基于词条组合的军事类文本分词方法

黄 魏[1] 高 兵[1] 刘 异[2] 杨克巍[1]
(国防科学技术大学信息系统与管理学院 长沙 410073)[1] (湖南师范大学文学院 长沙 410081)[2]

摘 要 针对传统的分词方法切分军事类文本存在未登录词多和部分词条特征信息不完整的问题,提出把整个分词过程分解为若干子过程,以词串为分词单位对军事类文本进行分词。首先基于词典对文本进行双向扫描,标识歧义切分字段,对切分结果一致的字段进行停用词消除,计算第一次分词得到的词条间的互信息和相邻共现频次,根据计算结果判定相应的词条组合成词串并标识,最后提取所标识的歧义字段和词串由人工对其进行审核处理。实验结果表明,词条组合后的词串的特征信息更丰富,分词效果更好。

关键词 军事,文本,分词,词条
中图法分类号 TP391.3 **文献标识码** A

Word Segmentation Approach in Military Text on the Basis of Word Combination

HUANG Wei[1] GAO Bing[1] LIU Yi[2] YANG Ke-wei[1]
(School of Information System and Management, National University of Defense Technology, Changsha 410073, China)[1]
(School of Liberal Arts, Hunan Normal University, Changsha 410081, China)[2]

Abstract Since the unknown word in military texts is excessive, and the feature of some words is incomplete, the word segmentation method which is based on lexical chunk as the unit was provided. word segmentation was divided into some sections: bidirectional scanning in the text in the base of dictionary, marking the various and segment the words; deleting the stop-words which share the same segmentation results, then count words mutual information and adjacency frequency by the first time's word segmentation, according to this counting result, the lexical chunk with relevant words can be judged and signed. At last, picked up the signed various segment and lexical chunks to audit and deal with them artificially. The experimentation shows that after the word combination, the lexical chunk bears much more feature information which shares a better effect of the process.

Keywords Military, Text, Word segmentation, Words

随着信息技术的迅猛发展,军事类文本数量急剧增长,对这些文本实行有序的分类并挖掘其有效信息已经成为军事信息处理的关键,对它们进行处理就必须将其转化为适合计算机处理的中间形式,即以结构化形式描述军事类文本信息。在文本挖掘中文本通常被表示为向量空间模型,而要将军事类文本表示为空间向量,首先就要对其进行分词处理。文献[1]面向作战命令自动理解研究了作战命令的分词技术,最终得出采用多个成熟的汉语分词系统进行作战命令的分词是可行和有效的;文献[2]面向装备信息管理的自动化水平探讨了军用词典库的设计。本文主要从面向分类的文本特征表示的角度探讨军事类文本分词方法。

汉语自动分词是中文信息处理的基础,通过分词可以将文本转化为一个词语的集合,文本特征就是从这个集合中产生。当前主流的汉语自动分词方法主要包括基于词典和基于统计的分词方法[3,4],在分词过程中发现,采用这些方法切分军事类文本所获得的词语集合难以满足文本特征抽取的要求,这主要表现在两个方面:一是军事类文本的文本内容专业性较强,其专业词汇或新词的数量较多,在目前缺少军事类分词词典的情况下,分词时出现未登录词的数量要比普通文本分词更多,而且这些未登录词对文本语义的完整表示有着一定的影响;二是当前的汉语自动分词方法多是针对普通文本,要求能快速识别和匹配词条,对其语义信息表示的要求不高,而采用这些方法切分军事类文本所得到的一些词条不符合特征词的要求,其作为特征信息不完整,文本语义的表示性不强。

基于词典与基于统计的分词方法有一个共同点,它们都是基于字符串构词对文本进行切分[4,5],其对文本分全率有一定的要求,这也是其分词后的待选特征词在文本语义表示上不够完整的主要原因。因此,为使军事类文本分词后的词语特征信息更丰富和完整,本文提出了以词串为分词单位对军事类文本进行分词,把整个分词过程分解为若干子过程,通过多步处理策略解决军事类文本的分词问题。

