文章编号：1003-0077(2010)01-0048-06

中文核心领域本体构建的一种改进方法

谌贻荣，陆 勤，李文捷，崔高颖

（香港理工大学 计算学系，香港 666666）

摘 要：核心本体对最基本的领域知识建模，并在上位本体和领域本体之间建立联系。上位本体是领域无关的而核心本体是领域相关的，因此在自动创建中文核心本体过程中，映射中文核心术语到上位本体概念有很多的错误。本文提出的改进方法首先找到共享后缀术语集内被共享的术语条数更多、与各术语的意义更接近的上位概念；然后用其来改进词集中的核心术语和概念之间的映射。实验证明，该方法有效的提高了核心本体自动创建的精确度。

关键词：计算机应用；中文信息处理；本体构建；领域核心本体；上位本体；领域本体；上位关系

中图分类号：TP391 文献标识码：A

An Improve Method for Chinese Core Ontology Construction

CHEN Yirong, LU Qin, LI Wenjie, CUI Gaoying

(Department of Computing, the Hong Kong Polytechnic University, Hong Kong, China)

Abstract: A core ontology models fundamental domain knowledge and bridges the gap between an upper ontology and a domain ontology. Since the upper ontology is domain independent, many errors are introduced when mapping core terms to the upper ontology concepts in automatic Chinese core ontology construction. This paper proposes an extraction method making use of terms sharing the same suffixes to find the hypernyms: the term that is more frequently shared by other terms and are closer in meanings to those terms. These hypernyms are then used to improve the mapping of these terms to the correct concepts. Experiments show that a significant improvement is achieved in terms of accuracy for core ontology construction.

Key words: computer application; Chinese information processing; ontology construction; core ontology; upper ontology; domain ontology; hypernymy

1 引言

因特网的飞速发展带来了海量的信息，但如何有效地利用这些信息，成为一个新技术所带来的新问题。一个现今被广泛研究的解决这个问题的方法就是用本体对领域建模。在信息科学领域，本体是一个形式化的，有明确描述的共享的概念化模型[1]。领域本体可以广泛服务于各种信息应用，如信息检索、信息抽取、摘要和问答系统等。为满足各种基于知识信息的不同领域应用的需求，快速准确地建立和更新领域本体意义重大。

本体作为一个概念化模型，有多种分类方法。按照模型的复杂程度可以分为轻量型的本体和重量型的本体。轻量型的本体只包含概念和关系，重量型的本体还包含了公理和推导系统。按照本体建模的范围差别，又可以分为上位本体[2]、核心领域本体和领域本体。上位本体对通用的概念建模。一个比较著名的上位本体就是 SUMO（Suggested Upper Merged Ontology）[3-4]。领域本体对特定领

收稿日期：2009-05-24 定稿日期：2009-11-06

基金项目：香港理工大学教育资助委员会（UGC），研究资助局（RGC）角逐研究用途补助金资助项目(CERG)(PolyU 5225/05E, PolyU 5190/04E, PolyU 5246/08E)

作者简介：谌贻荣(1978—)，男，博士生，主要研究方向为本体学习，术语提取；陆勤(1960—)，女，教授，主要研究方向为中文信息系统和自然语言处理；李文捷(1966—)，女，助教，主要研究方向为自然语言处理。

域建模。而核心领域本体是对领域中的核心概念建模，并作为一个中间层本体，为上位本体中的抽象概念和领域本体中的特定概念建立联系。概念大多用自然语言词汇来表达，然而词汇和概念的对应并不是一对一的。词汇及新增词汇和概念之间的多对多映射关系，就构成了词汇本体。典型的词汇本体有 WordNet[5-6]，HowNet[7]，SinicaBOW[8] 等。

为了快速从中文文本和已有资源中有效地学习领域本体，我们所要自动创建的目标本体是重量型的领域核心概念本体。由于中文语义资源有限，我们的实现方法是从中文领域术语库中先提取核心术语，然后使用双语术语库把核心术语通过英文 WordNet 的同义词词义映射到一个上位本体中，最后继承上位本体中和领域相关的概念、关系以及公理，从而构建中文领域的核心概念本体，简称核心本体。核心本体也是领域本体半自动创建过程中人工介入的关键点。自动创建的核心本体经过人工的整理，可以以较小的代价大大提高领域本体到上位本体的映射质量。

