服务于内容侧面发现的框架识别*

王 荀，李素建，宋 涛，姜伯平
北京大学 计算语言学教育部重点实验室，北京 100871
E-mail: {lisujian,wangxun,songtao,jiangboping}@pku.edu.cn

摘 要：文本中的内容通常包含多个侧面，全面地识别这些内容侧面对自然语言处理有重要的意义。但是传统的使用简单特征的统计方法难以识别出所有的内容侧面。以自动摘要为例，传统的抽取式方法多以词频为主要特征，一些重要的句子常因重复度不高被舍弃。要想全面地覆盖原始文本的重要信息，就要识别出文本描述的内容侧面。本文以框架语义学为指导，使用 FrameNet 语料库作为知识库，综合多种特征来标注文本描述的框架，在此基础上识别文本所包含的内容侧面。该方法在新闻语料上取得了较好的结果，达到了 61%的正确率。

关键词：FrameNet 语料库；内容侧面发现；框架识别

Frame Identification for Aspect Recognition

Wang Xun, Li Sujian, Song Tao, Jiang Boping
Key Laboratory of Computational Linguistics (Peking University) Ministry of Education, Beijing 100871
E-mail: {lisujian,wangxun,songtao,jiangboping}@pku.edu.cn

Abstract: Texts usually contain various aspects of information. In natural language processing, many tasks would benefit from the recognition of these aspects. For example, in the summarization task, traditional method of extracting sentences mainly bases on features of words frequency. Sentences of great importance would be ignored if they appear too infrequently. However an understanding of the text will help system to identify important information. In this paper, we use FrameNet corpus as ontology to annotate sentences based on combined features. The frame of the annotated sentence explains its aspect information. This method works well when tested on TAC2010 summarization task test corpus and the precision rate reaches 61%.

Keywords: FrameNet; aspect recognition; frame identification

1 前言

在现阶段自然语言处理研究中，统计学方法应用比较广泛。但是[8][9]单纯基于统计，而不对文本信息进行理解不能保证信息覆盖的全面性。此处以及下文，我们以自动摘要为例，说明理解文本信息的重要性。自动摘要是自然语言处理的一个重要任务，它试图将长文档压缩为不损失重要信息的短文档。常用的方法可以有抽取式和生成式两种，其中抽取式方法应用比较广泛，它主要是从文中抽取句子构成摘要。

抽取式方法多以词频为主要特征[1]，倾向关注出现频次较高的信息。但是人们关注的侧面不同，重要信息并不完全等同于出现频次较高的信息。常常有某些信息虽然重要，却由于频次的关系而不能被选入摘要。特别是在多文档新闻摘要中，新闻报道同质化比较严重，一部分信息的冗余度较高，而另一部分人们关注的信息又由于出现频次不够而被舍弃。在我们参加的 TAC 2010[7]的自动摘要任务中，就提出要求：好的摘要须尽可能全面地报道整个事件，要涵盖不同内容侧面（aspect）的信息。

解决问题的一个方法是减少冗余，这样可以增加信息量，从而可以地提高信息覆盖的全面性。其中 MMR[6]是一种常用的策略，但 MMR 也主要是从词语层的相似度出发，并没有进行语义的分析，因而不能彻底解决这个问题。一个较好的办法是对文本进行语义分析，然后根据文本所描

* 本工作受到国家自然科学基金项目（项目号：60875042、90920011）以及国家社会科学基金项目（项目号：10CYY023）的支持。

述的信息进行取舍。但是完全的语义分析难度较大，正确率较低，实用性较差。本文中，我们提出使用 FrameNet[2]语料库作为知识库，对文本进行语义分析，识别出文本所属的框架，从而帮助判断文本所描述的内容侧面。

本文的内容组织如下：第二节介绍内容侧面和框架语义学，以及作为知识库的 FrameNet 语料库。第三节介绍我们的系统设计，即使用 FrameNet 作为知识库，综合使用多种特征对文本的框架进行识别。第四节说明实验方法和结果，给出相应的分析；并尝试使用框架进行内容侧面的发现。最后由第五节进行总结，并展望未来的工作。

