统计与规则相结合的指代消解在事件自动文摘中的应用*

刘茂福[1] 金可佳[1] 姬东鸿[2] 张晓龙[1]

[1]武汉科技大学 计算机科学与技术学院 武汉 430065

[2]武汉大学 计算机学院 武汉 430072

E-mail:e_mfliu@163.com

摘要：本文利用基于规则和统计相结合的方法对自动文摘源语料中的代词进行消解。首先使用单纯的规则方法进行消解，通过对召回率和准确率以及消解后的语料进行分析，发现其不足在于不能很好的确定哪些代词指代命名实体。针对这一问题本文将统计中的最大熵方法和规则方法相结合，准确确定哪些代词需要消解，提高消解的准确率和召回率；同时增加语料中命名实体的数量，尽可能多的抽取语料中的事件项来提高基于事件的自动文摘的性能。实验结果表明利用消解后的语料生成的摘要比利用消解前的语料生成的摘要的性能提高了近8.4%，而且文摘的可读性、连贯性以及信息量也有明显的提高。

关键词：指代消解；规则；最大熵；命名实体；基于事件的自动文摘

Application of Anaphora Resolution Based on Statistics and Rules in Event-Based Automatic Summarization

Liu Maofu[1], Jin Kejia[1], Ji Donghong[2], Zhang Xiaolong[1]

[1]College of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430065

[2]School of Computer, Wuhan University, Wuhan 430072

E-mail:e_mfliu@163.com

Abstract: This paper uses the combination of rule-based and statistic-based anaphora resolution in event-based summarization. We find the problem that the rule-based anaphora resolution method can not locate precisely the pronoun indicating name entity by experiments. So we combine the rule-based and maximum entropy to solve this problem and confirm which pronouns should be replaced by name entities. The experiment results show that the anaphora resolution based on statistics and rules can make about 8.4% improvement comparing to the method without anaphora resolution. On the other hand, the readability, fluency and information of the summary have also improved.

Keywords: Anaphora resolution, Rule, Maximum entropy, Name entity, Event-based summarization

1 引言

近年来，越来越多的研究者用事件代表概念并提出了基于事件的自动文摘方法[1]。在基于事件的自动文摘中，一致认为事件应包含一个或多个参与者、发生事件和地点，因此，一般将事件在句子级别上形式化为"[Who] did [What] to [Whom] [When] and [Where]"[2]；其中，"did [What]"指示行为，是事件的核心要素，称为事件项，可以完全或部分标识事件的发生，它一般被定义为位于两个命名实体间的动词或具有动词意义的名词。从源语料中抽取出事件项后，可以利用外部语义资源的语义信息得到事件项语义关系图，进而计算事件项和句子的重要性[3]。基于事件项语

*本文承湖北省教育厅科学技术研究计划(项目号 500064)和国家自然科学基金重大研究计划(项目号 90820005)的资助。

义关系的这种自动文摘方法得到的文摘质量很大程度上取决于语义资源和抽取出的事件项数量。

通过分析 DUC 自动文摘源语料，发现其中包含了大量的代词，按照目前的事件形式，位于指代命名实体的两个代词或代词与命名实体间的事件项是无法抽取出来的。但若能从源语料中尽可能多的抽取出事件项，就可以更充分的利用语料中的语义信息来生成语义关系图，进而提高生成摘要的质量。因此，本文在基于事件项语义关系的自动文摘中引入了指代消解，重点消解那些指代命名实体的代词，在获取更多事件项语义信息的同时，也可以提高最终生成的摘要的可读性。

指代消解实际上是建立概念关联的过程，是文本处理的核心问题之一。与计算语言学的大多数问题一样，指代消解的实现技术主要分为规则方法和统计方法两类[4]。规则方法是从句法和语法层面提出的消解方法[5,6]，统计方法则是从语料库方面进行的消解方法[7,8]，包括最大熵[9]、决策树[10]和聚类等方法。

