第 40 卷 第 10 期 计 算 机 工 程 2014 年 10 月
Vol. 40 No. 10 Computer Engineering October 2014

·人工智能及识别技术· 文章编号：1000-3428(2014)10-0150-05 文献标识码：A 中图分类号：TP391

基于 URL 的中文多语义名词在线语义标注

刘一正，杨 静，李 强

（华东师范大学计算机科学技术系，上海 200241）

摘 要：中文语义标注在自然语言处理领域有广泛的应用，其目的在于挖掘并标注出中文多语义名词的多个语义。提出一种新颖的语义标注算法，通过在线 URL 分类目录，构建得到 URL 分类器。借助于 URL 分类器，对搜索引擎返回的多语义名词的搜索结果（包括网页 URL 及摘要）进行分类，得到多语义名词的初始语义分类结果。对初始语义分类结果按其网页摘要聚类，提取聚类特征后得到多语义词的语义标注结果。该算法利用基于 URL 的网页分类方法，能在线对中文多语义名词进行语义标注。实验结果证明，该语义标注算法可以取得 70% 的准确率及 80% 的召回率，适用于网络热词语义标注。

关键词：语义标注；自然语言处理；中文多语义名词；URL 分类器；文本聚类；热词

中文引用格式：刘一正，杨 静，李 强. 基于 URL 的中文多语义名词在线语义标注[J]. 计算机工程，2014，40(10)：150-154.

英文引用格式：Liu Yizheng，Yang Jing，Li Qiang. Online Semantic Annotation of Chinese Multi-semantic Nouns Based on URL[J]. Computer Engineering，2014，40(10)：150-154.

Online Semantic Annotation of Chinese Multi-semantic Nouns Based on URL

LIU Yi-zheng，YANG Jing，LI Qiang

（Department of Computer Science and Technology，East China Normal University，Shanghai 200241，China）

【Abstract】 Chinese semantic annotation aims to find out the senses of a multi-semantic word，which is widely applied in natural language processing. This paper proposes a novel approach for semantic annotation of multi-semantic Chinese nouns. Given a multi-semantic Chinese noun，the proposed method can get its senses. The procedure is described as follows. The search results including URLs and abstracts of related Web pages are acquired through a search engine. The initial semantic classes are automatically generated by an online URL classifier using information gotten at the first step. Clustering algorithms are adopted to make full use of the Web page abstracts to get the final semantic classes. Experimental results demonstrate the proposed novel approach can obtain a considerable precision and recall rate with little manual intervention.

【Key words】 semantic annotation；natural language processing；Chinese multi-semantic noun；URL classifier；text clustering；hot word

DOI：10.3969/j.issn.1000-3428.2014.10.029

1 概述

语义知识学习在人工智能领域中具有重要应用，一直以来都是自然语言处理研究中的热点问题。语义标注研究多语义词的语义信息获取，在相关度计算、查询扩展等领域得到广泛应用[1]。名词的多语义特征表现尤为明显，所以它是语义标注的研究重点。对于中文多语义名词，语义标注应能较为全面地标注出其最新语义。例如对于多语义名词“苹果”，语义标注应能标注出其包括“公司”、“水果”、“电影”等在内的多个语义。

传统的语义分类方法大多仅涉及文本或 html 文件的语义信息处理，通过对网页正文或语料文本进行词法或语法分析，标注出多语义词的多个语义。

由于涉及到分词、词法分析及语法分析等文本处理步骤，传统的语义标注效率较低。对于在线语义标注，网页下载耗时，使得标注过程尤为缓慢。因此，以往的语义标注方法并不高效。文献[2]利用

基金项目：上海市国际科技合作基金资助项目(11530700300)；上海市科委科研基金资助项目“面向 NGB 的智能业务分析关键技术研究及系统研制”(12dz1500205)。

