汉语词法分析中上文和下文孰重孰轻*

于江德[1]，王希杰[1]，樊孝忠[2]
[1]安阳师范学院 计算机与信息工程学院，河南 安阳 455002
[2]北京理工大学 计算机科学技术学院，北京 100081
E-mail：jiangde_yu@tom.com

摘 要：汉语词法分析是中文信息处理的基础，现阶段汉语词法分析的主流技术是基于统计的方法，这类方法的本质都是把词法分析过程看作序列数据标注问题。上下文是统计语言学中获取语言知识和解决自然语言处理中多种实际应用问题必须依靠的资源和基础。汉语词法分析时需要从上下文获取相关的语言知识，但上文和下文是否同样重要呢？为克服仅凭主观经验给出猜测结果的不足，我们对汉语词法分析的分词、词性标注、命名实体识别这三项子任务进行了深入研究，对比了上文和下文对各个任务性能的影响，在国际汉语语言处理评测 Bakeoff 多种语料上进行了封闭测试，采用分别表征上文和下文的特征模板集进行了对比实验，结果表明，上文和下文对汉语分词和中文命名实体识别性能的影响差别较大，对汉语词性标注的性能影响差别较小。

关键词：汉语词法分析；上下文；分词；词性标注；命名实体识别；特征模板

Which Is More Effective for Chinese Lexical Analysis: Above-context Versus Below-context?

Yu Jiang-de[1], Wang Xi-jie[1], Fan Xiao-zhong[2]
[1] School of Computer and Information Engineering, Anyang Normal University, Anyang 455002
[2] School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081
E-mail：jiangde_yu@tom.com

Abstract: Chinese lexical analysis is a foundational task for Chinese information processing. At the current, the mainstream technology of Chinese lexical analysis is based on statistical methods. These methods treat the analysis process as a sequence data tagging problem. Context is the necessary resource not only for obtaining linguistic knowledge in statistical linguistics but also for solving the problem in natural language processing. Chinese lexical analysis needs the help of correlative context. However, is above and below the same important? To overcome the lack of giving the result by the subjective experience, we studied the contribution of above and below for Chinese lexical analysis via the large number of experiments about word segmentation, POS tagging and named entity recognition. Closed evaluations are performed on many kinds of corpus from the international Chinese language processing Bakeoff, and comparative experiments are performed on different feature templates which describe above-context and below-context. Experimental results show that the performance is very different by the below-context and by the above-context.

Keywords: Chinese lexical analysis; context; word segmentation; POS tagging; named entity recognition; feature template

1 引言

汉语词法分析主要包括汉语自动分词、词性标注和命名实体识别三项子任务，它是中文信息处理领域的一项基础性研究课题[1,2]。现阶段，基于统计的方法是汉语词法分析的主流技术。对于汉语分词任务而言，自从 2002 年 Xue 等[3]在第一届国际计算语言学会下属的汉语语言处理特别兴趣研究组（special interest group on Chinese language processing，SIGHAN）研讨会上提出基于字标注（character-based tagging）的汉语分词技术后，近年来，基于类似思想的汉语分词技术得到了广泛关注和研究[4-6]。对于汉语词性标注任务，由于基于规则的方法适应性较差，所以已有的研究中

*基金项目：由河南省高等学校青年骨干教师项目（2009GGJS-108）和河南省教育厅自然科学研究项目（2011B520004）支持。

基于统计语言模型的方法居多，已采用的统计语言模型主要有 N 元语法模型（N-gram）[7]、隐马尔科夫模型（hidden Markov model, HMM）[2]、最大熵模型（maximum entropy, ME）[8]、条件随机场（conditional random fields，CRFs）[9]、支持向量机（support vector machine, SVM）[10]等。在中文命名实体识别任务上，基于 HMM 的方法[2,11]、最大熵模型的方法、CRFs 的方法[12]以及 SVM 的方法是目前比较常见的方法。综合分析这些已有的汉语词法分析研究，都是将汉语词法分析的本质看作是一个序列数据标注问题，借助于统计语言模型实现。

