基于条件随机场的冠词选择研究

宁 伟，蔡东风，季 铎

（沈阳航空工业学院 知识工程中心，辽宁 沈阳 110034）

Email：ningwei3000@yahoo.com.cn

摘要：冠词选择需要综合考虑语言知识、语义知识以及世界知识，是汉英翻译中的一个难点。针对传统的基于规则和机器学习的方法只考虑名词短语前冠词选择的问题，本文将冠词看作一种标记，将该问题形式化的描述为一个序列标注任务，提出一种基于条件随机场的解决策略，选取特征时充分利用词、词性等多层次资源，并引入前后词的互信息。实验采用包含 91106 个冠词的专利摘要做测试语料，F 值达到 80%。

关键词：冠词选择；条件随机场；序列标注；互信息

Research of Article choice Based on Conditional Random Fields

NING Wei, CAI Dong-feng, JI Duo

(Knowledge Engineering Center, Shenyang Institute of Aeronautical Engineering, Shenyang, Liaoning 110034)

Abstract: Article choice poses a difficult problem for Chinese to English translation for involving complex knowledge about grammar, semantics and the world. Traditional researches based on rule or machine learning only deal with articles used in the noun phrases. This paper considers article as a label, describes the problem as a sequence labeling task and investigates a strategy based on Conditional Random Fields. In the process of feature extraction we make good use of the word and part of speech of the word, especially the mutual information feature. Experimental results on testing corpus composed of patent abstracts containing 91106 articles show that the algorithm yields F-score of 80%.

Key words: article choice; Conditional Random Fields; sequence labeling; mutual information

1 引言

在翻译工作中，译文审校是一个关键环节，是译文质量的保证。翻译人员所犯的错误种类繁多，翻译公司一般将这些错误分为语法错误、语义错误、单词拼写错误及冠词使用错误四类。本文对专利摘要翻译项目中 300,000 件译文校对前后的错误进行了统计分析，得出各类错误所占的比例分别为 36.5%，14.8%，32.8%和 15.8%。由此可见，冠词错误是影响翻译人员译文质量的一个关键因素。冠词错误又可分为缺少冠词错误，多加冠词错误和误用冠词错误三种类型，其中缺少冠词错误所占比例最大，达到 45.6%。所以冠词选择是目前翻译审校的一个重要考察点。

冠词是在英语中普遍使用的一种限定词[1]，只能在名词或名词短语之前使用，并且不同的冠词可使得名词含义不同，例如 out of question 和 out of the question，前者“question”不加冠词表示“毫无疑问”的意思，而后者“question”加定冠词“the”表示的则是“不可能，办不到”的意思，两个短语唯一的区别就是使用了不同的冠词，然而表达的意思却完全不同。因此，冠词的使用可以使名词含义明确。

在汉英翻译中冠词选择问题难以解决的原因，一方面是汉语中没有相应的语言范畴，另一方面是冠词在英语中的使用需要综合考虑复杂的语言知识、语义知识以及世界知识，

基金项目：国家 863 计划课题(the National High-Tech Research and Development 863 plan of China under Grant No. 2006AA01Z148)

作者简介：宁伟(1983-)，男，研究生，主要研究方向为自然语言处理、译文质量自动评测；蔡东风(1958-)，男，博士，教授，主要研究方向为人工智能、自然语言处理。

用法规则难以形式化[4]；还有就是冠词的用法相对灵活而难以掌握[2]，例如：(1)A talk will be given on Friday about NLP; The talk will last for one hour.和(2)Friday's NLP talk will last for one hour;这两句话表达相同的意思，然而(2)把(1)中两个句子组合而导致名词短语形式发生改变，其相应的冠词也跟着发生改变。

有关英语冠词的使用，许多英语语法论著都总结了详细的用法规则来帮助人们学习和使用冠词，但要使计算机能够理解这些规则，表示和编码存在一定的困难，针对人的冠词用法规则显然不适于计算机的处理。对于翻译人员的冠词选择错误，现在普遍采用的策略还是人工校对，但是人工校对的成本相对较高，而且效率低，不可复用。既使在机器翻译中，目前也没有一种统一有效的冠词自动选择策略，冠词选择问题没有得到足够重视。本文提出一种基于条件随机场的冠词自动选择方法，从词汇，语法方面选取特征，并且引入前后词的互信息来训练模型，实验结果表明本文的方法可以有效的提高冠词选择的正确率，F 值达到 80%。