将统计与规则相结合的方法一直研究的比较少，2006 年王智强[11]等人提出的基于决策树的方法采用了规则和统计相结合的算法，利用规则的方法，先把单复数和性别不一致的情况过滤掉，然后再利用决策树的方法，来确定共指关系，其准确率达到 89.69%；若不利用规则进行过滤，直接应用决策树其准确率为 88.04%。相比之下，先用规则的方法过滤使准确率有很大的提高。本文将最大熵方法和规则方法相结合对给定语料进行消解，先利用规则对第一人称代词和第三人称指人的代词进行消解，再利用最大熵方法对其余的代词进行消解。在统计与规则相结合的指代消解中，特征库和规则库的确定至关重要，本文所使用的特征和规则是从英语的语法角度出发，针对指代命名实体的代词消解这一特定任务来确定的。

2 统计与规则相结合的指代消解方法

2.1 指代消解系统模型

基于统计与规则相结合的指代消解可以分为预处理、代词过滤、统计与规则相结合的指代消解以及实验评估四个步骤，具体的指代消解系统模型如图 1 所示。

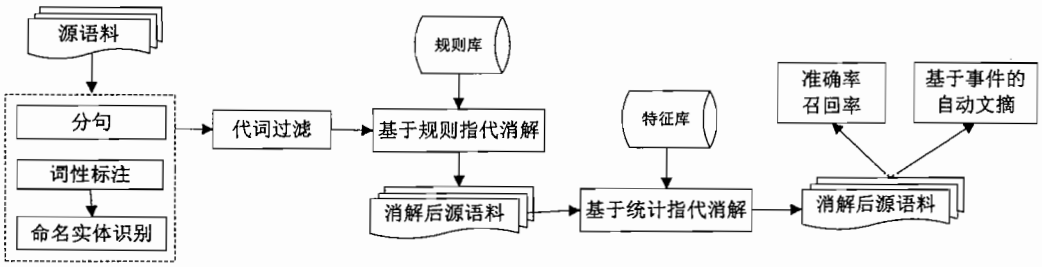

图 1 指代消解系统模型图

本文将 2001 年的 DUC 测试语料作为处理对象，使用 GATE 工具完成句子分割、词性标注和命名实体识别等预处理阶段的工作。由于是对指代命名实体的代词进行消解，为了保证消解的准确率和召回率，将消解分为两个阶段进行，第一个阶段利用规则库中的规则对语料中的代词进行消解，第二个阶段利用最大熵方法进行消解。因此，确定哪些代词使用规则方法进行消解，哪些代词使用统计方法进行消解是非常重要的。消解后的语料一方面被用来计算准确率和召回率；另一方面作为基于事件自动文摘的源语料来测试指代消解对其性能是否有提高。

2.2 代词过滤

本文的指代消解任务是消解那些代指命名实体的代词，对于那些指代一般名词的代词不予消解，所以进行消解前首先需要确定消解对象。通过对DUC语料的分析，我们发现如下规律。

（1）语料中大多数的第二人称代词都是表示读者或者没有实际所指对象。

（2）“it”和“its”在文中多指该文所讲到的中心词或文章中的句子，而大多数文章的中心词和句子都不是命名实体；“it”在文中有一少部分是用于强调句或是英语中的固定用法；只有极少数指代文中的命名实体。

（3）文中的指示代词多指文中叙述的一句话、观点或是普通名词，指代命名实体的情况很少。

基于以上规律，本文对第一人称代词和第三人称指人代词采用基于规则的消解方法；指示代词和第三人称指物代词的一部分采用最大熵方法进行消解，其他的不指代命名实体的代词不进行消解；第二人称代词不消解。

2.3 规则集确定

进行指代消解的另一个关键问题是规则集的确定，针对表示命名实体的代词消解这一特定任务，从英语的句法和语法角度得到了五个规则，即单复数一致、性别一致、语义类一致、句法搭配和距离属性。

规则1：若代词表示单数，则找到表示人、机构、地点、距离、时间、长度等表示单一概念的名词；若代词表示复数，则找到“and”连接的并列结构或其他表示复数概念的命名实体。

规则2：若代词为“he”、“his”等表示男性的代词，则找到命名实体的标记为“<Person>”并且性别属性为“male”的词进行消解；对代词“her”、“she”等进行类似处理。