统计语言建模中，上下文扮演着解决问题所需语言知识和资源提供者的重要角色[13]。通常情况下，上下文的选取是基于当前位置左右一定范围进行的，这个固定的范围称为“窗口”。一般情况下，选取的这个“窗口”是对称的，即上文和下文有相同的宽度。但是，这种对称地选取上下文是否合适呢？这取决于在解决具体问题时从宽度相同的上文和下文中获取的语言知识的多少，或者说上文和下文对具体问题解决的贡献情况。为了克服当前仅凭主观经验对称地选择上下文的不足，本文首先对基于统计语言模型的汉语词法分析中上下文的建模进行了详细解析，然后深入研究了汉语词法分析的三个子任务中上文和下文对性能的影响情况，并在国际汉语语言处理评测 Bakeoff 的分词、词性标注、命名实体识别三种语料上进行了封闭测试，采用表征上文和下文的不同特征模板集进行了对比实验，结果表明，上文和下文对汉语词法分析的影响大不相同，而且在三个不同的子任务中上文和下文对性能的影响差别又各不相同。

2 基于统计的汉语词法分析中上下文的建模

基于统计方法的汉语词法分析的实质是将汉语词法分析转化为序列数据标注问题，该问题可使用统计语言模型之一来实现。例如，N-Gram、HMM、ME、CRFs、SVM 等。在纷繁复杂的自然语言中，蕴含着一些内在的语言规律（某种程度上也可以称为语言知识），对一个具体的自然语言处理任务，这些语言规律是不可或缺的。对汉语词法分析这一具体任务而言，这些语言规律外在的表现形式就是句子中字之间、词之间、字与词之间、词与词性之间、字与命名实体之间、词与命名实体之间所存在的内在关联。统计语言模型是采用概率论与数理统计的方法从训练语料中统计出字、词、命名实体等不同粒度语言单位（linguistic unit）及其相应标记之间的内在联系，即在训练语料中这些不同粒度语言单位之间存在的概率分布规律，也就是训练语料中存在的语言规律。由此可见，利用统计语言模型对训练语料上下文进行建模的本质就是要准确描述和刻画上下文中存在的语言规律，其中，上下文“窗口”的设定和窗口中上下文特征的刻画是两个关键问题。

2.1 上下文“窗口”的设定

通常情况下，上下文的选取是基于当前位置前后一定范围进行的，这个固定的范围称为“窗口”。由于语言规律要从“窗口”的上下文中统计获取，所以上下文“窗口”开设的大小要依据具体的任务来设定。图 1 示意了汉语词法分析中可能的上下文“窗口”。图中方框中示意的上下文窗口是“3 个语言单位窗口”，即设定的上下文窗口包含当前语言单位及其前后各一个语言单位共 3 个。具体到基于字标注的汉语分词而言，此时的语言单位是字，上下文窗口所表现出来的上下文

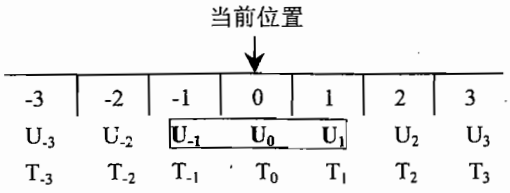

图 1 可能的上下文“窗口”

特征是指由当前字本身以及当前字前后各一个字所构成的特征。当然，上下文窗口也可以是“5语言单位窗口”、“7语言单位窗口”等。一般情况下所设定的上下文窗口都是关于当前位置对称的，即取当前位置前后相同宽度的上文和下文。可是，在汉语词法分析中这样对称地选取上下文窗口是否合适？这将是本文要回答的问题。