作者简介：刘一正(1990－)，女，硕士研究生，主研方向：自然语言处理；杨 静，副教授；李 强，博士。

收稿日期：2013-07-25 **修回日期**：2013-09-19 **E-mail**：lyzheng2011@163.com

SVM 模型对日语多语义词进行语义标注,该方法针对某些多语义词准确率可达 90% ,然而针对全部实验数据集,平均准确率只有 60% 左右,并且 SVM 模型的运用使得该方法的语义标注过程较为耗时,效率低。

与在线获取网页 html 文件相比,获取网页 URL 速度较快。基于网页 URL 特征的分类方法已在网页主题分类及查询分类领域得到广泛应用。然而,基于 URL 的分类方法还从未用于语义标注。本文利用基于 URL 的分类方法得到多语义词搜索结果的初始语义分类,即根据其 URL 对搜索结果按语义分类;对初始语义分类的网页摘要进行聚类,最终得到中文多语义词的语义标注结果。最终语义标注结果由一组与该语义相关的标签表示。

2　相关工作

语义标注方法主要可以划分为 3 大类:基于模板的方法,基于主题模型的方法以及基于百科的方法。对基于模板的方法,模板主要从文本集[3]或网页 html 文件[4]中训练得到,该方法常用于在线语义标注,效率较高,但召回率较低。基于主题模型的方法中常用到的模型有 LSA 模型及 LDA 模型[5],还产生了一些针对语义标注的新模型,如文献[6-7]在 LDA 中加入一个标签层。基于主题模型的方法准确率及召回率较高,但语义标注过程较为耗时。第 3 种方法基于在线百科[8],通过解析在线百科的语义信息得到多语义词的语义标注结果。这种方法高度依赖于在线百科,不能标注出在线百科未收录的语义信息。已有研究表明,网页文本及其对应的 URL 间存在内在的语义联系。文献[9-10]表明,网页 URL 字符包含其对应网页的部分语义信息。因此,基于 URL 特征的网页分类方法应运而生。不同于基于内容的分类方法,该方法通过解析 URL 字符的语义信息对网页进行分类[11-13]。本文将利用此方法对多语义词搜索结果进行初步语义分类。

3　中文多语义名词的语义标注方法

搜索引擎的在线搜索结果能高度反映出多语义词的语义信息。因此,本文将在线搜索引擎[14]返回的多语义词搜索结果作为语义标注原材料。对于返回的搜索结果,首先将其投入由在线网址分类目录构建的 URL 分类器,得到多语义词的初始语义分类。每个初始语义分类包括此类别下的网页 URL 及对应的网页摘要。随后,对初始语义分类中的网页摘要进行聚类,并从聚得的每类中抽取特征词,得到最终的语义标注结果。此过程可描述如图 1 所示。

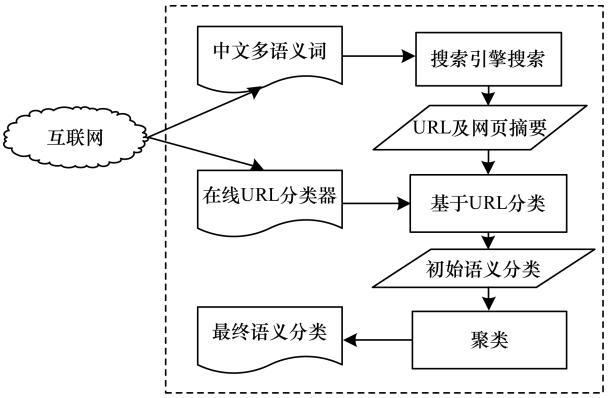

图 1　基于 URL 的中文多语义词在线语义标注过程

3.1　分类器

网页 URL 蕴含其对应网页正文的语义信息。根据多语义词搜索结果的 URL,对其进行初始语义分类。

一些中文权威网站发布或更新 URL 网址分类目录,此目录会对其收录的网站 URL 按语义类别归类。通常情况下,目录为树状结构,树中叶子结点即为其收录的网站主页 URL,非叶子结点则为其子结点的语义类别。雅虎网页目录含 3 层 ~4 层,首层对应 16 个语义类别,如图 2 所示。一些中文权威网站会发布与雅虎目录类似的网页目录,本文将以多个权威网站的网页目录为基础,构建 URL 分类器。