2 内容侧面发现和框架语义

2.1 框架语义学和 FrameNet 语料库

框架语义学（Frame Semantic）最早由 Fillmore[4][5]在 1976 年提出。它认为一个框架（frame）可以表示一个场景，在框架中，动词是核心，其他成分是这个动词的配价成分。而动词和框架之间的关系是多对多的。同一个框架可以使用不同的核心动词，而每个动词也可以在不同的框架中被使用。我们对文本的理解就是确定文本的核心动词以及其对应的框架，了解了这个框架就知道了文本描述的内容，从而帮助我们发现文本所描述的内容侧面。

FrameNet 语料库[1]是在框架语义学指导下建立的，它给出了一个词语的所属的框架、在每个框架下的配价表示并附有例句。我们将利用这些例句和配价表示作为特征，通过相似度计算完成对文本的框架识别，实现对句子的浅层理解，在此基础上再判断其是否含有摘要所需要的内容侧面。

2.2 内容侧面和框架之间的语义映射

文本中的每一个句子都可以根据其语义，[3]将其投射到一个框架。本文将进一步利用识别出来的句子框架，基于框架内部含有的语义角色，将框架映射到不同的内容侧面上。

这儿我们采用 TAC2010 自动摘要任务定义的内容侧面，以袭击事件（Attacks: Criminal/Terrorist）为例，定义了如下可能的内容侧面：

WHAT/事件	what happened
WHEN/时间	date, time, other temporal placement markers
WHERE/地点	physical location
PERPETRATORS/袭击者	individuals or groups responsible for the attack
WHY/原因	reasons for the attack
WHO_AFFECTED/受害者	casualties(death, injury), or individuals otherwise negatively affected by the attack
DAMAGES/损失	damages caused by the attack
COUNTERMEASURES/预防措施	countermeasures, rescue efforts, prevention efforts, other reactions to the attack (e.g. police investigations)

其中 WHAT、WHEN、WHERE 等内容侧面可以使用命名实体识别等方法抽取。但是 WHY、DAMAGES、COUNTERMESURES 等信息则需要对句子内容进行分析然后才能做出判断。这些信息一般只在很少的几篇报道中以较小的篇幅出现，单纯依靠统计很难抽取出来。在进行文本理解，识别出其中所描述的框架之后，这个判断就很容易做出。例如损失（DAMAGES）侧面的内容，一般由损坏、后果等框架进行描述。而预防措施类（COUNTERMESURES）侧面的内容，可由后续、措施等框架描述。这样就可以将框架进一步映射到内容侧面。下文我们将以袭击事件之 WHO_AFFECTED 为例，进行内容侧面的识别。

[1] http://framenet.icsi.berkeley.edu/index.php

3 系统设计

系统如图 1 所示。首先对测试文本进行预处理，对测试文本中的每一句 S，抽取其中的动词 v_1、v_2、v_3……。在 FrameNet 语料库中进行查询，每个动词都对应有若干框架，取出这些动词对应的全部框架记为 f_1、f_2、f_3……，为备选框架集合。对每个框架f_i，取出它对应例句 S_{ij}。根据一系列特征计算例句 S_{ij} 和原句 S 的相似度，得分最高的例句 S_{ij} 对应的框架 f_i 就是原句 S 所对应的框架。框架识别完毕之后，就可以根据句子的框架判断出其描述的内容侧面。

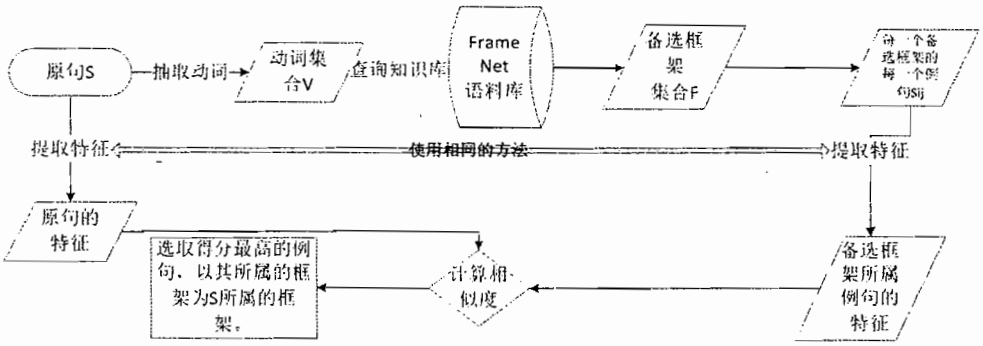


图 1　框架识别流程

我们选择的实验语料是新闻文本，识别内容侧面为自动摘要任务服务，预处理还要过滤掉不适合作为摘要的句子。如报头（Xinhua Agency，Beijing）等和其他一些不含动词的句子。