前人在核心本体自动建构上做的研究工作并不太多。大多数的核心本体是手工创建的[8-10]。在中文核心本体建构方面，由于中文本体资源的相对匮乏，工作更少[11]。前人研究比较少的原因至少包括几点：一是核心本体首先要求存在一个上位本体作为基础；二是如何界定领域中的核心概念并没有定论；三是核心本体的建构要求存在联系上位本体和领域本体之间的相关数据资源。

核心本体建构的一个巨大问题是上位本体是领域无关的，而核心本体是领域相关的并且其中的核心概念要映射到上位本体中。如何有效地利用领域信息来解决这个问题比较关键。我们以往的工作就是针对这一问题，并在一定程度上解决了这个问题[12-13]。这篇论文中我们将采用共享后缀词集特性来进一步提高性能。崔[14] 在其研究报告中指出超过 90％的中文术语词汇的中心构件位于后缀位置。由此我们提出基于共享后缀词的抽取来找到与中心构件相对应的词义，并基于该词义来改进相应词的词义映射。由于核心术语抽取算法和核心本体建构算法（Core Ontology Construction Algorithm，缩写为 COCA）[12-13] 是本文改进算法的基础。所以将在第 2、3 章分别定性介绍一下，详细的定量计算公式请参考原文。第 4 章具体给出基于共享后缀词集的算法，第 5 章用实验来验证方法的有效性并对实验结果作分析。第 6 章总结全文。

2　核心术语抽取算法

核心术语是术语库中能产性高，领域特定的术语。能产性高的术语可以作为术语构件在更多的术语中采用。能产性高的核心概念构成的核心本体更能发挥其作为领域本体和上位本体的中间层的作用。谌在文献[15] 的工作中提出了核心术语的抽取算法，本文的在该算法的基础上进行改进。该算法首先做后向最大词典切分，然后用词频排名做领域性过滤。

所谓后向最大词典切分算法，其输入是词典，被切分的对象是词条，输出的是切分后的词典，切分方法是以输入的词典为切分词典，同时对该词典的每一词条切分之前，暂时在切分词典中去掉当前被切分的词条，然后反向最大切分当前词条。这样就保证了词典的每一个多字词条都会被切成更小的词段，这些被切分的词段就是当前术语的最大术语构件。从另一方面看，由于是最大切分，就避免了父串对子串的频率叠加效应，有效地去除了构件嵌套产生的短构件淹没长构件的效应。比如，“计算机”内嵌了“机”，如果不采用最大切分法的话，就会造成构件“机”的频度排名更靠前。而实际上在 IT 领域，“计算机”作为构件直接合成术语比“机”更频繁，意义更明确。所以采用最大切分是必需的。

在切分后的词典里统计术语构件词频，按词频从高到底排名，就形成了一个术语构件词频表。另外取通用领域的词频表作为对照。一个术语词条，如果在领域中的排名比通用领域中的排名高出设定的一个阈值，将予以保留，否则删除。经过这两步后，一个核心术语词表就自动产生了，实验证明该列表的质量较高。

3　核心本体建构算法－COCA

核心本体建构算法是本文研究者之前用英文发表的工作[12-13]，英文名字是 Core Cotology Construction Algorithm，缩写为 COCA。该算法首先自动将中文核心术语映射到 WordNet 中的同义词词义，进而映射到上位概念继承上位本体的和领域相关的概念，关系以及公理。这个自动映射的算法分别有下面的三个子任务：1. 中文到英文的翻译（在中英文术语库的数据上实现）；2. 英文到同义词词义的消歧（在 WordNet 的数据上实现）；3. 找