规则3：若代词表示人，则匹配标记为“<Person>”的命名实体；若代词表示物，则匹配标记为非“<Person>”的命名实体。

规则4：将命名实体的搜索范围限制在该代词所在句子及前两句范围内，即距离值小于3。

规则5：若代词为第一人称代词，则在该句范围内搜索“<Person>+<VB>”的形式，若能找到，则用“<Person>”替换该代词，若找不到，则不予消解。

2.4 特征集确定

在利用规则进行消解时，主要针对第三人称指人代词和第一人称代词进行消解，其余的代词利用最大熵方法进行第二阶段的消解。在最大熵模型中，信息用特征的形式表达，一个特征是一个二值函数，它将事件空间中的事件映射到{0，1}。本文所使用的特征集包括语义类一致性、句子结构、指示词一致性、同位语一致性、距离属性共五个特征。

特征1：语义类一致性，比较候选集中的先行词和待消解代词的语义类，若属于同一语义类，则其特征函数值为1，否则其特征函数值为0。例如：

(1) A coalition of members of <AN>congress</AN> <VB>announced</VB> <Date>Wednesday </Date> that <PRP>they</PRP> <VB>plan</VB> to <VB>sue</VB> the <Organization>Census Bureau</Organization> in an <AN>effort</AN> to <VB>force</VB> the <AN>agency</AN> to <VB>delete</VB> illegal aliens from <PRP$>its</PRP$> <AN>count</AN> in <Date>1990</Date>.

这是DUC源语料在利用规则进行消解之后所得到的句子，在这个句子中可以提取先行词和待消解代词对有：(congress，they)、(Wednesday，they)、(congress，its)、(Wednesday，its)、(Census

Bureau，its)、(effort，its)、(agency，its)，由这些待消解对组成候选集的一部分，其中，“they”所表示的语义类可以是“Person”、“AN”、“Organization”，而“congress”的语义类为“AN”，所以待消解对(congress，they)表示的语义类相同，其语义类特征函数值为1，而待消解对(Wednesday，they)的语义类不同，其语义类特征函数值为0。

特征2：句子结构，主要用于判断代词“it”是指代具体实体，还是用于句子的固定搭配，将代词“it”所在的句子与所给的句子模型匹配，若该句子的结构和句子模型中的任意一个匹配成功，则属于英语固定句型，其特征函数值为0，否则其特征函数值为1。

这里的固定句型主要指强调句和宾语从句中出现“it”的情况，因为在这些情况中，代词“it”只是作为固定句型出现，并没有真正指代具体成分，是不需要消解的。例如：

(2) <PRP>It</PRP> is urgent that the same <AN>reassurance</AN> can be given about the lack of effect of BSE on human health," a consultative committee <VB>reported</VB> to the <AN>agriculture</AN> ministry.

(2)中的句型与强调句相匹配，所以该句中的“it”与任何命名实体和名词组成的代消解对的函数值都为0，所以特征函数值等于0。

特征3：指示词一致性，这里的指示词是指“this”和“that”等词，若句子以指示代词加名词或名词短语位于句首，则候选集中的命名实体与指示代词的特征函数值为0，否则其特征函数值为1。例如：

(3) <DT>this</DT> month, the <AN>government</AN> <VB>announced</VB> <PRP> it </PRP> would <VB>pay</VB> farmers 100 percent of <AN>market</AN> value or average <AN>market </AN> price, whichever is less, for each animal <VB>diagnosed</VB> with BSE.

(3)中指示代词“this”后面是“month”，符合上述特征中所描述的指示代词加名词这一要求，所以待消解对(government，this)的这一函数值为0。

如果“this”后面没有任何名词，而是动词的形式，例如：

(4) <VB>Look</VB> at <DT>this</DT>，the <AN>government</AN> <VB>is</VB> at the front of <PRP>us</PRP>.