要计算原句和例句之间的相似度，我们选用了词袋、词性序列、语法分析等特征。

(1) 词袋。两个句子含有的相同的词越多，二者就越相似。此处我们假设当两个句子的相似度高于一个阈值时，则二者所属的框架相同，因此系统赋予该特征一个较高的权重。

(2) 词性序列。词性序列比词袋更抽象，它可以反映句子本身的搭配信息和结构信息等。两个描述相似场景的句子常常因为实词不同而导致句子相似度较低，但是由于句子结构相似，词性序列的相似度一般会很高。因此我们将句子的词性序列作为一个重要的特征。

(3) 依存分析。使用 stanford parser 对原句进行分析，可以得到依存分析的结果，该结果可以说明句子内部成分之间的依赖关系。其中比较重要的是动词和它所支配词语之间的依赖关系。通常相同的依存关系越多，两个句子也越相似。

我们使用以上三种特征计算例句和原句之间的相似度，然后将其综合起来，作为最终的语义相似度，以衡量两个句子——例句和原句之间的语义相似性。

比较各个例句与原句的语义相似度，我们选择得分最高的例句，把该例句对应的框架作为原句对应的框架，即：

$$Sim(S_i,F_j)=max_k\{Sim(S_i,S_{jk})\};Frame(S_i)=argmax_j(Sim(S_i,F_j))$$

S_i 表示第 i 个句子，F_j 表示第 j 个框架。S_{jk} 表示第 j 个框架的第 k 个例句。Frame (S_i) 表示 S_i 所属于的框架编号。Sim(S_i, S_{jk}) 由三部分构成：

1．基于词袋的相似度计算

该相似度使用两个句子的共现词比例来度量。记句 S_i 和 S_{jk} 之间共现词为 m，两个句子的长度分别为 $n1, n2$。另外设置一个阈值 $smax$，定义其相似度为：

$$SemSim(S_i,S_{jk})=\begin{cases}max\left(\dfrac{m}{n1},\dfrac{m}{n2}\right),others\\ MAX,\max\left(\dfrac{m}{n1},\dfrac{m}{n2}\right)>smax\end{cases}$$

2．基于词性序列的相似度计算

词性序列的相似度定义为两个序列之间的Levenshtein distance，即从序列A1变化到序列A2需要的最少的增、删、移位操作步骤数。

$$PosSim(S_i, S_{jk}) = L\_dis(Pos_i, Pos_{jk})$$

其中POS_i是第S_i个句子的词性标注序列。POS_{jk}是第j个框架的第k个例句的词性标注序列。

3．基于依存分析结果之间的相似度

依存语法关注句子成分之间的依存关系。我们使用Stanford Parser来对句子进行依存分析。依存分析给出的结果形如：*relationship* (*word*1,*word*2)。其中word1和word2是依赖对，relationship是依赖的关系。相似度为原句和例句中共现的依赖对以及关系占全部分析结果的比例。

4．总的相似度计算

例句和原句之间的总相似度由以上三个相似度组成，我们采用简单的策略，使用线性加权平均的方法综合这三个得分得到一个最终得分：

$$\mathrm{Sim}(S_i, F_j) = \max\{\lambda_1 \mathrm{SenSim}(S_i, S_{jk}) + \lambda_2 \mathrm{PosSim}(S_i, S_{jk}) + \lambda_3 \mathrm{DepSim}(S_i, S_{jk})\}$$

$$\mathrm{Frame}(S_i) = \arg\max_j(\mathrm{Sim}(S_i, F_j))$$ 其中F_j遍取所有可能对应的框架。

下文的实验中将根据实验的结果对三个参数进行调整。

4 实验

我们的训练语料和测试语料都选自TAC2010自动摘要任务的测试语料。使用FrameNet语料库作为知识库，综合采用多个特征对每个句子所描述的框架进行识别。使用的工具有Porter Stemmer[2]、Stanford Parser[3]等。评测指标采用常用的准确率。

