Ith_Peroson	若 A 是第一、二、三人称, 中性, 未知, 分别对应取值 First, Second, Third, Neutral, Others
6	A_Role	若 A 是主语、宾语或者其他, 分别对应取值 Subject, Object, Others
7	A_Has_Anc_NP	若 A 有一个 NP 祖先节点则 T, 否则 F
8	A_Has_Anc_NP_In_IP	若 A 有一个 NP 祖先节点且该节点是 A 最低 IP 祖先节点的后代则 T, 否则 F
9	A_Has_Anc_VP	若 A 有一个 VP 祖先节点则 T, 否则 F
10	A_Has_Anc_VP_In_IP	若 A 有一个 VP 祖先节点且该节点是 A 最低的 IP 祖先节点的后代则 T, 否则 F
11	A_Has_Anc_CP	若 A 有一个 CP 祖先节点则 T, 否则 F
12	A_Grammatical_Role	若 A 在句子中所承担的语法角色是主语、宾语或其他, 则特征值取为 S, O 或 X
13	A_Cluse	若 A 在主句、独立分句、附属从句或这 3 种情况之外, 特征值分别对应取 M, I, S, X
14	A_Is_ADV	若 A 是状语 NP 节点则 T, 否则 F
15	A_Is_TMP	若 A 是一个时间 NP 则 T, 否则 F
16	A_Is_Pronoun	若 A 是一个代名词则 T, 否则 F
17	A_Is_NE	若 A 是一个命名实体则 T, 否则 F
18	A_In_Headline	若 A 存在文本的标题中则 T, 否则 F
19	Z_Has_Anc_NP	若 V 有一个 NP 祖先节点则 T, 否则 F
20	Z_Has_Anc_NP_In_IP	若 V 有一个祖先节点且该节点是 V 的最低 IP 祖先节点的后代节点则 T, 否则 F
21	Z_Has_Anc_VP	若 V 有一个 VP 祖先节点则 T, 否则 F
22	Z_Has_Anc_VP_In_IP	若 V 有一个 VP 祖先节点, 并且该节点是 V 的最低 IP 祖先节点的后代节点则 T, 否则 F
23	Z_Has_Anc_CP	若 V 有一个 CP 祖先节点则 T, 否则 F
24	Z_Grammatical_Role	若零指代项 Z 的语法角色是主语则 S, 否则 X
25	Z_Clause	若 V 在主句, 独立分句, 附属从句, 或者三种以外句子中, 特征值分别对应为 M, I, S, X
26	Z_Is_First_ZP	若 Z 是所在句子第一个零指代项候选则 T, 否则 F
27	Z_Is_Last_ZP	若 Z 是所在句子最后一个零指代项候选则 T, 否则 F
28	Z_In_Headline	若 Z 存在文本的标题中则 T, 否则 F

式的对照成分(代词、名词或与其他子句共享主语)。这种现象可以在中文零指代项识别的过程中协助判定是否为零指代项, 从而提高零指代项识别的准确率。

2) 中文里的长句通常会被翻译成英文的多句, 而中文句子与英文句子两者间的一对多关系恰好体现了中文长句中标点符号, 特别是逗号在不同上下文中承担的不同功能。利用中英对等句将长句切分为若干子句(Clause), 可以提高句法分析的效果。

3) 单个子句通常最多包含一个零指代项(在 OntoNotes5.0 语料中比例大于 99%), 因此, 基于子句进行零指代项的识别和消解可以获得更高的准确率。

4) 由于零指代消解在很大程度上依赖于句法分析和语义角色标注(semantic role labeling, SRL)的性能, 句法分析和 SRL 的精准与否会直接影响零指代消解的性能, 而英文句法分析和 SRL 的准确率相较于中文有很大提高, 因此利用中英文对等句, 对确定的中文子句进行调整, 可以提高零指代项消解的性能。

