

文章编号: 2096-1472(2016)-11-17-03

基于句法结构分析的目标词识别研究

张立凡

(太原工业学院计算机工程系, 山西 太原 030008)

摘要: 对于一条给定的句子, 目标词识别就是识别出句子中能够激起语义场景的目标词。针对目标词识别中的特征选择问题, 本文把目标词识别任务看作是一个分类问题, 在传统的词法特征和上下文特征基础上, 加入了句法结构特征设计特征模板, 识别句子中的目标词。在汉语框架网的标注语料集上进行测试, 实验结果表明, 相比于传统的词特征, 基于句法结构分析的目标词识别率有显著地提升。

关键词: 框架语义分析; 目标词识别; 句法结构分析; 汉语框架网

中图分类号: TP391 **文献标识码:** A

A Study on Target Word Identification Based on Syntactic Structure Analysis

ZHANG Lifan

(School of Computer Engineer, Taiyuan Institute of Technology, Taiyuan 030008, China)

Abstract: For a given sentence, target words identification is to identify a target word that can evoke the semantic scene. Aiming at the feature selection problem in target words identification, it is considered as a classification problem. On the basis of traditional lexical features and contextual features, the syntactic structure feature template is added to identify the target words in the sentence. Compared to the traditional word features, the results show that the rate of target words identification based on syntactic structure analysis is significantly improved.

Keywords: frame-semantic analysis; target word identification; syntactic structure analysis; Chinese FrameNet

1 引言(Introduction)

框架语义分析的任务是从语义角度, 在给定的文本中自动地识别目标词, 为目标词分配框架, 如果分配时有歧义, 进行框架排歧, 最终识别出框架中的论元, 为论元标注语义角色。在整个任务中, 首要解决的就是目标词的识别任务。例如, 文本“根据周恩来的指示, 贺龙等绕道香港去上海寻找党中央。”的框架语义分析结果如图1所示。要得到这个分析结果, 首先要识别出文本的目标词“去”和“寻找”。

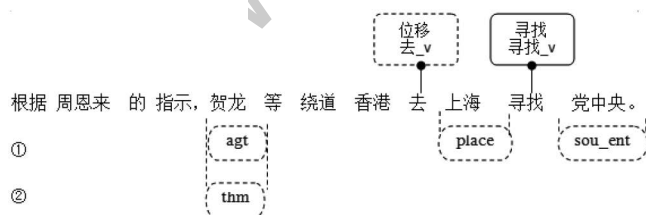


图1 框架语义标注示例

Fig.1 Sample of frame semantic labeling results

近年来, 越来越多的学者关注框架语义分析在某些特定领域的应用^[1-3]。目前, 与目标词识别任务相关的研究, 主要有目标词扩展和目标词识别两方面的工作。针对目标

词扩展, Jahansson^[4]与Das等^[5]认为待识别的目标词包含在例句和训练集词表中, 利用规则筛选目标词。文献分别利用 WordNet和SIM^[6]词典, 识别和收集语义近似的目标词, 实现目标词的扩展。针对目标词识别, 文献[7]使用基于规则的目标词过滤方法进行判定, 若判定为非目标词, 利用基于监督学习的方法予以修正。文献[8]利用分类模型, 在特征模板中加入同义词词林编码信息来识别句子中的核心目标词, 但同义词词林信息的加入会导致识别性能下降。在上述文献中, 大多数是把句子中与目标词相近的词都作为特征来进行分类, 而没有考虑到句子结构特征对目标词识别的影响。事实上, 句子中的很多词对目标词识别并没有帮助, 反而会给目标词识别带来噪音, 如何合适地选择句子中的某些词作为分类特征是目标词识别的一个研究重点。

本文利用句法结构来抽取目标词识别的特征, 结合分类模型, 提出了如何选择句子中合适的词语作为特征的方法。这种方法利用句法依存分析的结果, 分析句子中各个词之间的依存关系和关系类型, 去掉那些对目标词识别没有作用的词语, 选择句子中关键的词作为特征来进行识别目标词。

2 目标词识别模型(Target word identification model)

2.1 目标词识别任务形式化描述

对于一个待识别的文本 $S = \{w_1, w_2, \dots, w_n\}$ ，其中， n 表示文本 S 中词的个数。目标词识别，首先要识别出句子中的词元 $L = \{l_i | l_i \in S\}$ ， k 表示 S 中的词元个数。然后，在词元中识别出目标词 $T = \{t_j | t_j \in L\}$ ， m 表示 S 中的目标词个数。