同义词词义对应的上位概念(在上位本体SUMO和WordNet的映射数据上实现)。

第一个子任务实现的假设是,给定一个中文词,如果对应的英文越长,该中文和对应英文同时作为其他中英文术语对的子串的次数越多,那么取该英文作为翻译的概率越大。第二个子任务实现的假设是,给定英文词,如果该英文词在语料库中取一个同义词词义的频率越高,这个英文取该同义词词义的概率越大。第三个子任务是在数据中直接获得的,并且每个同义词词义只有一个对应的上位概念,所以不存在最优选择的问题。

COCA除了基本的三个子任务计算外,还运用了独立事件并集概率的框架方法来集成各种特性。所谓独立事件的并集概率就是多个独立的事件任意一个出现的概率,其公式表达如下:

$$\underset{x\in E}{U}=\begin{cases}0 & |E|=0\\ p(x)+\underset{y\in E^{(x)}}{U}(p)-p(x)\times\underset{y\in E^{(x)}}{U}(p) & |E|>0\end{cases}$$

这里E是独立事件集;$p(x)$是概率函数用于返回事件x的对应概率值。如果E是空集{},那么$U(p)=0$;如果E是集合$\{x\}$,那么$U(p)=p(x)$;如果E是集合$\{x,y\}$,那么$U(p)=p(x)+p(y)-p(x)p(y)$。

之前的工作中已经提出了三个用来提高性能的特性[13],其中包括(1)合成了多路特性的算法,(2)合成了下位词特性的算法,以及(3)合成了词性标记特性的算法。多路特性利用一个中文术语可能通过多个英文翻译映射到同一概念的现象,把多条路径叠加权重来计算中文术语到概念的映射权重。下位词特性利用一个词的下位词来改进该词的概念映射精准率。词性标记特性利用一个词经常取特定词性的偏向性来提高映射精准率。这些特性将和本文的基于共享后缀词集方法在同一数据集上作性能比较。

本文提出的共享后缀词集也是作为一个附加的特性在同样的框架下进行集成的。

4 基于共享后缀词集的核心本体建构改进算法—COCA_SE

很多共享后缀的中文词都会有着共同的上位概念。举例来说,"驱动器"对应的英文"driver"有歧义,一般表示一种"人"(human)——"司机",也可以表示一种"设备"(device)。而"驱动器"(driver)有很多共享后缀的词如"服务器"(server),"传感器"(sensor)等。在通用领域,这些词更多的是"人"的下位词,而在信息科学领域中,这些词都应该是"设备"的下位概念,并且都以"器"作为后缀。这个例子提示我们可以用后缀词来改进词义的正确映射,从而改进自动建构的核心本体的质量。下面的问题就是如何找到并利用重要的上位概念来改进下位词在特定领域中的词义映射。

重要的上位概念有两个方面的特性。一方面,在共享后缀词中一个上位概念的下位词越多,该上位概念就越重要。另一方面,上位概念越抽象,其对下位概念的辨别区分能力就越弱;也就是说,一个上位概念和下位概念之间距离越近,越具体,该上位概念越重要。基于以上两点,我们提出了基于共享后缀词的本体构建改进算法(COCA_SE)。该算法在已有的核心本体算法(COCA)框架上把共享后缀词特性集成在后处理模块中来提高性能。

对于每一个中文核心术语T_C,COCA_SE算法的处理方法如下:

输入:1)共享的后缀T_H和共享后缀词集,2)从COCA中得到的中文和候选概念之间的映射权重,3)WordNet中的概念继承结构

输出:调整后的中文和候选概念之间的映射权重

详细步骤:

(1)共享上位概念权重(Weight of Shared Hypernym,缩写为SHW)的计算如下:

$$SHW(S_H|T_H)=\sum_{T_C\in ext(T_H)}\ \sum_{\substack{s_i\in synset(T_C)\wedge\\ s_i\in hyponym(S_H)}}\frac{dep(S_H)}{dep(s_i)}COCA(s_i|T_C)$$

其中后缀T_H是输入的中文共享后缀词集所共享的后缀,同义词词义S_H是被共享的上位概念,函数$ext(T_H)$返回的是T_H的父串中所有术语的集合,函数$dep(s)$返回同义词词义s在WordNet概念继承中的继承深度,$synset(T_C)$返回中文词T_C的候选同义词词义集合。