(4)中“this”后面没有名词，所以待消解对(government，this)的这一函数值为1，由此可以看出，相同的待消解对，在语言环境不同时，其特征函数值也会不同。

特征4：同位语一致性，指若候选集中的先行词和待消解的代词是同位语关系，则其特征函数值为1，否则其特征函数值为0。例如：

(5) <PRP>We</PRP>，<PERSON>John</PRESON>，<PERSON>Mary</PERSON> and <PRP>I</PRP> <VB>put</VB> forward a <AN>proposal</AN>.

(5)中“We”和“John、Mary、I”是同位语的关系，所以这些候选集中的代消解对的函数值为1，而“I”和“John、Mary”组成的代消解对的同位语一致性的函数值为0。

特征5：距离属性，这作为规则集的一部分已经定义了，但在特征集中重复使用这个属性，是因为距离对于消解来说起很重要的作用。另外，在特征中定义的距离属性是个多值函数，代词和待消解的命名实体或名词位于同一行时，其特征值为3，若它们之间相差一句，则特征是为2，若相差两句，则其特征值为1，其他情况其特征值为0。

在计算了候选集中的代消解对的所有特征函数值之后，根据最后计算出来的特征值并按照最大熵模型计算决定是否进行消解。这次消解主要针对利用规则的方法没有消解的那部分代词，

通过规则和最大熵方法的结合，使语料中可以消解的代词大都进行了消解。

3 实验与结果分析

从 DUC 源语料中随机抽取了 30 篇文章，对其中的代词进行人工消解，然后和本文方法所消解后的结果进行比较，得出本文方法的准确率(1)和召回率(2)如表 1 所示。

$$\text{准确率}=\frac{\text{正确消解的代词数目}}{\text{欲消解的代词数目}} \quad (1) \qquad \text{召回率}=\frac{\text{正确消解的代词数目}}{\text{系统识别的代词数目}} \quad (2)$$

表 1 准确率和召回率

消解方法	代词总数	待消解数	正确消解数	错误消解数	准确率	召回率
规则	425 个	328 个	309 个	19 个	94.21%	72.71%
规则与统计	425 个	352 个	334 个	18 个	94.89%	78.59%

从表 1 数据可以看出，统计与规则相结合的方法比单纯规则方法的召回率有较大的提高。

利用统计与规则相结合方法所生成的语料进行基于事件的自动文摘，并通过 ROUGE 对所生成的摘要进行了评测，具体评测结果如表 2 所示。

表 2 消解前后自动文摘性能比较

LOCAL+OTAC+SUM	消解前	规则	统计与规则	GLOBAL+OTAC+SUM	消解前	规则	统计与规则
ROUGE-1	0.27341	0.28864	0.29132	ROUGE-1	0.24140	0.25719	0.25722
ROUGE-2	0.05335	0.05038	0.05129	ROUGE-2	0.04009	0.03968	0.03968
ROUGE-W	0.09596	0.10121	0.10123	ROUGE-W	0.08118	0.08805	0.08705
LOCAL+OCAT+SUM	消解前	规则	统计与规则	GLOBAL+OCAT+SUM	消解前	规则	统计与规则
ROUGE-1	0.31966	0.32231	0.33542	ROUGE-1	0.30538	0.31722	0.32175
ROUGE-2	0.05750	0.05257	0.05536	ROUGE-2	0.04893	0.05260	0.05306
ROUGE-W	0.11127	0.11327	0.11328	ROUGE-W	0.10629	0.11082	0.12052
LOCAL+OTAC+MAX	消解前	规则	统计与规则	GLOBAL+OTAC+MAX	消解前	规则	统计与规则
ROUGE-1	0.24463	0.25920	0.26351	ROUGE-1	0.23138	0.26175	0.27514
ROUGE-2	0.04414	0.04074	0.04123	ROUGE-2	0.03242	0.04374	0.04544
ROUGE-W	0.08351	0.08894	0.08994	ROUGE-W	0.07776	0.08979	0.08699
LOCAL+OCAT+MAX	消解前	规则	统计与规则	GLOBAL+OCAT+MAX	消解前	规则	统计与规则
ROUGE-1	0.31146	0.32070	0.33525	ROUGE-1	0.30805	0.32468	0.33869
ROUGE-2	0.05404	0.05110	0.05235	ROUGE-2	0.05029	0.05402	0.05402
ROUGE-W	0.10889	0.11243	0.11301	ROUGE-W	0.10786	0.11570	0.11570