基于以上原因，我们提出基于中英文可比较语料的中文零指代方法。针对零指代项的识别和消解，融入双语信息，分别提出基于自动对齐机制的英文对等句识别方案以及基于简化的语义角色标注框架在英文对等句中识别子句的方法。改进后的方法流程如图 1 所示。

3.1　基于自动对齐机制的英文对等句识别

基于对 OntoNotes5.0 语料库的观察和分析，本文提出英文对等句的概念。我们认为中文长句中以标点符号分割的各小句间，若其英文翻译以独立句子出现，则中文也可作为独立句子；若其英文翻译以从句形式出现，则中文也与其他小句合并成独立语句。按此方式得到的新的语句划分称为英文对等句。考虑到标点符号的使用频率，本文仅考虑逗号和分号。

借助 OntoNotes5.0 语料中提供的部分中英文可比较语料和自动对齐技术，我们将中文的各语句对齐到英文翻译中，再提取其中一对多形式的中英文句对，用人工干预的方式，对其中出现的逗号和分号是否承担句子分割标记进行标注。可比较语料中的其他逗号和分号均认为是普通标点符号，不承担英文对等句标识的作用，以此形成中英文对等句语料库。

我们以语料库中的逗号和句号为分析对象，参考 Xue 等[9]的工作，借助机器学习方法生成英文对等句识别器。对给定的文档，首先进行句子识别、分词和句法分析；以句子为单位，提取其中的逗号和分号列表；针对提取的每一个标点，抽取其所在句子的上下文特征；根据中英文对等句子语料中标注的正负例信息，确定每个标点符号的类别；借助最大熵分类算法进行训练，获得对等句识别器模型或在测试阶段进行对等句分类。

从上述流程可以看到，对等句识别器构建的核心是上下文特征的提取。下面，逗号概指逗号或者分号。本文使用的特征集合包括：1) 从逗号到前一逗号或句首的范围内，前面 *N* 个词的词性及词，后面 *N* 个词的词性及词；2) 逗号之后第一个词的词性和词；3) 逗号左右兄弟及其组合的句法信息标记；4) 逗号左右兄弟及逗号父亲的句法信息组合；5) 从逗号到前一逗号或句首是否包含“是”(VC)、表语形容词(VA)、“有”(VE)、其他动词(VV)和从属连词(CS)，从逗号到后一逗号或句尾是否包含“是”(VC)、表语形容词(VA)、“有”(VE)、其他动词(VV)和从属连词(CS)；6) 逗号的父节点是否为并列的 IP 结构，逗号是否为第一层子节点，逗号的父节点是否是第一层并列 IP 结构；7) 逗号所在句子标点的集合；8) 逗号到前一标点或句首句子长度是否小于 5，逗号左右两边句子长度差是否大于 7；9) 逗号所在句法树层次；10) 逗号的父节点、左兄弟节点、右兄弟节点是否为 NP；11) 从本标点到前一个标点或句首的句子中，第一个词及最后一个词的词性及词的组合；12) 逗号前后单元包含的相同词及词性信息。