本文组织结构如下，第二节为相关研究工作的概述；第三节介绍基于条件随机场的冠词选择，包括条件随机场的相关介绍和实验所选特征；第四节介绍本文的实验以及实验结果分析；最后是结论和展望。

2 相关研究工作

冠词选择是机器翻译译文生成时需要解决的问题之一，早期针对机器翻译的冠词选择研究(Murata and Nagao[14], 1993; Bond and Ogura[4], 1998; Herine[5], 1998)普遍采用的是基于规则的思想，其关键在于人工制定规则，再综合考虑词汇信息来选择合适的冠词，然而人工制定规则库需要计算机专家和语言学家的共同努力，是一项耗时耗力的任务。刘群等人提出一种基于转换的错误驱动的学习(transformation-based Error-Driven Learning)策略[3]用于机器翻译的冠词选择。

与基于规则的思想不同，Kinght 和 Chander(1994)提出了一种基于决策树的学习方法来选择冠词[6]，他们对 1600 个常用名词构建决策树，以从华尔街时报中获取的 400,000 个名词短语作为训练实例，提取 30，000 个词汇，语法，句法方面的特征，实验中名词短语的冠词选择正确率为 78%。另外 Guido Minnen 等人提出一种基于记忆学习的方法来研究名词短语的冠词选择问题[2]；John Lee 根据 WordNet 和句法分析的特征，采用对数线性模型恢复遗漏冠词[7]。

这些研究的一个共同点是只针对名词短语进行冠词选择判断，而未考虑冠词使用的其它情况，例如 the more the better 和 the stronger the china is...,显然 more 和 stronger 前面的冠词都不是限定名词短语的。而且在实际应用中冠词选择问题是针对一篇或者一段完整的文章，所以这些研究普遍采用的策略是先对文章进行句法分析，从中提取出名词短语然后再针对名词短语进行冠词选择。然而文献[13]在比较中文和英文句法分析时，指出目前英语句法分析的性能只能达到 90%的水平[12]，而且对于缺少或者存在错误冠词的句子进行句法分析更容易出现错误，导致不能将文中出现的所有名词短语都抽取出来，对于没有抽取出来的名词短语就不能进行冠词选择判断。本文提出的冠词选择策略是针对译文中的所有词按所加冠词的不同进行分类，一方面可以避免名词短语抽取错误，另一方面可以全面考虑冠词选择的各种情况，冠词选择正确率达到 80.4%，F 值为 80%。

3 基于条件随机场的冠词选择

3.1 问题描述

冠词是一种限定词，可以把它看作所限定词的一个标记，本文将冠词选择任务形式化

的描述为词序列标注问题，每个词可以选择的标记包括 a、an、the 以及不加冠词，其中冠词 a 和 an 的选择可以采用规则的策略解决，为降低标记冗余性对 CRF 标注正确率的影响，把这两个标记看作是同一个标记，然后再对标注结果进行后处理确定 a/an。词的标记集记为 T={A，the，NO}，其中 A={a,an}，NO 表示不加冠词。本文采用条件随机场解决冠词选择问题，输入 n 个词组成的句子 $S=w_1,w_2,\ldots\ldots,w_n$,目标是给出一个相应的标注序列 $T^*=t_1,t_2,\ldots\ldots t_n$，且满足

$$T^* = \arg\max_T p(T \mid S) \qquad (1)$$

3.2 特征选择

在基于 CRF 的标注分类问题中，特征函数的选择通常起着关键性作用，特征选择的好坏直接决定着 CRF 标注结果的优劣。本文主要采用词的上下文作为特征，根据 Knight 的实验结果[6]，95%左右的冠词可以由其所处的上下文预测出来。