生活服务	女性	彩票	汽车	电影票	股票	地图	母婴	查询	天气	美食	健康	
购物	京东	唯品会	凡客诚品	折800	梦芭莎	寺库奢侈品	1号店	苏宁易购				更多 »
彩票	中彩网	中国体彩网	彩票大赢家	爱彩网	快3	11选5						更多 »
旅游	携程旅行网	艺龙订酒店	同程机票	游多多住宿	云南旅游网	去哪儿网						更多 »
生活	赶集网	美食天下	支付宝缴费助手	12306	好药师网上药店							更多 »
女性	瑞丽女人	性感女人	女人世界	蘑菇街	太平洋女性							更多 »
团购	美团网	满座网	拉手网	大众点评团	1号店团购	聚美优品						更多 »
手机	太平洋手机	新浪手机	天极手机	机锋网	移动营业厅							更多 »
汽车	爱卡汽车	易车网	太平洋汽车	搜狐汽车	汽车之家							更多 »
银行	工商银行	招商银行	支付宝	农业银行	建设银行	中国银行	交通银行					更多 »

图 2　Yahoo 在线网址分类目录的部分分类

URL 分类器的训练过程如下:

(1) 将网页 URL 按标识符分段,提取分类特征。

(2) 将提取得来的 URL 分类特征同 URL 目录

下的叶子结点进行相似度匹配,若相似度超过阈值,则将此叶子结点的语义分类作为该网页的候选语义分类。若未达到阈值,则不做任何处理。

(3)按照一定的选择策略,为每个网页从候选语义分类中确定其语义分类。多语义词不同语义下的搜索结果对应的网址分类往往不同。如“苹果”有3个语义:水果,数码产品品牌及公司。使用上文提到的URL分类器,可以将“苹果”的搜索结果按语义主要划分为3类。各类所占百分比如表1所示,搜索结果取百度的前100条搜索结果。

表1 多语义词搜索结果在不同语义类别下的比例 %

类别	比例
美食	35.80
IT	39.60
股市	15.90
其他	8.96

从表1中可以看出:利用URL分类器可以把苹果的搜索结果分为3类(美食、IT和股市),其中,“美食”对应苹果所具有的“水果”语义;“IT”对应“数码产品”;“股市”则对应“公司”语义。由此可以看出,URL分类器能有效地对多语义词的搜索结果按其语义类别进行分类,且大部分类别能有效对应该多义词的某一语义。

然而,通过对单语义词搜索结果的考察,发现多个URL分类可能对应同一语义,如表2所示。“桔子”的搜索结果可以被划分为2类(美食、健康),但是它们都表示同一种语义,即水果。这种情况在多语义词上也有发生,如多语义词“小米”的其中一个语义——“粮食作物”对应的搜索结果同时分布在“健康”、“美食”2个URL分类下,即多个URL分类对应多义词的同一语义。本文通过对网页摘要聚类来解决这个问题。

表2 单语义词搜索结果在不同语义类别下的比例 %

类别	比例
美食	49.7
健康	32.1
其他	18.2

3.2 网页摘要聚类

搜索引擎中返回的多语义词搜索结果含有噪音,可能会影响聚类结果的准确度。因此,在聚类前,必须对网页摘要进行清洗。首先,采用一种基于统计的方法过滤掉含搜索结果条数过少的初始语义分类。然后,对过滤后的初始语义分类结果分词后,再进行去停用词处理。

在文本清洗过程完成后,就对网页摘要进行聚类。本文采用基于词频的方法得到初始分类结果的聚类特征,对传统的基于词频的方法进行改进,得到最能代表初始分类结果的聚类特征。定义 W 为:

$$W = \frac{TF}{Cnt}$$

其中,TF 为某一初始语义分类的词频数;Cnt 为某候选特征词在对应的初始语义分类下的词频数。在聚类过程中,考察不同分类的网页摘要的文本相似度,具有较高文本相似度的网页摘要的初始语义分类将归为同一语义。本文将采用2种聚类方法考察不同聚类方法对实验结果的影响。