3.2　基于子句的零指代识别和消解方法

参考 Kong 等[10]在空语类识别中的处理方法，我们从句法层面定义子句，将包含独立谓词及其驱

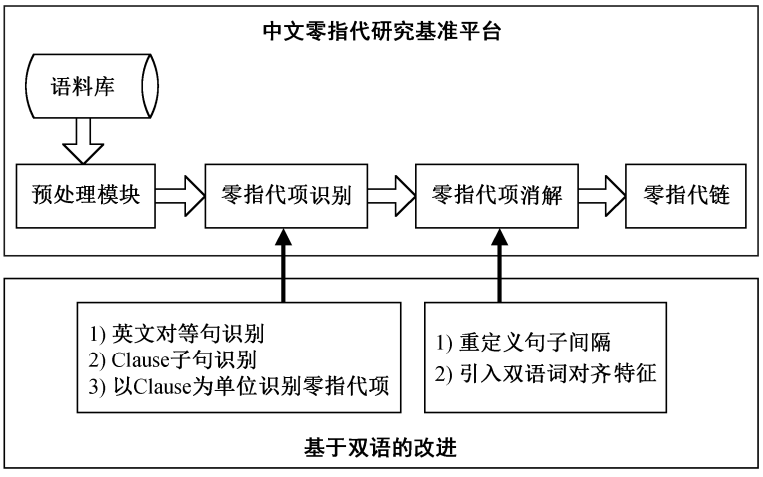

图 1　双语的零指代改进

Fig. 1　Improvement of bilingual zero pronoun

动的所有论元的最小子树称为一个独立的子句。这样就可以借助句法分析和 SRL 工具完成中文子句的识别。考虑到在中文句法分析及自动句法分析基础上的 SRL 性能，本文的子句识别方法有三方面的改进：1) 不再考虑完整的句子，仅在对等句上进行谓词及其论元的识别；2) 将 SRL 任务进行了简化，不再进行角色标注任务，简化后的子句识别由谓词识别(本文仅关注动词性谓词)、论元裁剪和识别以及最小子树提取 3 个模块组成；3) 将对等句借助工具自动翻译成英文，在英文结果上进行简化版的 SRL，并基于 SRL 的结果进行最小子树提取，确定英文中的子句。最后根据中英文词对齐信息，对确定的中文子句进行调整。

我们以得到的中文子句为单位，进行后续的零指代项识别和消解。同一个英文对等句可以包含多个子句，且多个子句间可能存在嵌套关系。根据识别出的子句是否包含其他子句，我们将子句分成元子子句和复合子句两种。首先提取元子子句的特征，确定它们是否包含零指代项(即正、负例)。若元子子句包含零指代项，则包含它的所有复合子句不再考虑，否则进一步考虑复合子句。在考虑复合子句时，将内嵌的其他子句看做一个特殊的不可分的句法成分。

与基准系统相比，后续零指代项识别和消解主要有以下三方面的改进：1) 零指代项的识别以子句为单位进行，每个识别出的子句对应一个训练实例；2) 对于基准平台消解过程中先行词候选集的选定，我们将跨越的子句数量作为一个约束条件，认为零指代关系跨越不会超过 6 个子句；3) 在零指代消解中引入英文翻译与其中文部分进行词对齐后的相关特征。

新引入的双语词对齐特征：1) 候选先行词的英文翻译及其对应的词性；2) 中文零指代项在英文翻译中对应的词及其词性；3) 中文原文中，零指代项与先行词候选间是否包含其他零指代项；4) 英文翻译中，零指代项与先行词候选间是否包含其他代词；5) 中文原文中，零指代项与先行词候选间的子句路径(我们选择每个子句在句法树上的根节点来表示一个子句)；6) 英文翻译中，零指代项与候选先行词间的子句路径。

4 实验与分析

4.1 实验设置

我们的零指代基准平台和改进后的零指代系统均在 CoNLL2012 shared task[11]的语料库 OntoNotes5.0 上进行实验。为了便于比较基准平台与本文提出的改进方法，我们的零指代基准平台在 CoNLL2012 shared task 提供的自动句法树(Auto)上分别对数据域 NW, MZ, WB, BN, BC 和 TC 进行实验，并对全部语料 Overall 进行实验评测。评价指标采用通用的召回率(*R*)、准确率(*P*)和 *F* 值。

4.2 实验结果分析

与其他场景相比，我们更关注完全自动情况下的性能。表 3 给出基准系统在自动句法树情况下，以全自动方式进行英文对等句识别，再借助自动 SRL 和机器翻译获得子句，在此基础上进一步进行零指代项的识别和消解的性能情况。为公平比较，表 3 也列出目前结果最好的 Chen 等[7]的和我们的基准平台在 OntoNotes5.0 语料上 NW，MZ，WB，BN, BC 和 TC 这 6 个数据集上的性能情况，以及这 6 个数据集整体上的性能情况。从表 3 可以得到以下结果。