表 2 中“LOCAL”是在事件项类中计算事件项的重要性，“GLOBAL”是在所有的事件项类中计算事件项的重要性；“OTAC”是从每个事件项类中抽取该类中最重要的事件项，“OCAT”

是选取最重要的事件项类中的所有事件项；“SUM”是利用句子中出现的所有事件项的值的和作为句子的重要性的值，“MAX”是利用句子中最重要的事件项的值作为句子的重要性的值。

从表 2 可以看出，利用消解后语料比利用消解前语料生成的自动文摘性能平均提高 8.42%，在 OCAT 中，消解前后生成自动文摘的性能都很高，利用消解后的语料比利用消解前的语料生成的自动文摘性能都有所提高，其中“GLOBAL+OCAT+MAX”使自动文摘的性能提高了近 9.95%；在 OTAC 中，相对来说，生成自动文摘的性能要低一些，但利用消解后的语料比利用消解前的语料有较大的提高，其中“GLOBAL+OTAC+MAX”方法使自动文摘的性能提高了 18.68%

另外通过人工比较利用消解前的语料和消解后的语料生成的摘要可以发现，用消解后的语料生成的摘要在可读性、连贯性和信息量方面都有所提高。

4 结论

本文主要介绍了利用统计与规则相结合的指代消解方法将 DUC 测试语料中指代命名实体的代词消解为它所指代的命名实体，并将消解后的语料作为自动文摘的源语料，从实验结果看，指代消解的准确率达到了 94.89%，召回率达到了 78.59%；从 ROUGE 的评价来看，利用消解后语料生成的摘要比利用消解前的语料生成的摘要有很大的提高，说明本文所述的方法具有可行性。

下一步的工作将会考虑如何消解指代一般名词的代词这一问题。找到合适的方法消解本文中未能消解的代词，在保证方法的运算速度的同时进一步提高准确率和召回率。

参 考 文 献

[1] Elena Filatova and Vasileios Hatzivassiloglou. Event-based Extractive Summarization. In Proceedings of ACL 2004 Workshop on Summarization, 2004, 104-111.

[2] Wenjie Li, Wei Xu, Mingli Wu, et al. Extractive Summarization using Inter- and Intra- Event Relevance. In Proceedings of ACL 2006, 369-376.

[3] Maofu Liu, Wenjie Li, Mingli Wu. Extrative Sumarization Base on Event Term Clustering. In Proceedings of the ACL 2007, 2007: 185–188.

[4] 王厚峰. 指代消解的基本方法和实现技术. 中文信息学报,2002, 16(6): 9-17.

[5] Shalom Lappin, Herbert J. Leass. An Algorithm for Pronominal Anaphora Resolution. Computational Linguistics, 1994, 20(4): 535-561.

[6] Christopher Kennedy, Branimir Boguraev. Anaphora for Everyone Pronominal Anaphora Resolution Without a Parser. In Proceedings of the 16th International Conference on Computational Linguistics, 1996: 113-118.

[7] Joseph F. McCarthy, Wendy G. Lehnert. Using Decision Trees for Coreference Resolution. In Proceedings of the 14th International Joint Conference on Artificial Intelligence, 1995:1050-1055.

[8] Niyu Ge, John Hale, Eugene Charniak. A Statistical Approach to Anaphora Resolution. In Proceedings of the 6th Workshop on Very Large Corpora, 1998: 161 -170.

[9] 钱伟, 郭以昆, 周雅倩等. 基于最大熵模型的英文名词短语指代消解. 计算机研究与发展, 2003, 40(9): 1337-1343.

[10] 郎君, 刘挺, 秦兵. 基于决策树的中文名词短语指代消解. 第二届全国学生计算语言学研讨会论文集, 2004:155-157.

[11] 王智强, 李蕾, 王枞. 基于决策树的汉语代词共指消解. 北京邮电大学报, 2006,29(4): 1-5.