2.2 上下文特征的刻画

上下文“窗口”确定之后，又会面临另一个问题：自然语言中的规律是通过“窗口”中一个个“上下文”样本表现出来的，如何将窗口中的上下文特征描述和刻画出来？在统计语言建模中，通常是采用特征模板来描述和刻画上下文中的语言特征。特征模板的主要功能是定义上下文中某些特定位置的语言成分与某类待预测事件的关联情况。习惯上，特征模板可以看作是对一组上下文特征按照共同的属性进行的抽象。在统计语言模型的训练学习中，每个特征都对应了一组特征函数，这些特征函数对该模型的训练学习至关重要。而这些特征又是通过特征模板扩展来的，所以，设定合适的特征模板就至关重要。

对于汉语词法分析而言，综合分析已有汉语分词、词性标注和命名实体识别三个子任务中特征模板的设定，可以发现，特征模板所刻画出的上下文特征主要是指上下文“窗口”中字、词等语言单位之间所组合成的特征。特征模板可以根据语言单位和当前位置的距离属性组合设定，可以是单个语言单位构成的特征模板。例如，U_{-1}，U_0，U_1等，其中U_{-1}表示当前位置前一个语言单位，U_0表示当前语言单位，U_1表示当前位置的后一个语言单位。也可以是窗口内的两个语言单位联合组成的特征模板。例如，$U_{-1}U_0$表示当前位置前一个语言单位和当前位置的语言单位构成的特征模板。依此类推还可以有三个语言单位联合组成的特征模板等。

为了回答汉语词法分析中对称地选取上下文是否合适？就需要研究上文和下文对汉语词法分析性能的影响是否大体相等，或者看谁对汉语词法分析的影响更大。为此，我们需要设定仅仅刻画上文（包括当前位置）和仅仅刻画下文（包括当前位置）的特征模板集，以便进行对比实验。在本文后面的实验中后缀-Above和-Below示意的特征模板集分别表示相应模板集的上文部分和下文部分。

3 汉语词法分析中上文和下文孰重孰轻的对比实验及其结果分析

3.1 实验设计

本文分别针对汉语词法分析的三个子任务：汉语分词、词性标注、命名实体识别进行了上文和下文孰重孰轻的对比实验。其中汉语分词中上文和下文的对比实验采用基于字标注的汉语分词技术，使用四词位标注集（B、M、E、S）和条件随机场来具体实现。汉语词性标注中上文和下文的对比实验采用最大熵模型具体实现。中文命名实体识别中上文和下文的对比实验采用基于字标注的命名实体识别技术，采用条件随机场具体实现。

3.2 汉语分词中上文和下文孰重孰轻

3.2.1 实验数据集和性能评估

本组实验采用的训练语料和测试语料是SIGHAN举办的第二届国际汉语分词评测Bakeoff2005所提供的语料中的简体中文语料，分别是北京大学（PKU）和微软亚洲研究院（MSRA）提供的训练语料和测试语料。

在对汉语分词性能进行评估时，采用了常用的5个评测指标：准确率（*P*）、召回率（*R*）、综合指标*F*值（*F*）、未登录词召回率（*OOV RR*）、词表词召回率（*IV RR*）。准确率表示在切分的全

部词语中，正确的所占的比值。召回率指正确切分的词语占标准答案中词语的比值。综合指标 F 值是综合准确率和召回率两个值进行评价的一种办法。*OOV RR* 和 *IV RR* 分别指测试中未登录词和词表词的召回率。

3.2.2 本组实验使用的特征模板集

为了研究汉语分词中上文和下文究竟谁对分词性能的贡献更大，我们设定了六组特征模板集，这六组特征模板集包含的特征模板见表 1（注：此时特征模板中的主要语言单位是汉字，故用字符 C 表示）。其中，TMPT-10 是在相关工作中最常用的一组特征模板，TMPT-6 是文献[4]中配合 6 词位标注集（B、B2、B3、M、E、S）使用的特征模板集。需要注意的是 TMPT-10 和 TMPT-6 都包含一个特征模板：$T_{-1}T_0$，该模板用于表征上下文中相邻两个字的词位转移特征。