同时引入前后词语的互信息作为特征，实验证明这是一种很有效的特征。在研究自然语言词汇的搭配问题中，经常使用互信息作为描述两个单词之间关联程度大小的量度，本文的互信息指的是点互信息，计算公式为：

$$I(x,y) = \log \frac{p(x,y)}{p(x)p(y)} \qquad (2)$$

其中 p(x)和 p(y)表示词 x，y 各自出现的概率，p(x,y)表示二元搭配(x,y)出现的概率。从训练语料中统计得到各二元词对互信息的值，该值的大小能很好的反映两个词之间的关联程度，但是在 CRF 训练时，它把各种特征都是当作字符串来理解的，不能将值的大小信息体现出来，所以本文将两个词之间的紧密程度定性的分为两种，即紧密与不紧密，该特征被描述为一个二值属性。首先从训练语料中统计得到一个阈值，然后比较测试语料中词对互信息值与该设定阈值的大小，如果它与前一个词的互信息大于该阈值时，特征取 1，否则特征取 0。

本文是对每个词进行标记选择，每个词的标记选择过程都被看作一个事件，由当前词及其上下文环境来确定一个事件的特征集合，根据影响冠词选择的各种因素，本文所选原子特征如表 1：

表 1 原子特征列表

特征	特征描述
1	当前词(WORD)
2	当前词词性(POS)
3	当前词前面的N个词W_{-i}，以及当前词后面的N个词Wi
4	当前词前面的N个词的词性P_{-i}，以及当前词后面的N个词的词性Pi
5	当前词的单复数(FORM)，单数为S，复数为P，非名词为!NOUN
6	当前词在前面是否出现过(OCCUR)，出现过为1，否则为0
7	当前词和前面一个词的互信息特征(MUL)

(其中 i=1,2……,N，N 表示上下文词窗口大小)

4 实验及实验结果分析

4.1 实验语料

本文的实验语料是国家专利局英文专利摘要，采用专利语料是因为在专利语料中使用了大量的冠词，统计发现大约 150 个词的摘要一般包含 15-20 个冠词。实验采用包含 132,769 个冠词

的 8000 个英文专利摘要作训练语料，用包含 91,106 个冠词的 5100 个英文专利摘要做测试。其详细数据如表 2 所示。

表 2 语料中所有词、名词以及冠词个数统计表

	所有词	名词	the	a	an
训练语料	931, 844	294, 187	72, 601	49, 555	10, 613
测试语料	625, 700	199, 932	49, 657	33, 557	7, 262

4.2 评价方法

本文采用准确率，召回率以及 F 值作为系统性能的评价指标，这是在分类问题中普遍采用的三种评价指标。给定数据集合包括类 C={A, the, NO}，正确率，召回率和 F 值的定义如下：

$$P_{Ci} = \frac{N_{Ci}}{M_{Ci}} \times 100\% \quad (3)$$

$$R_{Ci} = \frac{N_{Ci}}{T_{Ci}} \times 100\% \quad (4)$$

$$F_{Ci} = \frac{P_{Ci} \times R_{Ci} \times (\beta^2 + 1)}{R_{Ci} + \beta^2 P_{Ci}} \times 100\% \quad (5)$$

C_i 表示第 i 个类别，N_{Ci} 表示 CRF 标注正确的冠词 C_i 的个数，MCi 表示 CRF 标注出的冠词 C_i 的个数，T_i 表示测试语料中存在的冠词 C_i 的个数，β 是 P 和 R 重要性的加权系数，β 取 1，即把准确率和召回率同等看待。

4.3 实验设置

本文实验中主要采用词的上下文作为特征，所以先设定一个词窗口来确定上下文的范围，词窗口的大小对实验结果有很大影响。分别取当前词前后各 4，3，2 个词做实验，反复实验后，发现前后各取 3 个词时系统的测试性能相对最优。实验中首先将所有的冠词去掉，然后进行词性标注，按表 1 选定的原子特征，设计如表 3 所示的特征模板，其中复合特征是在分析了原子特征模板对实验结果的影响程度的基础上设计的，通过复合特征可以弥补原子特征在表示上下文环境上的不足，更有效用简单的特征和特征组合表示复杂的语言现象。根据选定的特征模板进行模型训练，然后将该模型用于测试语料的标记。