4 实验与分析

4.1 实验设定

4.1.1 评估标准

在实验中,根据百度百科、互动百科等中文权威在线百科构建了一个多语义词知识库,对于一些网络热词新出现的语义,如果在线百科还未收录,则手工加以补充,以保证该知识库的完备性。该知识库较为精准可靠,可作为一个有效的实验评估标准。实验中将以此知识库为基准,计算语义标注结果的准确率及召回率。表3为知识库的多语义词语义示例。

表3 知识库中的一些多语义词分类示例

多语义词	语义数	知识库语义分类
围脖	2	微博 围巾的一种
糯米	2	一种米 团购网站
…	…	…

对每个多语义词,c 为用本文方法进行实验标注到的语义数,c_1 为 c 中正确的语义数,c_0 为知识库中该多语义词的语义数。实验的准确率 P、召回率 R 和 F-值 F 定义如下:

准确率:$P = \frac{c_1}{c}$;

召回率:$R = \frac{c_1}{c_0}$;

F-值:$F = \frac{2PR}{P+R}$。

4.1.2 URL分类器

实验中使用基于3大中文权威网站(百度、搜狗、雅虎中国)的网址分类目录构建URL分类器,且保证分类器中的URL目录与这些在线目录保持同步更新。网页URL的特征提取方法如3.1节所述。

4.1.3　聚类算法

为了考察不同聚类算法对实验结果的影响,实验中将使用2 种不同的聚类算法,即 MKCLS 聚类和 Single-link 聚类,下面对这 2 种算法进行简单介绍。

(1)MKCLS 聚类:MKCLS 算法使用最大似然估计来训练词类,适用于处理语言模型或统计翻译模型。本文使用开源版本。

(2)Single-link 聚类:LingPipe 是一套常用的文本处理工具包,其中包括聚类、主题分类及命名实体识别等功能。Single-link 聚类是其中一种使用贪心策略的聚类方法。

在实验中,本文将分别采用这 2 种聚类算法实现中文多义词的语义标注,分别考察不同聚类方法下的实验效果,以考察聚类算法对实验结果的影响。

4.2　实验结果

4.2.1　基准实验

从实时在线得到的 500 个搜索热词中,任意选取 100 个词,作为基准实验的数据集。实验数据来自 2013 年 4 月 6 日的百度搜索结果。借助于 URL 分类器,可以得到初始语义的分类结果。例如,针对“围脖”这个多语义词,它的网络搜索结果可以被 URL 分类器分为 2 个类别,分别对应其 2 个不同的语义,每个类别下包含若干条网页信息(包括网页 URL 和网页摘要)。第 1 个类别中的网页信息有“围巾行情价格评价正品行货韩版纯色杂线…&http://www.360buy.com/products/1315-1…、多图单品,女装,服饰搭配购买美丽说狐狸…&http://www.meilishuo.com/attr/show/34…等等”,其中 & 前表示的该网页的摘要信息,& 后表示的是该网页的 URL。第 2 个类别中的网页信息有“南都周刊围脖女王姚晨的幸运与惊慌互联…&http://tech.sina.com.cn/i/2009-12-31/0945…、今天你围上围脖了吗互联网科技时代新浪网…&http://tech.sina.com.cn/i/2010-02-03/0745…等等”。

对初始语义分类结果进行聚类后,可以得到中文多语义名词的最终语义分类。对“围脖”的初始语义分类结果进行聚类后得到其最终语义分类,这里采用的聚类算法为 MKCLS 聚类。其结果包括 2 个语义类别:第 1 个语义类别中包括“围巾、时尚、价格、品牌、购物、评论”等语义词;第 2 个语义类别中包括“时代、科技、女王、周刊、成为、新浪网”等语义词。