4.1 试验准备

我们使用三组不同的参数，对1000个随机选择的句子进行框架识别，然后抽取出三组测试中标注结果一致的500个，经过人工检查后抽取出100个作为训练语料。测试语料同样选自TAC2010的自动摘要任务测试语料。随机选择100个句子。从中剔除较短的（长度不超过四个单词）和不含有动词的句子，再继续选取句子补足100个。

对语料进行预处理，需要去除不适合作为摘要的句子：如报头等；使用规则简化xx said that, It's said/believed… that等句式。使用stanford parser对原句进行处理，得到词性标注和依存分析的结果。根据词性标注结果，找出句中所有的动词并抽取词干，对FrameNet语料库进行同样的处理，以保证同一个动词被映射到同一个词根上面。

4.2 试验

4.2.1 分别基于三种特征的实验

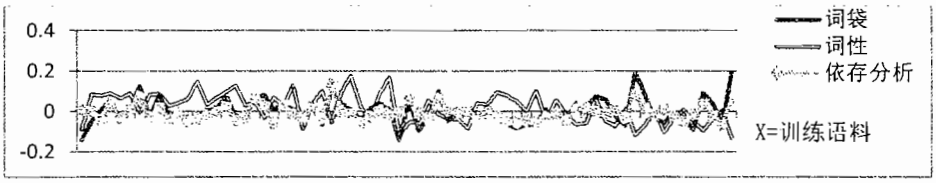
首先我们分别使用三个特征，对训练语料的框架进行识别。图2给出了分别基于三种特征进行框架识别的准确率。从图中我们可以看出，第2个特征效果比较明显。相比之下，另外两个特征效果较差，因此在综合三个得分的时候，可以考虑将第2个特征权重设置高一点。

4.2.2 三个特征之间关系的分析

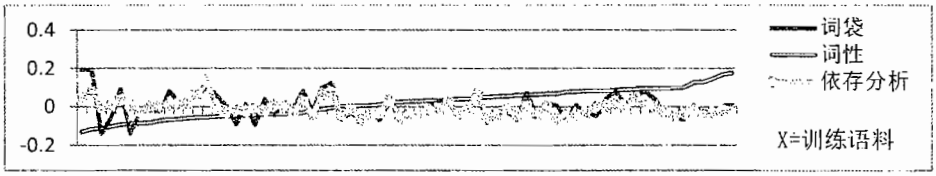
对于每个句子，根据三个特征可得到三个相似度得分。我们将训练语料上的三个相似度得分值归一化。然后做散点图，如图3（a），可以看出三个特征之间有一定关系。我们再将数据按照词性序列相似度得分排序，如图3（b），可以看出词袋和依存分析结果这两类特征的相似度得分关系

较紧密，说明二者有一定的相关关系，且与词性序列特征略有正相关[1]。

图2　分别基于三种特征的实验



(a)



(b)

图3　三个相似度得分之间的关系

4.2.3　参数调整

根据实验 1 我们控制参数 λ_2 使之最大，然后调整另外两个参数。调整参数的结果，如图所示：首先控制 λ_2=1, λ_1=0 调整 λ_3。

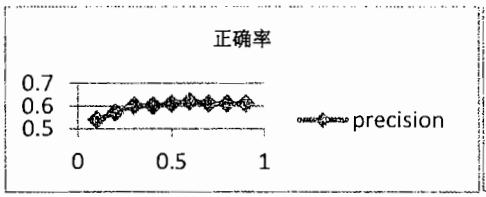


图4　调整参数 λ_3

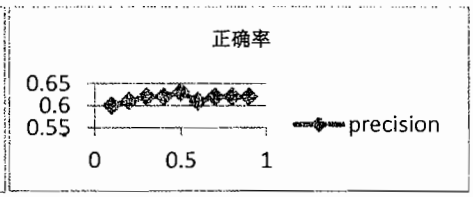


图5　调整参数 λ_2

由图 4，可令 λ_3=0.6；同样，在 λ_2=1, λ_3=0.6 情况下，调整 λ_2。

$$Sim(S_i,F_j) = \max\{\lambda_1 SenSim(S_i,S_{jk}) + \lambda_2 PosSim(S_i,S_{jk}) + \lambda_3 DepSim(S_i,S_{jk})\}$$