表 3 实验中采用的特征模板

模 板	原子特征					复合特征		
	Wi	Pi	FORM	MUL	OCCUR	MUL /OCCUR	W-1/WORD/ W+1	P-1/POS/ P+1
T0	O							
T1	O	O						
T2	O	O	O					
T3	O	O		O				
T4	O	O			O			
T5	O	O	O		O			
T6	O	O		O	O			
T7	O	O	O	O	O	O		
T8	O	O	O	O	O	O	O	
T9	O	O	O	O	O		O	O

4.4 实验结果及分析

本文分别采用最大熵模型和 CRF 模型做了对比实验，具体实验结果如表 5 所示。

表 5 最大熵和 CRF 标注结果

模板	最大熵(%)					CRF(%)				
	F_a	F_{the}	F_{NO}	F_{a+the}	$F_{a+the+NO}$	F_a	F_{the}	F_{NO}	F_{a+the}	$F_{a+the+NO}$
T0	58.2	59.6	96.1	58.5	90.5	63.2	65.1	96.9	64.3	91.7
T1	63.4	63.5	96.8	63.4	91.5	64.4	66.6	97.1	65.7	91.9
T2	64.2	64.2	97.0	64.2	91.7	65.3	66.7	97.2	66.1	92.0
T3	66.7	69.0	98.1	67.9	93.2	71.5	74.8	97.2	73.2	93.3
T4	70.4	72.5	97.2	71.6	93.0	72.0	74.5	97.3	73.3	93.3
T5	70.6	72.9	97.2	71.9	93.2	75.9	79.3	98.4	77.8	94.9
T6	71.0	76.7	97.8	74.1	93.4	76.7	80.1	98.5	78.6	95.2
T7	71.2	77.1	97.8	74.2	93.5	77.1	80.6	98.5	79.0	95.3
T8	71.4	77.2	97.8	74.5	93.6	77.2	81.8	98.5	79.2	95.5
T9	71.7	77.8	97.8	75.2	93.8	77.6	84.1	98.6	80.0	95.8

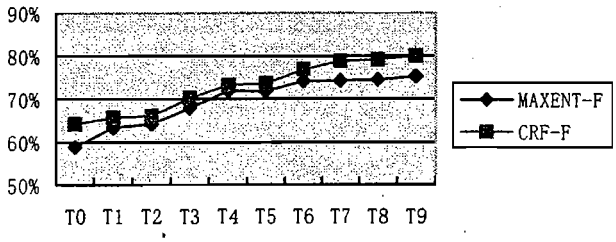

图 1　CRF 和最大熵标注结果 F 值比较

图 1 是 CRF 和最大熵标注结果 F 值比较，从中可以看出在选用同样特征的条件下，CRF 标注性能要优于最大熵的性能，说明 CRF 对特征的融合能力要优于最大熵模型，即使是给出很简单的特征，仍然能达到较高的性能，比如对于特征模板 T0，仅给出词特征，其 F 值更是明显的高于最大熵模型。对比实验中在特征模板 T9 二者都得到最优标注结果，F_{a+the} 值比最大熵高 4.8%。CRF 和最大熵两种模型都可以用于序列标注问题，二者都能够充分考虑上下文的特征，综合利用词、词性等多层次的资源，但 CRF 是将所有标记当作一个连续序列来处理，而最大熵将整个序列看作是一些离散点，各个标记之间是孤立的，这样就不能充分利用标记之间的相互影响关系，而且 CRF 对于长程关联(long distance dependency)有很好的描述能力。另外最大熵模型在实现时在每一结点都要进行归一化，所以只能找到局部的最优值，同时也带来了标记偏置(label bias)问题，即凡是语料中未出现的情况全部忽略掉，而 CRF 很好的解决了这一问题，它并不在每一结点上都进行归一化，而是对所有的特征进行全局归一化，因此可以得到全局最优值，性能优于最大熵模型。