1 词条组合

分词就是将连续的字(词)序列按照一定的规范重新组合

到稿日期:2009-08-26 返修日期:2009-11-09 本文受"十一五"武器装备预先研究项目(513300102)资助。

黄 魏(1982-),男,博士生,主要研究方向为体系工程、文本挖掘等,E-mail:wayewong@nudt.edu.cn;高 兵(1984-),男,硕士生,主要研究方向为体系工程、文本挖掘等;刘 异(1983-),女,硕士生,主要研究方向为汉英语言对比、计算语言学等;杨克巍(1977-),男,博士,副教授,主要研究方向为体系需求建模技术、系统管理与综合集成技术。

成词序列的过程[6]，我们称组合在一起的字(词)为词条。文献[7]认为词组(Phrase)比单个的词能更好地表示文档的内容语义，在文本分类时以词组为基本单位能够获得更好的分类效果，即词组更适合作为文本的特征词。英语文本词与词之间以空格作为自然分界符，英语文本中的“词”与汉语分词后的字串即“词条”在形式上是一致的。因此，将汉语文本切分后的字串组合成词串能更好地表示文档的内容语义，即词条组合。

当前汉语自动分词的方法较多且相对成熟，在某种程度上可以达到相当高的正确率，满足某些层面的需求[8]。但是，对于军事类文本而言，这种“合法”的切分方式往往会带来“不合理”的切分结果，这主要表现在一些被准确切分的词作为候选特征词时语义表示的效果相对较差。如“反空袭”在分词时会被切分为“反/ 空袭”，“反”与“空袭”分开均不能真实地表示其原来的语义；“大规模杀伤性武器”通常被切分成“大规模/ 杀伤性/ 武器”，而如果将“大规模”与“杀伤性”组合在一起，“大规模杀伤性/ 武器”或“大规模杀伤性武器”所含的特征信息更丰富，语义表示效果要更好。而从分词时容易出现的未登录词问题来看，词条组合可以有效地解决这个问题，如“河/ 战/ 行动”可以组合成“河战/ 行动”，此外，许多词典不能匹配的装备名称也可以通过词条组合在一起成为文本的特征词。

因此，为获得军事类文本“合法且合理”的分词结果，使特征词的特征信息更完整，可以将原有的以字串为单位进行分词的形式改成以词串为单位分词，即在传统分词方法切分文本的基础使一些关联性强的词条组合在一起，增加特征词的语义信息，同时在一定程度上也可以减轻文本中歧义切分和未登录词识别问题对分词的影响。

2 词条组合方法

2.1 基于词条组合的军事类文本分词

歧义切分是文本分词中所难以回避的问题，首先通过正、逆向最大匹配法对文本进行双向扫描切分，一些统计结果表明[9,10]，单纯使用正向最大匹配的错误率为 1/169，单纯使用逆向最大匹配的错误率为 1/245，因此，双向扫描结束后返回逆向切分结果，对双向扫描切分结果不一致的部分即歧义字段进行标识；其次，选择相应的停用词表，消除双向扫描切分结果一致的部分中可能影响到词条组合的噪音字或停用词；第三，根据词条间的互信息原理计算词条组合的可能性，同时标识组合在一起的词串，在这个过程中一些歧义字段将随着词条的组合被消除歧义；最后，为进一步提高分词的准确率，可以提取被标识的歧义字段和组合的词串由人对其进行审核和处理，对词条组合后仍然存在的歧义字段进行人工处理，若审核后组合的词串还存在问题，则通过人工或统计模型对词条进行再组合，审核通过后则生成文本的特征词集合。