多语义词的每个语义由一组与此语义高度相关的标签表示,如上面的“围巾、价格”等词就是“围脖”的第 1 个语义的标签。标签从聚类的特征词中产生,以每个特征词的 W 值为衡量标准,采用 top-N 选择策略。基准实验的实验结果如表 4 所示。

表 4　基准实验结果　%

评估指标	数值
准确率 P	62.4
召回率 R	58.2
F – 值	60.2

4.2.2　不同聚类算法及数据集下的语义标注结果

本文探索了不同数据集及不同聚类算法对实验结果的影响。不同于基准数据集 Dataset,Dataset-imp 选取了前 100 个搜索热度最高的多语义词作为数据集。实验结果如表 5 所示。

表 5　不同实验设定下的实验结果　%

方法	数据集	聚类算法	P	R	F
方法 1	Dataset	MKCLS	62.4	58.2	60.2
方法 2	Dataset	Single-link	68.1	60.2	63.9
方法 3	Dataset-imp	MKCLS	71.2	84.5	77.3
方法 4	Dataset-imp	Single-link	76.3	83.1	79.5

从表 5 中不难看出,聚类算法并不是影响实验结果的关键因素。当数据集相同时,使用不同的聚类算法得到了类似的实验结果。然而,数据集的选择策略则对实验结果有较大影响,显然,在 Dataset-imp 上可以得到更好的实验结果。对于热搜词,搜索引擎返回的搜索结果更为丰富有效,能提供更健壮的语义信息,能得到更高的准确率及召回率。从这一点也可以看出,本文方法更适合热词语义标注,具有较好的实时性,这一点正是在线百科及其他语义标注方法所欠缺的。

4.2.3　与其他语义标注算法的实验对比

实验还将本文方法(即方法 4)在同一数据集下(即数据集 Dataset-imp)与基于模板、基于百科的语义标注方法分别从准确率、召回率及方法能标注的多义词比率进行了对比,实验结果如表 6 所示。

表 6　不同语义标注算法对比　%

方法	P	R	覆盖率
方法 4	76.3	83.1	100.0
基于模板	62.1	45.2	100.0
基于百科	100.0	94.5	67.1

从表 6 可以看出,本文方法能对在线百科尚未收录的多语义词进行语义标注,且保证较高的准确率及召回率。

4.2.4　错误分析

实验过程中,主要有 2 种类型的错误,即语义标注结果漏掉某些语义、由聚类算法引起的错误。表 7 列出了这些错误的原因及其对应的百分比。

表 7 错误类型及原因分析 %

错误类型	错误原因	比例
丢失语义	互联网上无此语义	28.33
	URL 分类器不能解析网页	22.70
聚类错误	多个同一语义下的初始语义分类未能归为一类	26.31
	不同语义下的初始语义分类归为一类	12.56
其他错误	分词错误、网页摘要表述不当、URL 分类错误等	10.10

5 结束语

本文提出了一种新颖的中文多语义名词的语义标注算法,将基于 URL 的网页分类方法引入到中文多语义词的语义标注中。实验证明,该算法能得到多语义词的语义标注结果,且保证较高准确率及召回率。今后的研究主要集中在 2 个方面:(1)进一步研究多语义词的语义数量对实验结果的影响;(2)研究非名词的多义词语义标注方法。

参考文献

[1] Gabrilovich E, Markovitch S. Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis[C]//Proc. of International Joint Conference on Artificial Intelligence. Hydrabad, India: [s. n.], 2007: 1606-1611.

[2] De Saeger S, Kazama J, Torisawa K, et al. A Web Service for Automatic Word Class Acquisition[C]//Proc. of the 3rd International Universal Communication Symposium. Tokyo, Japan: ACM Press, 2009: 132-138.

[3] Pasca M. Acquisition of Categorized Named Entities for Web Search[C]//Proc. of the 13th ACM International Conference on Information and knowledge Management. Washington D. C., USA: ACM Press, 2004: 137-145.

[4] Shi Shuming, Liu Xiaokang, Wen Jirong. Pattern-based Semantic Class Discovery with Multi-membership Support[C]//Proc. of the 17th ACM Conference on Information and Knowledge Management. Napa Valley, USA: ACM Press, 2008: 1453-1454.