则目标词识别任务，可形式化地描述为

$$\hat{t} = \arg \max_{t_k \in T} R(t_k | S)$$

2.2 基本概念

框架^[8]：框架是指与一些激活性语境相一致的结构化范畴系统，它是储存在人类认知经验中的图式化情境，是理解词语的背景和动因。

词元：在框架语义学中，词通常被称之为词元。

目标词^[8]：当词元在句子中能够激起语义场景时，被称为目标词。在一条包含多个目标词的句子中，如果某个目标词激起的框架及其在句中所支配的框架元素依存项相比其他框架更能完整表达该句的核心语义，该目标词即为核心目标词。

2.3 特征抽取

本文将目标词识别任务看作是一个分类问题。解决分类问题最重要的两部分内容是特征选择和模型的选择，本节重点介绍目标词识别任务中特征的选择。本文利用句法结构来提取目标词识别任务中的分类特征，以依存句法分析的结果为基础，分析句子中与目标词存在依存关系的词，去掉句子中有噪音影响的词，从而提高识别的性能。

句子特征提取方法分为三步：

第一步，查找句子中的目标词。这一步主要是在词元库中抽取所有的词元构成候选词元列表。针对一条输入的句子中，句子经分词后，句子中的词逐个在候选词元列表中进行匹配，若匹配成功，则标记为目标词。

第二步，对句子进行句法分析，得到依存弧和关系类型。本文对句子句法分析利用依存句法分析工具，句子经依存句法分析后，可得到句子中各个词之间的依存关系和依存弧。

第三步，提取句子主干作为分类特征。以目标词为核心，在句子中抽取与目标词相关的句子主干特征，主要抽取的是句子中的主语、谓语和宾语成分。

针对目标词识别任务，本文抽取句子的词法特征、上下文特征和句法特征。

- (1)词法特征。词法特征主要抽取的是当前词的词性、词。
- (2)上下文特征。主要考虑的是词语搭配对目标词识别的

影响，一般考查的是在位置上与目标词相近的词。

(3)句法特征。句法特征抽取的是句子主干及其附属成分。句子主干抽取的是句子的主谓宾。主语成分抽取的是句子中依存关系标记为SBV的词，谓语成分抽取的是句子中依存关系为VOB的词。

例如，句子“这里四周有冰山阻隔，海上白雾茫茫，绝不会被人发现。”使用句法分析后的依存关系图如图2所示。

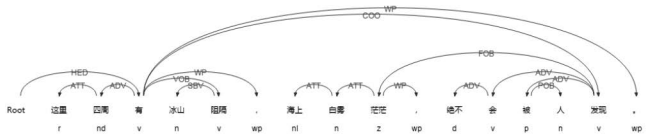


图2 依存句法结果示例
Fig.2 Sample of dependent syntactic results

目标词识别就是识别出这个句子中的目标词“阻隔”，具体的特征值如表1所示。

表1 分类特征抽取示例

Tab.1 Sample of feature extraction

特征类型	特征简称	特征描述	特征值
词法特征	TGT_LEMMA	目标词的词形	阻隔
	TGT_POS	目标词的词性	V
上下文特征	C_LEMMA	与目标词距离为k词的词形	冰山 海上
	C_POS	与目标词距离为k词的词性	n nl
	S_SBJ	句子主语成分	冰山in
句法结构特征	S_PRE	句子谓语成分	阻隔v
	S_OBJ	句子宾语成分	有v

3 实验(Experiment)

3.1 数据来源

实验语料来源于汉语框架网^[9]中的例句库。本文选取汉语框架网例句库中30个常用框架的例句进行实验，共包含2542条句子和2813个目标词。

3.2 评价指标

本文采用准确率 P (Precision)和 Avg_P 平均准确率评价目标词的识别性能。为了避免实验过程中由语料规模小引起的过拟合问题，采用3折交叉验证进行实验。

$$P = \frac{A}{A + B} \tag{1}$$

$$Avg_P = \frac{\sum_{i=1}^n P_i}{n} \tag{2}$$

其中，在式(1)中 A 表示正确识别的核心目标词个数， B 表示识别到的所有核心目标词个数， C 表示实验的句子总数。在式(2)中， P_i 表示使用第 i 份语料得到的准确率。由于基于分类

模型的核心目标词识别实验中,采用3折交叉验证,因此 n 的取值为3。

3.3 实验结果

本文将目标词识别看作是一个分类问题,采用最大熵模型进行实验。本节实验用张乐的最大熵工具包^[9]和表2中的特征模板进行目标词识别。特种模板中特征的抽取使用的是哈尔滨工业大学社会计算与信息检索研究中心的语言处理集成平台LTP^[10]。