2.2 基于互信息和共现频次的词条组合

经过双向扫描切分和停用词消除后，文本成了一个由若干词条组成的集合。词条组合主要分为两个词条的组合以及两个以上的词条组合。

(1)两个词条的组合

两个词条的组合采用的是分词中统计方法的思想：文本中邻近的字同时出现的次数越多，就越可能是一个词。基于统计的原理主要有互信息原理、N-Gram 统计模型原理、t-测试等，其中，N-Gram 方法强调的是文本中的第 n 个词的出现只与前面的 $n-1$ 个词相关，而与其它任何词都不相关，t-测试方法强调的是字与前后相邻字相连的趋势比较，而词条组合主要关注相邻的两个词之间的相关性，即组合的可能性。因此，N-Gram 和 t-测试方法不适合用于词条组合，而互信息的原理比较适合词条组合，即可通过对文本中邻近出现的各个字的组合频度进行统计来计算它们的互信息量，然后通过设定阈值的方法来判断若干个字是否组成一个词[11]。传统的基于统计分词方法是以字为统计单位的，而在词条组合时，频数的统计是将每一个词条视为一个统计单位，即统计词条相邻出现的概率，相邻的词条出现的次数越多，其就越有可能组合成一个新的词条。可以通过对文本中相邻出现的各个词条的组合频度进行统计，计算它们的互现信息，当互现信息值高于某一阈值时，便可认为这些词条可以组合成一个新的词条。

对有序词条 A,B，A 和 B 联合出现的概率表示为 $P(A,B)$，词条 A 的出现概率表示为 $P(A)$，词条 B 出现的概率表示为 $P(B)$，它们在文本中出现的次数分别计为 $n(A)$，$n(B)$，$n(AB)$，n 是词频总数，则有公式：

$$P(A,B)=\frac{n(AB)}{n},P(A)=\frac{n(A)}{n},P(B)=\frac{n(B)}{n} \tag{1}$$

词条 A 和 B 之间的互信息定义如下：

$$I(A,B)=\log_2\frac{P(A,B)}{P(A)P(B)} \tag{2}$$

互信息反映了词条 A,B 之间的相关程度。如果 $I(A,B)\geqslant 0$，即 $P(A,B)\geqslant P(A)P(B)$，则 A,B 间是正相关的，随着 $I(A,B)$ 增加，相关度增加，如果 $I(A,B)$ 大于给定的一个阈值 ε_1，这时可以认为 A,B 可组合成词串，否则词条 A 和 B 不能被组合成词串。

与以字为单位的统计方法相比，词条的组合缩小了统计的范围，缩减了统计模型的规模，采用互信息的方法也能有效地消除歧义切分[5]。但是，分词后许多词条的出现频次都为 1，称这些词条为单频词条，单频词条的大量出现影响了互信息的计算结果。在分词测试时发现，单频词条所产生组合的互信息值相对较高，尤其是两个相邻的词条在分词样本中只出现一次时，$I(A,B)=\log_2 n$，即其互信息值随着样本规模的扩大而不断增大，其结果将致使许多关联不强的词条组合在一起。因此，对于任意的两个相邻的词条 A_x,A_y，当 $F(A_x)=1$ 或 $F(A_y)=1$ 时，令 $W(A_x,A_y)=0$，即 A_x 和 A_y 不构成组合关系，当 $F(A_x)>1$ 且 $F(A_y)>1$ 时，令 $W(A_x,A_y)=1$，即 A_x 和 A_y 可能被组合：

$$W(A_x,A_y)=\begin{cases}1 & F(A_x)>1 \text{ and } F(A_y)>1\\ 0 & F(A_x)=1 \text{ or } F(A_y)=1\end{cases} \tag{3}$$

其中，$F(A_x)$，$F(A_y)$ 表示词条出现的频次。当 $W(A_x,A_y)=1$ 时，计算 A_x 和 A_y 的互信息，即计算式(2)。

(2)两个以上的词条组合

对于两个以上的词条组合可以通过词条的多次组合来实现，如“非/ 战斗/ 人员”先组合为“非/ 战斗人员”，而后两个词条再组合为“非战斗人员”。但是，采用这种多次组合的方式需要对文本进行多次扫描，一定程度上降低了分词效率。如果是计算两个以上的词条的互信息，则实现起来较为复杂。

如果应组合在一起的词串被切分为两个以上的词条，其多为未登录词，从文本特征词抽取的角度来看，如果未登录词在文本集中出现的频次较低，则其对特征词抽取并不会产生负面影响。因此，以两个以上的词条在文本中共现的频次来进行判断其是否应组合[12]：

$$W(A_1,A_2,\cdots,A_m)=\begin{cases}1 & F(A_1,A_2,\cdots,A_m)\geqslant\varepsilon_2\\ 0 & F(A_1,A_2,\cdots,A_m)<\varepsilon_2\end{cases}\quad(4)$$