表 3 系统性能比较

Table 3 Comparison of system performance %

数据域	Chen 等[7]			基准系统			引入双语的改进系统		
	R	*P*	*F*	*R*	*P*	*F*	*R*	*P*	*F*
NW	11.9	14.3	13.0	13.4	15.7	14.5	18.2	23.9	20.7
MZ	4.9	4.7	4.8	8.9	7.8	8.3	9.7	13.4	11.3
WB	20.1	14.3	16.7	14.2	11.4	12.6	16.2	14.5	15.3
BN	18.2	22.3	20.0	18.5	24.1	20.9	19.7	21.4	20.5
BC	19.4	14.6	16.7	21.6	14.3	17.2	22.9	19.2	20.9
TC	31.8	17.0	22.2	30.1	15.6	20.5	29.7	19.4	23.5
整体	19.6	15.5	17.3	20.3	15.8	17.8	25.4	20.7	22.8

1) 我们的改进系统在不同数据集(除 WB 外)上的表现都优于目前性能最好的 Chen 等[7]的系统，改进系统的总体性能比 Chen 等[7]系统的 *F* 值高出约 5.5%。具体来看，在 NW 数据集上，改进系统的 *F* 值为 20.7%，相较于 Chen 等[7]系统的 13.0%，有 7.7 个百分点的提升。在 MZ 数据集上，改进系统的 *F* 值有 6.5%的提升。在 WB 数据集上，改进系统的 *F* 值却有 1.4%的降低。在 BN 数据集上，改进系统的 *F* 值提升幅度最小，仅有 0.5%。在 BC 数据集上，改进系统的 *F* 值有 4.2%的提升。在最后的 TC 数据集上，改进系统的 *F* 值也有 1.3%的提升。总体而言，我们在基准平台上引入双语的改进系统性能已超过目前已知性能最好的 Chen 等[7]的系统。

2) 与我们的基准平台相比，引入双语的改进系统在 BN 集上的性能有 0.4%的轻微下降，而在其余各数据集上都有不同程度的提升。总体而言，相对于基准平台的 *F* 值 17.8%，改进系统提高约 5%。具体来看，在 NW 数据集上，相对于基准平台的 *F* 值 14.5%，改进系统提升 6.2%。在 MZ 数据集上，相对于基准平台的 *F* 值 8.3%，改进系统提升 3.0%。在 WB 数据集上，改进系统的 *F* 值提升 2.7 个百分点。在 TC 与 BC 数据集上，改进系统的 *F* 值分别提升 3.0%和 3.7%。

为了进一步分析引入的双语信息对中文零指代消解的两个环节(零指代项识别和零指代项消解)的贡献，我们在标准句法树和自动句法树两种情况下进行实验。

表 4 给出基准系统和改进后系统在标准句法树和自动句法树情况下零指代项识别的性能。从表 4 可以看出，与使用自动句法树的情况相比，引入的双语在使用标准句法树情况下的作用大大减小，这主要是因为句法信息的正确与否会极大地影响零指代项识别的性能。在一定程度上，双语信息的引入降低了自动句法分析错误对该任务的影响。

表 5 给出使用自动句法树时，在已知零指代项(标准零指代项)和自动获取零指代项的情况下，零指代消解的性能。从结果可以看到，不论是基准系统还是改进后的系统，中文零指代消解的性能都极大的依赖于零指代项识别的性能。此外，相比基准系统，在使用标准零指代项时，新引入的双语信息能提升零指代消解的 *F* 值约 2%。在使用自动零指代项情况下，由于改进系统在零指代项识别上的性能优于基准系统，整个中文零指代系统的 *F* 值提升了 5%。