上述公式利用了前面阐述的共享后缀词集中的重要上位概念的两个特性。对于第一个特性,该公式对词集中匹配的每个词求和,被更多词所共享的上位概念自然的会得到更高的权重。对于第二个特性,越抽象的概念其概念深度dep越低,最后算出来的权重也就越低;反之,越具体其深度越高,公式返回的权重也就越高。

(2)重要上位概念对术语取同义词词义的影响力(Weight Under Hypernym,缩写为WUH)计算

如下

$$WUH(S \mid T_C) = WUH(S \mid T_C, T_H) = \frac{\underset{S_i \in hpr(S)}{MAX} SHW(S_i \mid T_H)}{\sum_{S_j \in synset(T_C)} \underset{S_k \in hpr(S_j)}{MAX} SHW(S_k \mid T_H)}$$

其中函数 $hpr(s)$返回一个同义词词义 s 的全部上位概念构成的集合。

这个公式计算术语 T_C的一个候选词义 S 对应的最重要上位概念的正规化权重。这里最重要的上位概念就是权重最大的上位概念。公式中的分子部分计算权重最大的上位概念的权重，公式的分母部分计算全部候选概念的最大上位概念权重的和，这样公式 $WUH()$返回的值被正规化为一个百分比，其值介于 0 到 1 之间。

(3) 利用独立事件并集概率公式，集成重要上位概念权重（Core Ontology Construction Algorithm with Suffix Enhancement，缩写为 COCA_SE)的计算如下

$$COCA\_SE(S \mid T_C) = \underset{x \in \{COCA(S \mid T_C), WUH(S \mid T_C)\}}{U}(x)$$

其中函数 U 就是前面提及的独立事件并集概率公式。两个独立事件概率就是通过 *COCA* 算法得到的概率和通过 *WUH* 算法得到的概率，展开式子后这个概率就等于 $COCA(S \mid T_C) + WUH(S \mid T_C) - COCA(S \mid T_C) \times WUH(S \mid T_C)$。

5　实验和分析

核心本体创建算法首先从中文术语库中抽取核心术语，然后对应的英文词条映射每一个核心术语到最好的同义词词义候选，并通过该词义连接到上位本体概念。这些核心术语，对应的同义词词义，上位本体概念以及继承自词汇本体 WordNet，上位本体 SUMO 的各种关系和公理，就构成了自动创建的重量型中文领域词汇核心本体。在自动创建的过程中，为核心术语找到最好的词义和上位本体概念是关键的步骤，其精准率直接影响自动创建的核心本体的质量。因此算法的选择核心术语最佳词义的精准率和选择最佳对应上位概念的精准率将作为后面实验的性能指标。可以看到这两个精准率的要求都比较高，因为在一个领域中，核心术语还是有可能有几个相近的意义（比如“网络”在 IT 领域中既可以指电子通讯网络，也可以指抽象的由节点和节点间的边构成的一种数据结构），而这两个指标要求算法必须选择领域中最合适的才算正确。

首先核心术语抽取算法在来自北京大学计算语言学研究所的中英文双语 IT 领域术语库（Chinese and English IT Term Bank，缩写为 **CEITTBank**，包括大约 13 万 IT 领域中文术语）[16] 上自动抽取了 1 500 个能产性高的领域特定的核心术语[15]，这些核心术语作为术语构件大约覆盖了 50%的全部术语。为保证 95%置信度时 5%的误差范围[17]，两个 IT 领域的研究人员随机抽取了 400 个核心术语来分别人工制作并互相校验答案。

核心术语建构算法在映射词义和上位概念采用的数据源包括 CEITTBank WordNet 1.6 和 WordNet 1.6 与上位本体 SUMO 的映射数据[18]。之所以采用 WordNet 1.6 的数据是因为上位本体 SUMO 只在 WordNet 1.6 上有完全的映射，也就是每一个 WordNet 中的同义词词义都被赋予了一个上位本体概念。因为核心术语有时即使对 WordNet 的映射不是最好，但对 SUMO 的映射却仍然正确，例如“灵敏度”(sensitivity)，算法映射到生理上的“敏感度”，答案应是物理上的物理灵敏度，但到了上位本体都是一种能力(capability)，所以对这两个资源映射的评估要分别进行。