其中，$W(A_1,A_2,\cdots,A_m)$表示 m 个词条组合，$m\geqslant3$，$F(A_1,A_2,\cdots,A_m)$表示这些词条的共现频次，ε_2 为设定的阈值。

一些歧义字段在词条组合后，其歧义将被消除，如“常/驻地”和“常驻/地”这两种歧义切分经过词条组合为“常驻地”，其歧义自然消除。歧义切分消除后其相应的标识也应同时消除，即其已经被视为正确的切分。对于词条组合而成的词串也需要进行标识，其标识符区别于歧义字段的标识。

最后，为了提高分词结果尤其是词串的准确率，应提取所有被标识的歧义字段和词串，由人工审核和处理，从而形成特征词选择所需要的分词结果。组合的词串将出现 3 种情况：可组合、组合或不组合皆可、不组合，人工审核中人的主观性对结果的判断有着一定的影响，且关于词条是否组合并不存在统一的标准，因此，对于组合或不组合皆可的词串在人工审核时都判定其为不组合。人工审核后把认可的词串加入军事类文本分词词典，这样将有助于下一次文本切分时准确地识别这些词串，提高分词的效率。

2.3 算法描述

输入：军事类文本、分词词典和停用词表

输出：待选特征词集合

Step 1　读入军事类文本，搜索词典，执行逆向最大匹配法切分军事类文本，得到逆向切分结果集 Converse{}，切分结果集中每个词条的属性标记为 Inconsistent；

Step 2　读入军事类文本，搜索词典，执行正向最大匹配法扫描军事类文本，得到正向切分结果集 Direct{}，在 Converse{}中标记与 Direct{}中一致的切分词条的属性为 Consistent；

Step 3　读入停用词表，对 Converse{}中属性为 Consistent 的字段执行停用词消除；

Step 4　对 Converse{}执行式(3)，对 $W(A_x,A_y)=0$ 的词条不进行词条组合处理；

Step 5　设定 ε_2，对消除停用词后的 Converse{}执行式(4)，结果大于 ε_2 时进行词条组合，组合后的词串属性标记为 Combination；

Step 6　设定 ε_1，对 Converse{}中属性不为 Combination 的词条执行式(1)、式(2)，结果大于 ε_1 时进行词条组合，组合后的词串属性标记为 Combination；

Step 7　输出 Converse{}。

人工修正过程：

分词人员修正 Converse{}中属性为 Inconsistent 的词条以及属性为 Combination 的词串，将修正后的结果录入分词词典。

3　实验结果与比较

本文采用基于词条组合的分词方法对美军联合能力清单的译本进行了分词实验，同时从互联网中选取医学类文本和普通文本进行了分词测试，主要从 4 项指标对本文所提算法在普通文本和军事类文本中分词的效果进行了比较，结果如表 1 所列。

表 1　与其它类别文本分词比较

	军事类	医学类	普通类
文本规模	12074	8361	15265
词串数	433	275	477
组合准确率	75.52%	71.63%	59.11%
组合错误率	9.01%	8%	11.53%
歧义消除率	86.7%	80%	86.96%
识别未登录词	17	19	9

其中，文本规模指采用逆向最大匹配得到的词条数；词串数指通过词条组合形成的词串数量；词条组合的准确率指人工审核词条可组合的词串比率；词条组合的错误率指人工审核词条不可组合的词串比率；歧义消除率指标识的歧义切分中被消除的比率；识别未登录词是词条组合后识别的未登录词数量。

各项数据对比后表明，本算法在军事类或其它专业性较强的文本中分词效果较好。

(1)词条组合后缩减了文本规模，被组合的词条约占文本总词条数的 10%；

(2)词条组合后词串的语义表示效果更好，如“防扩散”、“快反”、“稳定性作战”以及“大规模杀伤性”，其中专业性较强的文本词条组合准确率较高，这是因为专业性较强的文本中词条间的关联性更强；

(3)普通文本中“组合或不组合皆可”的比例较高，这与普通文本向量分布均匀有关，这些词串在文本特征的表现中可能产生噪音，如“时间地点”、“从而达到”等；

(4)对歧义切分的消除较好，其主要原因在于本算法结合了词典与统计的方法进行分词，词条的组合也有利于对这些字段进行第二次的判定，但是在对测试文本进行分析时发现本算法在未登录词的识别上效果一般，其原因在于词条组合时排除了单频词条；