由图可以看出当上式的参数设置为{0.5,1,0.6}时，可以取得最好的结果。

4.2.4　框架识别和服务于内容侧面识别

本小节的实验是为了验证框架的识别有利于内容侧面的发现。测试语料为人工标注了框架的 100 个句子。在人工标注结果中，这 100 个句子属于 65 个框架。我们使用上文选定的{0.5,1,0.6}作为参数，在测试语料上进行试验。在测试结果中，100 个句子被识别到 72 个框架中，其中识别正确的句子有 61 个，正确率为 61%。

我们以哥伦比亚中学枪击案事件的 WHO_AFFECTED 内容侧面识别作为例子。通过对框架的定义和框架内部包含的语义角色的研究，根据专家知识预先人工判断出其中和 WHO_AFFECTED 相关的框架有：killing、Shoot_projectiles、attacking 等。报道该事件的 10 篇新闻经过预处理后一共

[1] 相关分析在相似度得分值归一化之后进行。

有230个语句，被识别后属于100个框架。其中属于和WHO_AFFECTED相关的框架有两个：killing、Shoot_projectiles。识别为killing的句子有8个，Shoot_projectiles的句子有3个。搜索空间缩小为原来的4.7%。而经过人工检查，230个句子中共有9个和WHO_AFFECTED相关。在识别为killing和Shoot_projectiles的句子中包含了其中的6个。正确率75%。通过这个例子可以看出，识别句子所属的框架、判断句子所属的内容侧面对完成自然语言处理任务很有帮助。

4.3 结果分析

从实验的结果可以看出在综合使用多个特征的时候，在新闻语料上框架识别能达到63%的正确率。我们对分析错误的句子进行了研究，发现有以下几点规律：

第一，复句比单句更容易出错。复句结构复杂，包含的动词比较多，有时候主要信息在从句里面，有时候信息在复句里面，增加了识别的难度。第二，常用动词的意义种类繁多，FrameNet语料库中没有也很难标出它们所属的全部框架。根据统计，出错最多的动词有go\come\take\bring等等。这些动词应用广泛，意义灵活，可以用在很多框架中，难以穷举。第三，句子包含的内容侧面信息和动词的意义指向并不统一，这是因为动词描述的场景和句子本身表达的意思并不统一。另外，句子描述的框架有时候并不是唯一，但我们的系统只能给每个句子赋予一个框架。通常只有意义明确的动词识别效果比较好，因为这样的动词在不同的框架中差别较大，特征明显。

5 总结和展望

我们在实验中综合采用多个特征来标注句子所属的框架，取得了较好的效果，说明这些特征较好地反映了句子的语义。下一步可以考虑加入PCFG语法树特征，来进一步提高识别的正确率。在此基础上我们还将使用机器学习等方法将框架映射到内容侧面，完成内容侧面的自动识别，以更好地服务于自动摘要等自然语言处理任务。

参 考 文 献

[1] A. Nenkova and L. Vanderwende, The Impact of Frequency on Summarization. Microsoft Research Technical Report. 2005, MSR-TR-2005-101.

[2] Baker, Collin F., Charles J. Fillmore and Beau Cronin, The Structure of the Framenet Database, International Journal of Lexicography, 2003, Volume 16. 3: 281-296.

[3] Boas, Hans C., A frame-semantic approach to identifying syntactically relevant elements of meaning. Steiner, Petra, Boas, Hans C., and Stefan Schierholz (eds.), Contrastive Studies and Valency. Studies in Honor of Hans Ulrich Boas. Frankfurt/New York: Peter Lang,2006: 119-149.

[4] Fillmore, Charles J., Frame semantics and the nature of language. Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech, 1976, Volume 280: 20-32.

[5] Fillmore, Charles J., Frames and the semantics of understanding. Quaderni di Semantica, 1985, Volume 6. 2: 222-254.

[6] Jaime Carbonell and Jade Goldstein, The use of MMR, diversity-based reranking for reordering documents and producing summaries, SIGIR '98 Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, 1998: 335-336.

[7] TAC 2010 Guided Summarization Task.

[8] 刘挺，吴岩，王开铸. 自动文摘综述. 情报科学，1998，16(1): 63-69.

[9] 秦兵，刘挺，李生. 多文档自动文摘综述. 中文信息学报，2005，19(6): 13-20.