5　结束语

本文构建了中文零指代研究基准平台，并在此基础上融入双语信息来进一步提升零指代消解的性能。提出基于自动对齐机制的英文对等句识别方案，以及基于简化的语义角色标注框架在英文对等句中识别子句的方法。英文对等句识别方案用于识别中文长句中的逗号和分号，判断是否将该长句进行裁剪断句，即将该长句断成几个短句。我们在对等句中识别子句方法的改进过程中发现，通常情况下单个的子句最多包含一个零指代项(OntoNotes5.0 语料中比例大于 99%)。据此，我们将零指代项识别以动词和介词为中心的方法转换成以 Clause 为中心的识别方法，然后以单个子句为基本单元，融入双语信息，实施零指代项的识别和消解。实验结果表明，本文提出的基于中英文可比较语料库的中文零指代识别和消解方法可以明显提高中文零指代系统的性能。

尽管本文提出的方法在中文零指代的识别和消解方面取得一定进展，但中文零指代的性能仍然有很大的提升空间。本文只考虑了一部分中英文对等句的特征，在接下来的工作中，我们将进一步挖掘中英文可比较语料之间的相互关联性，将中英文词

表 4　中文零指代项识别性能

Table 4　Performance of Chinese zero pronoun detection　%

系统	使用标准句法树			使用自动句法树		
	P	*R*	*F*	*P*	*R*	*F*
基准系统	72.9	58.2	64.7	62.7	39.4	48.4
改进后系统	70.5	62.4	66.2	60.4	50.2	54.8

表 5　使用自动句法树时中文零指代消解的性能

Table 5　Performance of Chinese zero pronoun resolution using automatic parse trees　%

系统	标准零指代项			自动零指代项		
	P	*R*	*F*	*P*	*R*	*F*
基准系统	44.8	44.8	44.8	20.3	15.8	17.8
改进后系统	46.7	46.7	46.7	25.4	20.7	22.8

对齐等特征更深入地融入中文零指代系统。

参考文献

[1] Weischedel R, Palmer M, Marcus M, et al. Ontonotes release 5.0 LDC2013T19 [EB/OL]. Philadelphia: Linguistic Data Consortium. (2013−10−16) [2015−03−23]. https://catalog.ldc.upenn.edu/LDC2013T19

[2] Kim Y J. Subject/object drop in the acquisition of Korean: a cross-linguistic comparison. Journal of East Asian Linguistics, 2000, 9(4): 325−351

[3] Zhao S H, Ng T H. Identification and resolution of Chinese zero pronouns: a machine learning approach // Proceedings of EMNLP-2007. Prague: Association for Computational Linguistics, 2007: 541−550

[4] Kong Fang, Zhou Guodong. A tree kernel-based unified framework for Chinese zero anaphora resolution // Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Massachusetts: Association for Computational Linguistics, 2010: 882–891

[5] Chen C, Ng V. Chinese zero pronoun resolution: Some recent advances // Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle: Association for Computational Linguistics, 2013: 1360–1365

[6] Chen C, Ng V. Chinese overt pronoun resolution: a bilingual approach // Proceedings of the 28th AAAI. Conference on Artificial Intelligence. Québec City, 2014: 1615–1621

[7] Chen C, Ng V. Chinese zero pronoun resolution: a joint unsupervised discourse-aware model rivaling state-of-the-art resolvers // Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). Beijing: Association for Computational Linguistics, 2015: 320–326

[8] Chung T, Gildea D. Effects of empty categories on machine translation // Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Massachusetts: Association for Computational Linguistics, 2010: 636–645

[9] Xue Nianwen, Yang Yaqin. Chinese sentence segmentation as comma classification // Proceedings of ACL-2011: Short Papers. Portland, Oregon: Association for Computer Linguistics, 2011: 631−635

[10] Kong Fang, Zhou Guodong. A clause-level hybrid approach to Chinese empty element recovery // Proceedings of the 2013 International Joint Conference on Artificial Intelligence. San Francisco: Morgan Kaufmann, 2013: 2113–2119

[11] Pradhan S, Moschitti A, Xue Nianwen, et al. CoNLL-2012 shared task: modeling multilingual unrestricted coreference in OntoNotes // Proceedings of the Shared Task: Modeling Multilingual Unrestricted Conference in OntoNotes, EMNLP-CoNLL 2012. Jeju Island, 2012: 1−40