表 1 汉语分词的六组特征模板集

特征模板集名称	包含的特征模板
TMPT-10	C_{-2}，C_{-1}，C_0，C_1，C_2，$C_{-2}C_{-1}$，$C_{-1}C_0$，C_0C_1，C_1C_2，$C_{-1}C_1$，$T_{-1}T_0$
TMPT-6	C_{-1}，C_0，C_1，$C_{-1}C_0$，C_0C_1，$C_{-1}C_1$，$T_{-1}T_0$
T10-Above	C_{-2}，C_{-1}，C_0，$C_{-2}C_{-1}$，$C_{-1}C_0$
T10-Below	C_0，C_1，C_2，C_0C_1，C_1C_2
T6-Above	C_{-1}，C_0，$C_{-1}C_0$
T6-Below	C_0，C_1，C_0C_1

3.2.3 实验结果及其分析

在表 1 中，T10-Above 和 T10-Below 这两组特征模板集分别对应了 TMPT-10 中的上文部分和下文部分，T10-Above 特征模板集仅仅刻画了当前字本身以及当前字前面两个字所组成的特征，而 T10-Below 特征模板集仅仅刻画了当前字本身以及当前字后面两个字所组成的特征。从上下文“窗口”来看分别是原“5 字窗口”的上文“3 字窗口”和下文“3 字窗口”。相应地 T6-Above 和 T6-Below 这两组特征模板集分别对应了 TMPT-6 中的上文部分和下文部分。通过设定这四组特征模板集，然后进行对比实验，实验结果见表 2。从表 2 可以看出，对于使用四词位标注集，采用条件随机场作为词位标注器，使用 T10-Below 特征模板集在 PKU、MSRA 两个语料上的综合指标 F 值比 T10-Above 分别高出了 13.7、14.7 个百分点；使用 T6-Below 特征模板集在 PKU、MSRA 两个语料上的综合指标 F 值比 T6-Above 分别高出了 13.8、14.8 个百分点。这表明，在汉语分词中下文比上文蕴含的有效语言知识要多的多，对汉语分词的性能贡献更大。

表 2 不同特征模板集的分词结果

不同特征模板集	PKU 语料上评测结果					MSRA 语料上评测结果				
	P	*R*	*F*	*OOV RR*	*IV RR*	*P*	*R*	*F*	*OOV RR*	*IV RR*
TMPT-10	0.935	0.923	0.929	0.620	0.941	0.960	0.962	0.961	0.694	0.968
TMPT-6	0.936	0.922	0.929	0.594	0.942	0.964	0.961	0.963	0.686	0.969
T10-Above	0.772	0.763	0.768	0.334	0.789	0.791	0.791	0.791	0.374	0.803
T10-Below	0.912	0.898	0.905	0.483	0.924	0.939	0.938	0.938	0.559	0.948
T6-Above	0.747	0.743	0.745	0.274	0.771	0.766	0.771	0.769	0.317	0.784
T6-Below	0.889	0.878	0.883	0.420	0.906	0.916	0.918	0.917	0.466	0.931

3.3 汉语词性标注中上文和下文孰重孰轻

3.3.1 实验数据集和性能评估

本组实验采用的训练语料和测试语料是 SIGHAN 举办的第四届国际汉语语言处理评测

Bakeoff2007 所提供的汉语词性标注部分的简体中文语料，分别由北京大学（PKU）、国家语委（NCC）和美国科罗拉多大学（CTB）提供的训练语料和测试语料。

在对汉语词性标注性能进行评估时，采用了常用的评测指标：标注精度（*Accuracy*）。标注精度表示在全部词语的标注词性中，正确标注的词语所占的比值。计算公式如下：

$$Accuracy = \frac{\text{正确标注词性的词语数}}{\text{所有待标注词性的词语数}}。$$

3.3.2 本组实验使用的特征模板集

为了研究汉语词性标注中上文和下文对其性能的影响情况，我们设定了六组特征模板集，这六组特征模板集包含的特征模板见表 3（注：此时特征模板中的主要语言单位是词，故用字符 W 表示）。需要注意的是本组实验的六组特征模板集中都包含特征模板：$T_{-1}T_0$，该模板用于表征上下文中相邻两个词的词性转移特征 $T_{i-1} \rightarrow T_i$。