本文实验中将当前词的前后三个词作为一个基本特征，然后不断的加入其它特征来做实验，其中模板 T2-T4 分别在 T1 的基础上引入一个新的特征得到。可以看出，对于两种模型，F_{NO} 值变化不明显，一方面是因为在语料中有大约 85%以上的词都是不加冠词的，CRF 已经很好的拟合了这种特征，另一方面针对冠词选择任务引入的特征对于不加冠词标记基本上是没有信息增益的。

同时可以看出，不同特征的引入对冠词选择结果的影响是不同的，有些特征的引入并不能使实验结果产生太大的改进，而“当前词是否出现过”和“互信息”特征的引入使 CRF

和最大熵标注结果的 F_{a+the} 值分别提高 9%，8.9%和 13.1%，9.4%，可见这两个特征对两种模型的训练都有很大的信息增益，原因是这两个特征分别从两个方面体现了冠词使用的规律。按照英语冠词的用法规则，在一篇文章中某个名词首次出现时一般加不定冠词，当该词再次出现时需要使用定冠词 the，所以“当前词是否出现过”对确定当前词是否需要加定冠词有很大的指导作用；“互信息”反映的是两个词的紧密程度，其值越大表明两个词的耦合度越大，当前词加冠词的可能性越小，这在一定程度上反映了是否需要加冠词的概率。所以在模板 T6 中同时引入这两个特征，CRF 和最大熵标注结果的 F_{a+the} 值分别提高 14.3%和 15.6%，可见这两个特征对于冠词选择任务是很有效的。

CRF 对多元复合特征也有较好的融合能力，模板 T7，T8 和 T9 是分别引入不同的复合特征得到的，加入复合特征实验结果也略有提高，其中冠词选择的正确率，召回率和 F 值分别达到 80.4%，79.6%和 80.0%，表明 CRF 可以充分的利用多层次的资源，对于长程关联有很好的描述能力。

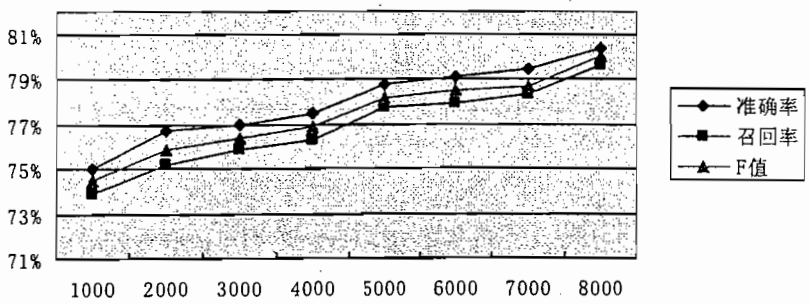

图 2 不同语料规模的实验结果比较

为了考察训练语料对标注结果的影响，本文分别采用不同规模的语料做了实验，结果如图 2 所示，从图中可以看出随着训练语料的增加，系统的准确率，正确率和 F 值都在提高。在现有语料情况下，语料数量与系统性能基本呈线性增长关系，说明数据稀疏还比较严重，随着语料的增加，系统性能还会有一定的提升空间。

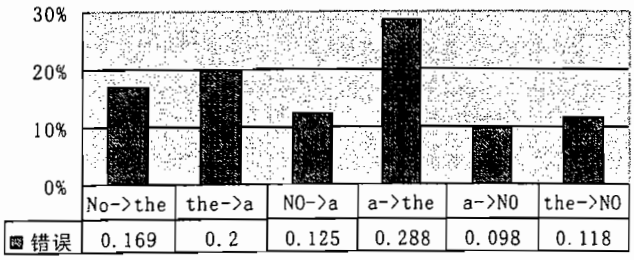

图 3 CRF 标注错误结果分析

对于 CRF 的标注错误分析如图 3 所示，可以看出冠词 a 和 the 误用情况最为严重，所占比例接近 50%，错误的原因分析如下：

(1)CRF 是一种基于统计的思想，所以存在数据稀疏问题。本文实验采用专利语料，所用词汇领域性强，对于训练语料中没有出现的专利领域，测试时标注错误严重。

(2)冠词 a/the 的选择涉及语义知识和人的世界知识，而这种知识很难形式化的表示出来。比如“The utility model relates to a damper door……, the handle of the door……”，这里的 handle 在文中是首次出现而加 the 表示的是一种特指，特指它是前面所述“door”的一部分，但是