[5] Arora R, Ravindran B. Latent Dirichlet Allocation Based Multi-document Summarization[C]//Proc. of the 2nd Workshop on Analytics for Noisy Unstructured Text Data. Singapore: ACM Press, 2008: 91-97.

[6] Li Fang, Shen Huiyu, He Tingting. Tag-topic Model for Semantic Knowledge Acquisition from Blogs[C]//Proc. of the 7th International Conference on Natural Language Processing and Knowledge Engineering. [S. l.]: IEEE Press, 2011: 221-226.

[7] 何婷婷,李 芳. 基于主题模型的博客标签语义知识获取[J]. 中国通信,2012,9(3):38-48.

[8] Liu Yang, He Tingting, Tu Xinhui, et al. Obtaining Chinese Semantic Knowledge from Online Encyclopedia[C]//Proc. of International Conference on Natural Language Processing and Knowledge Engineering. [S. l.]: IEEE Press, 2010: 1-7.

[9] Baykan E, Henzinger M, Marian L, et al. A Comprehensive Study of Features and Algorithms for URL-based Topic Classification[J]. ACM Transactions on the Web, 2011, 5(3).

[10] Devi M I, Rajaram D R, Selvakuberan K. Machine Learning Techniques for Automated Web Page Classification Using URL Features[C]//Proc. of International Conference on Computational Intelligence and Multimedia Applications. [S. l.]: IEEE Press, 2007: 116-120.

[11] Baykan E, Henzinger M, Marian L, et al. Purely URL-based Topic Classification[C]//Proc. of the 18th International Conference on World Wide Web. [S. l.]: ACM Press, 2009: 1109-1110.

[12] 张 宇,宋 巍,刘 挺,等. 基于 URL 主题的查询分类方法[J]. 计算机研究与发展, 2012, 49(6): 1298-1305.

[13] 张 宇,宋 巍,谢毓彬,等. 利用 URL 类别改进查询主题分类[C]//第六届全国信息检索学术会议论文集. 哈尔滨:哈尔滨工业大学出版社,2010:157-166.

编辑 顾逸斐

(上接第 149 页)

[7] 康轶非. 平方根容积卡尔曼滤波在移动机器人 SLAM 中的应用[J]. 机器人,2013,35(2):186-193.

[8] 张国良,汤文俊. 基于线段特征匹配的 EKF-SLAM 算法[J]. 控制工程,2012,10(6):119-124.

[9] Wang X, Zhang H. A UPF-UKF Framework for SLAM[C]//Proceedings of IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2007: 1664-1669.

[10] Kim C, Sakthivel R, Chung W K. Unscented Fast SLAM: A Robust and Efficient Solution to the SLAM Problem[J]. IEEE Transactions on Robotics, 2008, 24(4): 808-820.

[11] Shojaie K, Shahri A M. Iterated Unscented SLAM Algorithm for Nacigation of an Autonomous Mobile Robot[C]//Proceedings of IEEE/RSJ Internation Conference on Intelligent Robots and Systems. Piscataway, USA: IEEE Press, 2008: 1582-1587.

[12] Julier S J, Uhlmann J K. Unscented Filtering and Nonlinear Estimation[J]. Proceedings of the IEEE, 2004, 92(3): 401-422.

[13] Julier S J, Uhlmann J K. A Consistant, Debiased Method for Converting Between Polar and Cartesian Coordinate Systems[C]//Proceedings of the 11th International Symposium on Aerospace Defense Sensing, Simulation and Controls. Orlando, USA: Press, 1997: 110-121.

[14] Mohamed A H, Schwarz K P. Adaptive Kalman Filtering for INS/GPS[J]. Journal of Geodesy, 1999, 73(4): 193-203.

[15] Australian Centre for Field Robotics. Source Code[EB/OL]. (2008-06-10). http://www-personal.acfr.usyd.edu.au/tbailey/.

编辑 索书志