表 3　汉语词性标注中特征模板集列表

序号	特征模板集名称	包含的特征模板
1	TMPT-10+B	W_{-2}，W_{-1}，W_0，W_1，W_2，$W_{-2}W_{-1}$，$W_{-1}W_0$，W_0W_1，W_1W_2，$W_{-1}W_1$，$T_{-1}T_0$
2	T10-Above+B	W_{-2}，W_{-1}，W_0，$W_{-2}W_{-1}$，$W_{-1}W_0$，$T_{-1}T_0$
3	T10-Below+B	W_0，W_1，W_2，W_0W_1，W_1W_2，$T_{-1}T_0$
4	TMPT-6+B	W_{-1}，W_0，W_1，$W_{-1}W_0$，W_0W_1，$W_{-1}W_1$，$T_{-1}T_0$
5	T6-Above+B	W_{-1}，W_0，$W_{-1}W_0$，$T_{-1}T_0$
6	T6-Below+B	W_0，W_1，W_0W_1，$T_{-1}T_0$

3.3.3 实验结果及其分析

在表 3 中，序号为 2 和 3 的两组特征模板集分别对应了 TMPT-10+B 中的上文部分和下文部分。序号为 5 和 6 的两组特征模板集分别对应了 TMPT-6+B 中的上文部分和下文部分。通过设定这四组特征模板集，然后进行对比实验，实验结果见表 4。

表 4　不同特征模板集的词性标注结果

模板集序号	特征模板集名称	PKU 语料上评测结果	NCC 语料上评测结果	CTB 语料上评测结果
		Accuracy（%）	*Accuracy*（%）	*Accuracy*（%）
1	TMPT-10+B	93.36	90.17	92.12
2	T10-Above+B	92.81	90.28	90.73
3	T10-Below+B	93.65	90.82	91.48
4	TMPT-6+B	94.25	91.26	92.61
5	T6-Above+B	93.53	91.17	91.18
6	T6-Below+B	94.40	91.64	92.04

综合分析表 4 中的数据可以得出如下结论：在 PKU、NCC、CTB 三个语料库的对比实验中下文所对应特征模板对词性标注性能的贡献要比上文所对应特征模板的贡献大，但相差不大，不超过 1 个百分点。

3.4 中文命名实体识别中上文和下文孰重孰轻

3.4.1 实验数据集和性能评估

本组实验采用的训练语料和测试语料是 SIGHAN 举办的第四届国际汉语语言处理评测 Bakeoff2007 所提供的中文命名实体识别的简体中文语料，由微软亚洲研究院（MSRA）提供。

在对中文命名实体识别进行评估时，采用了常用的 3 个评测指标：准确率（*P*）、召回率（*R*）、

综合指标 F 值（F）、准确率表示在识别的全部命名实体中，正确的所占的比值。召回率指正确识别的命名实体占标准答案中的比值。综合指标 F 值是综合准确率和召回率两个值进行评价的一种办法。

3.4.2 本组实验使用的特征模板集

为了研究中文命名实体识别中上文和下文究竟谁对识别性能的贡献更大，我们设定了三组特征模板集，这三组特征模板集包含的特征模板见表 5（注：此时特征模板中的主要语言单位是字，没有用词特征,故用字符 C 表示）。需要注意的是这三组特征模板集都包含特征模板：$T_{-1}T_0$，该模板用于表征上下文中相邻两个字的词位转移特征。

表 5 中文命名实体识别中的三组特征模板集

特征模板集名称	包含的特征模板
TMPT-10+B	C_{-2}，C_{-1}，C_0，C_1，C_2，$C_{-2}C_{-1}$，$C_{-1}C_0$，C_0C_1，C_1C_2，$C_{-1}C_1$，$T_{-1}T_0$
T10-Above+B	C_{-2}，C_{-1}，C_0，$C_{-2}C_{-1}$，$C_{-1}C_0$，$T_{-1}T_0$
T10-Below+B	C_0，C_1，C_2，C_0C_1，C_1C_2，$T_{-1}T_0$