这是人根据自己世界知识判断出来的，计算机还不能理解。
(3)词性标注的错误也是不可忽略的。

5 总结和展望

冠词选择对于机器翻译或者是译文质量的评测都有着重要的应用价值，但是冠词选择涉及语法，语义以及世界知识等，处理难度较大，是汉英翻译中一个公认的难点。本文将冠词的选择看作是词的序列标注问题，提出一种基于条件随机场的冠词自动选择策略，并做了与最大熵模型的对比实验。条件随机场模型结合了条件概率模型的优势，并克服了标记偏置问题，实验结果优于最大熵模型，其冠词选择正确率达到 80.4%，F 值为 80%。

本文方法与前人研究最大不同是针对所有词而不只是名词短语进行冠词选择，一方面可以避免名词短语抽取错误，另一方面可以全面考虑冠词使用的各种情况，实验表明这是一种有效的方法。但同时也增加了冠词选择的“噪音”干扰。本文下一步打算继续扩充训练语料的规模，尽量减少数据稀疏的影响，并进行适当的预处理，优化特征的选取，进一步改善实验结果。

参考文献

[1] Biber, Johansson, Leech, Conrad. Longman grammar of Spoken and Written English[M]. London: Longman Press, 1999. 260-263

[2] Guido Minnen, Francis Bond, Ann Copestake. Memory-based Learning for Article Generation[A]. In Proce- edings of the 4th Conference on Computational Language Learning and the 2nd Learning Language in Logic Workshop[C]. Lisbon, Portugal, 2000. 43-48

[3] 常宝宝, 刘颖, 刘群. 汉英机器翻译中的冠词处理策略[J]. 中文信息学报, 1998, 1998年第3期

[4] Francis Bond, Kentaro Ogura. Reference in Japanese-to-English machine translation[J]. Machine Translation. 1998, 13(2-3):107-134

[5] Julia E. Heine. Definiteness predictions for Japanese noun phrases[A]. In 36th Annual Meeting of the Assoc- iation for Computational Linguistics and 17th International Conference on Computational Linguistics:COLI- NG/A CL-98[C]. Montreal, Canada, 1998. 519-525

[6] Kevin Knight, Ishwar Chander. Automated postediting of documents[A]. In Proceedings of the 12th National Conference on Artificial Intelligence[C]. Seattle, Washingtion, United States, 1994. 779-784

[7] John Lee. Automated article restoration. In proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics. 2004

[8] John Lafferty, Andrew McCallum, Fernando Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data[A]. In International Conference on Machine Learning[C]. 2001

[9] Fei Sha, Fernando Pereira. Shallow Parsing with Conditional Random Fields[A]. Proceedings of HLT-Na4-CL[C]. 2003, Edmonton, Canada

[10] Andrew McCallum, Wei Li. Early Results for Named Entity Recognition with Conditional Random Fields Feature Induction and Web-Enhanced Lexicons[A]. Proceedings of the 7th Conference on Natural Language Learning[C]. Edmonton, Canada, 2003

[11] Masaki Murata, Makoto Nagao. Determination of referential property and number of nouns in Japanese sentences for machine translation into English[A]. In Fifth International Conference on Theoretical and Methodological Issues in Machine Translation: TMI-93[C]. Kyoto, July, 1993. 218–225.

[12] 米海涛, 熊德意, 刘群. 中文词法分析与句法分析融合策略研究[J]. 中文信息学报, 2008, 22(2):10-17

[13] Craig G. Nevill-Manning, Ian H. Witten. Identifying Hierarchical Structure in Sequences: A linear-time algorithm[J]. Journal of Artificial Intelligence Research. Vol 7, 1997. 67-82