(5)单个词条的频数对词条组合有着直接的影响，在不考虑文本语义通顺的情况下对一些词条的频数进行调整后发现：单个词条频数越高，词条组合的准确率越高。

结束语　当前关于汉语自动分词方法的研究较多，各种算法相对成熟，但采用词条组合的思想对军事类文本分词也不失为一次有益的尝试，通过实验发现其对于军事类文本分词的难题解决有一定的帮助，尤其是从分词的角度增强了待选特征词的语义信息。但是，囿于作者的水平，本方法还存在一些待完善或进一步改进的地方。

(1)组合错误率偏高，在下一步的研究中可通过与其它算法相结合来对本文的方法进行改进，如结合 t-测试的方法判断词条前后组合趋势的问题，同时也有助于解决歧义切分的消除问题，此外在研究中发现，利用词性搭配的规律可以降低词条组合的错误率，如连词和动词不组合等，这就需要在第一次切分时标注词条的词性；

(2)将所有单频词条排除在组合之外对于词条组合的正确率有很大的帮助，但是它也阻止了一些准确组合的出现，如“通信链”、“短时间”、“全谱”等，同时对歧义切分消除尤其是未登录词的识别问题也产生了一定的影响；

(3)从词串的数目来看，人工干预的工作量相对较大，人工审核词条组合也缺乏统一的标准。

因此，下一步关于词条组合的研究将聚焦于词条组合规则的建立，基于词条的频数研究文本中词条分布的规律，并通过调整不同频数词条的分布发现词条组合的规则。从现在的研究来看，关联规则挖掘的相关算法对词条组合规则的建立有一定的帮助。

参 考 文 献

[1]　姜文志，范洪达，聂心东，等. 作战命令的分词技术研究[J]. 海军航空工程学院学报，2008，23(1)：52-54

[2] 姜文志，蒋伟俊，张金乙，等. 军用词典库的设计[J]. 兵工自动化，2007，26(8)：50-51
[3] 滕少华. 基于 CRFs 的中文分和短文本分类技术[D]. 北京：清华大学，2009
[4] 许高建，胡学钢，王庆人. 文本挖掘中的中文分词算法研究及实现[J]. 计算机技术与发展，2007，17(2)122-124
[5] 孙铁利，李晓微，张妍. 信息过滤中的中文自动分词技术研究[J]. 计算机工程与科学，2009，31(3)：80-82
[6] (美)Allen J. 自然语言理解(第二版)[M]. 刘群，等译. 北京：电子工业出版社，2003
[7] Furnkranz J，Mitchell T，Riloff E. A Case Study in Using Linguistic Phrases for Text Categorization on the WWW[R]. AAAI Technical Report WS-98-05. www. aaai. org，1998
[8] 曲维光. 汉语自动分词的方法选择[J]. 计算机科学，2002，29(9)：54-56
[9] Ma Yuchun，Song Hantao. Research of Chinese word segmentation based on the Web[J]. Computer Application，2004，24(4)：134-136
[10] 陈平，刘晓霞，李亚军. 基于字典和统计的分词方法[J]. 计算机工程与应用，2008，44(10)：144-146
[11] Sun Maosong，Shen Dayang，Tsou B K. Chinese Word Segmentation without Using Lexicon and Handcrafted Training Data[C]// Proceedings of the 36th Annual Meeting on Association for Computational Linguistics. Montreal：Association for Computational Linguistics，1998：1265-1271
[12] 张仰森，曹大元，俞士汶. 基于规则与统计相结合的中文文本自动查错模型与算法[J]. 中文信息学报，2006，20(4)：1-7

(上接第 112 页)

在 Windows xp 平台下，利用 API 函数，通过直接计算 CPU 主频测得一次 COMMAND 和 POCKET 交互所需时间，如图 4 所示，测试 C 编码如下：

```
QueryPerformanceFrequency(&Freg);
QueryPerformanceCounter(&Count1);
//……交互过程。
QueryPerformanceCounter(&Count2);
double d = (double)(Count2. QuadPart
    -Count1. QuadPart) / (double)Freg. QuadPart;
```