3.4.3 实验结果及其分析

在表 5 中，T10-Above+B 和 T10-Below+B 这两组特征模板集分别对应了 TMPT-10+B 中的上文部分和下文部分，T10-Above+B 特征模板集仅仅刻画了当前字本身以及当前字前面两个字所组成的特征，而 T10-Below+B 特征模板集仅仅刻画了当前字本身以及当前字后面两个字所组成的特征。通过设定这两组特征模板集，进行对比实验的结果见表 6。从表 6 可以看出，使用下文特征模板集的综合指标 F 值比上文高出了 6.9 个百分点。这表明下文比上文对中文命名实体识别的贡献要大。

表 6 不同特征模板集的命名实体识别结果

不同特征模板集	MSRA 语料上评测结果		
	P	R	F
TMPT-10+B	0.926	0.898	0.912
T10-Above+B	0.902	0.783	0.838
T10-Below+B	0.922	0.892	0.907

4 小结

汉语词法分析作为中文信息处理领域一项基础性研究课题，近年来得到了广泛的关注，其中基于统计的方法成为当前汉语词法分析的主流技术。本文深入研究了汉语词法分析的三个子任务中上文和下文对性能的影响情况，并在国际汉语评测 Bakeoff 的分词、词性标注、命名实体识别三种语料上进行了封闭测试，采用表征上文和下文的不同特征模板集进行了对比实验，结果表明，上文和下文对汉语词法分析的影响大不相同，其中，汉语分词中下文的贡献高出上文贡献 13 个百分点以上；下文对汉语词性标注性能的贡献要比上文的贡献大，但差别不超过 1 个百分点；下文对中文命名实体识别的贡献比上文大 6 个百分点以上。

参 考 文 献

[1] 姜维，王晓龙，关毅，等. 基于多知识源的中文词法分析系统[J]. 计算机学报，2007, 30(1): 137-145.

[2] 刘群，张华平，俞鸿魁，等. 基于层叠隐马模型的汉语词法分析[J]. 计算机研究与发展, 2004, 41(8): 1421-1429.

[3] XUE N W, CONVERSE S P. Combining classifiers for Chinese word segmentation [C]// Proceedings of the First SIGHAN Workshop on Chinese Language Processing. Taipei , Taiwan, China: AS Press, 2002: 20-27.

[4] 黄昌宁，赵海. 中文分词十年回顾[J]. 中文信息学报，2007, 21(3): 8-19.
[5] 宋彦，蔡东风，张桂平，等. 一种基于字词联合解码的中文分词方法[J].软件学报，2009, 20(9): 2366-2375.
[6] 罗彦彦，黄德根. 基于 CRFs 边缘概率的中文分词[J]. 中文信息学报，2009, 23(5): 3-8.
[7] 魏欧，吴健，孙玉芳. 基于统计的汉语词性标注方法的分析与改进[J]. 软件学报，2000, 11(4): 473-480.
[8] 赵岩，王晓龙，刘秉权，等. 融合聚类触发对特征的最大熵词性标注模型[J]. 计算机研究与发展, 2006, 43(2): 268-274.
[9] 洪铭材，张阔，唐杰，李涓子. 基于条件随机场(CRFs)的中文词性标注方法[J]. 计算机科学，2006, 33(10): 148-155.
[10] 王丽杰，车万翔，刘挺. 基于 SVMTool 的中文词性标注[J]. 中文信息学报，2009, 23(7): 16-21.
[11] 刘非凡，赵军，吕碧波，等. 面向商务信息抽取的产品命名实体识别研究[J]. 中文信息学报, 2006, 20(1): 7-13.
[12] 周俊生，戴新宇，尹存燕，等. 基于层叠条件随机场模型的中文机构名自动识别[J]. 电子学报, 2006, 34(5): 804-809.
[13] 鲁松，白硕. 自然语言处理中词语上下文有效范围的定量描述[J]. 计算机学报，2001, 24(7): 742-747.