表2 特征模板及结果

Tab.2 Feature template and results

特征模板编号	特征组合	平均准确率/%
T1	TGT_LEMMA+TGT_POS+C_LEMMA+C_POS	77.75
T2	T1+S_SBJ	77.83
T3	T1+S_PRE	77.58
T4	T1+S_OBJ	77.75
T5	T2+T3+T4	82.55

为了验证窗口大小和最大熵模型的迭代次数对核心目标词识别性能的影响,使用表2中的T1模板,窗口大小分别设置为 $[-1,1]$ 、 $[-2,2]$ 和 $[-3,3]$,迭代次数设置范围为 $[50,500]$,相邻迭代次数相差50,并与工具包中的默认参数30做对比。具体的结果如图3所示。

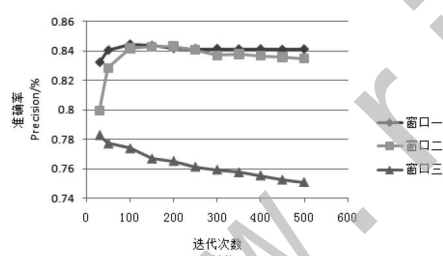


图3 迭代次数和窗口大小的散点图

Fig.3 Scatter plots of iteration times and window size

从图3中可以看出,窗口大小为 $[-1,1]$ 的时候识别性能最好,且迭代次数为100的时候,实验结果基本趋于稳定。因此,在后续的实验中,窗口大小设定为 $[-1,1]$,最大熵模型的迭代次数设定为100。

上述实验证明,采用句法结构分析的方法抽取句子特征,能够达到较高的识别率。但由于采用的训练语料比较小,限制了模型的性能。同时,从实验结果可以看出,系统对于单一的主干成分,如句子的主语,识别的效果不是很好,主要原因是单一的句子主干成分不能完整地表示整个句子的语义。

4 结论(Conclusion)

利用句法结构分析得到的特征识别句子中的目标词,

实验证明是有效可行的。在训练语料较小的情况下,模型取得了比较高的识别准确率。由于单个主干成分在句子中起到的作用不大,因此识别性能不显著。当主干成分与基本特征组合识别目标词时,句子主干能较完整地表示整个句子的语义,因此识别效果较好。在以后的研究中,将不断地扩充质量较高的语料,改善数据稀疏,丰富例句库。同时,将尝试加入语义特征,提升目标词识别的性能。

参考文献(References)

- [1] Dipanjan Das, Desai Chen. Frame-Semantic Parsing[J]. Computational Linguistics, 2014, 40(1): 9-56.
- [2] A Lakhfif, MT Laskri. A Frame-Based Approach for Capturing Semantics from Arabic Text for Text-to-Sign Language MT[J]. International Journal of Speech Technology, 2016, 19(2): 1-26.
- [3] AO Bayer, G Riccadri. Semantic Language Models with Deep Neural Networks[J]. Computer Speech & Language, 2016, 40: 1-22.
- [4] RICHARD Johansson, PIERRE Nuges. LTH: Semantic Structure Extraction Using Nonprojective Dependency Trees[C]. Proceedings of the 4th International Workshop on Semantic Evaluation. Stroudsburg: Association for Computational Linguistics, 2007: 227-230.
- [5] DIPANJAN Das, et al. Probabilistic Frame-Semantic Parsing[C]. Human Language Technologies: the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2010: 948-956.
- [6] LIN Dekang. Automatic Retrieval and Clustering of Similar Words[C]. Proceedings of the 17th International Conference on Computational Linguistics—Volume 2. Stroudsburg: Association for Computational Linguistics, 1998: 768-774.
- [7] 陈亚东, 等. 面向框架语义分析的目标词自动识别方法[J]. 山东大学学报(理学版), 2015(7): 45-53.
- [8] 石佼, 李茹, 王智强. 汉语核心框架语义分析[J]. 中文信息学报, 2014(6): 48-55.
- [9] 刘开瑛. 汉语框架语义网构建及其应用技术研究[J]. 中文信息学报, 2011(25): 46-53.
- [10] 张乐. 最大熵工具包 MaxEnt (2004). http://homepage.inf.ed.ac.uk/s0450736/maxent_toolkit.html.

作者简介:

张立凡(1987-), 女, 硕士, 助教. 研究领域: 中文信息处理, 人工智能与模式识别.