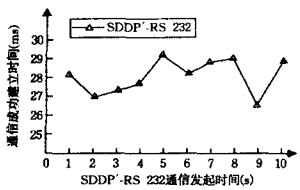

图 4 SDDP′-RS232 的建立时间

4.3 SDDP-OPC 分析

OPC 是工业控制中一个重要的软件接口标准，被广泛应用于测控系统中[10]。随着工业控制信息化程度的不断提高，传感器等一些设备的重要参数和数据需要集中保存在管理层的数据库中。现场设备在初始化或者控制过程中需要获得这些参数，而已有的接口广泛使用了 OPC。

OPC 接口主要由服务器(Server)、组(Group)和项目(Item)组成，而 SDDP 描述的就是传感器节点对象、通道和属性，两者能够较好地对应起来。在 MiniCS 系统中，用一个 OPC Item 来描述 SDDP 中的属性行，例如“a=channel:1”，将 OPC Item 的名称设为 channel。读取 channel 值的时间性能如图 5 所示。

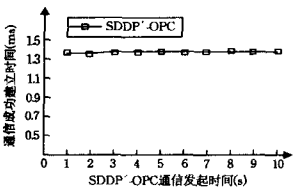

图 5 SDDP′-OPC 的建立时间

从上述几个案例可以看出，SDDP 作为一种与具体承载协议无关的数据描述协议，可以根据实际系统的情况而得以应用，具有较好的开放性。除数据包大小差异外，主要是承载协议的不同会导致 SDDP 传输效率的不同。可以预见工业以太网正作为一种测控网络互联的标准逐步深入到现场层。选择可以灵活改变通信状态的通信协议，也是一大优势。如 SIP 协议，其建立时间本实验系统也只要 10ms。

结束语 本文提出了一种扩展 SDP 的传感器数据描述协议 SDDP，并用 ABNF 语言对其做了形式化描述。实验结果表明其在一定程度上能够满足实际应用需求。与其它协议相比较，扩展 SDP 与 SIP 协议相结合，在开放性上更具有优势。通过 SDDP-SIP 协议，网络化传感器之间可以迅速建立初始化连接，从而引导后续测控任务的建立。扩展 SDP 协议更容易形成新的 RFC 草案，极易建立网络化测控系统接口标准。完备 SDDP 协议内容，扩大 SDDP-SIP 应用范围，寻求科学的 SDDP-SIP 协议形式化描述方法，将是下一步工作的内容。

参 考 文 献

[1] Lee K. Open standards for homeland security sensor networks[J]. IEEE Instrumentation & Measurement Magazine，December 2005：14-21
[2] Botts M，Robin A. OpenGIS® Sensor Model Language (SensorML) Implementation Specification[R]. OGC® 07-000. July 2007
[3] Akyildiz I F，Su Weilian，Sankarasubramaniam Y，et al. A Survey on Sensor Networks[J]. IEEE Communications Magazine，August 2002：102-114
[4] Aboelaze M，Aloul F. Current and future trends in sensor networks：a survey[J]. Wireless and Optical Communications Networks，March，2005：551-555
[5] 李凤保，杨黎明. 网络化测控系统技术[M]. 成都：四川大学出版社，2004
[6] 夏继强，邢春香，耿春明，等. 工业现场总线技术的新进展[J]. 北京航空航天大学学报，2004(4)：358-362
[7] Lee K，et al. IEEE 1451.2 A Smart Transducer Interface for Sensors and Actuators—Transducer to Microprocessor Communication Protocols and Transducer Electronic Data Sheet (TEDS) Formats[EB/OL]. IEEE Standards Department，www. ieee. com
[8] Rosenberg J，Camarillo G，Schulzrinne H. Session Initiation Protocol (SIP)[R]. RFC 3261. June 2002
[9] 刘进，雷为民，杨易凡. 支持视频广播的 SIP 媒体服务器设计与实现[J]. 小型微型计算机系统，2006，27(12)：2264-2267
[10] Xu Hong，Wang Jianhua. Using standard components in automation industry：A study on OPC Specification[J]. Computer Standards & Interfaces，2006，28(4)：386-395