为了测试共享后缀词集的改进，我们引入一个基准算法，标记为 *B*。*B* 算法选择只使用通用领域词汇本体 WordNet 中的词汇词义频度（也就是语料库中某个词汇取某个词义的次数），通过选取最高频的词汇词义频度来选择最佳的词义和上位概念。前述第三节讲述的不加载任何其他特性只实现三个基本任务的算法标记为 *S*；共享后缀词集特性标记为 4；以前的论文中提出的其他三个特性[13]：多路特性，下位词特性和词性标记特性，分别标记为 1，2，3。

特性之间可以任意组合，但必须和基本算法结合在一起。由此，我们确定运行如下算法：基准算法 *B*，基本算法 *S*，合成了多路特性的算法 *S1*、合成了下位词特性的算法 *S2*、合成了词性标记特性的算法 *S3*，合成了共享后缀词集特性的算法 *S4*，合成了前三个特性的算法 *S123*，合成了全部特性的算法 *S1234*。以精准率(Accuracy)作为衡量性能的指标，测试结果如图 1 所示。

从图 1 可以看到，在同义词词义选择上，基于共享后缀词集的改进算法 *S4* 取得了最高的性能，*S4* 的精准率比基准算法 *B* 的精准率提高了 78.9%。

在上位概念选择上，*S4* 取得了次高的性能。也可以看到合成算法 *S1234* 在上位本体概念的选

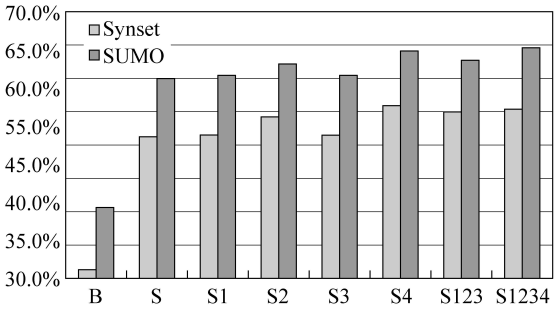

图1 各种特性组合的性能直方图

择上正确率稍高,但并不明显,和只用 *S4* 差别非常微小。采用了下位词特性算法 *S2* 的性能也是比较高的,这个方法主要使用了父串的术语集合来改进子串术语的映射性能。本文所述的基于共享后缀词方法实际上是方法 *S2* 的扩展版本,因为后缀词特性不仅改进了作为子串的术语的映射,同时也反过来改进了作为父串术语的映射。

通过对错误的分析可以发现问题主要有三种来源。第一就是通用词典的引入—WordNet。例如:"电阻"(resistance)总是会被错误的翻译成代表"反对你不赞同的事物的行为"的同义词词义"resistance, opposition"。在IT领域,正确的同义词词义应该是"材质对电的阻抗;单位是欧姆",其对应的上位本体概念是测量单位 UnitOfMeasure (SUMO 中的继承路径为"/实例-Entity /抽象物-Abstract /数量-Quantity /物理量-PhysicalQuantity /测量单位-UnitOfMeasure")。基于共享后缀词的改进算法在一定程度上解决了这类问题。例如,"驱动程序"(driver)在通用领域,如果不用共享后缀词特性的话会由于翻译是 driver 而错误的映射到同义词词义"car driver"(司机),用之后则被正确地映射到"driver program"(驱动程序)。第二个问题就是有些领域核心术语的词义并不存在于 WordNet 中。比如术语"多路存取"(multi-access)就是这样。在我们制作测试答案时发现大约4%的核心术语在 WordNet 中是找不到词义的,所以也就不能自动的被映射到 SUMO 中上位概念。第三种来源就是缺少上下文信息导致的翻译错误。这是因为双语术语库只是一个词典,缺少足够的上下文信息。这就导致术语的几个词义在领域中可能都正确,而没有语境信息无法判定到底那一个最恰当。图2是自动创建的中文核心本体的一个片段。最顶层的是 SUMO 中对应的上位本体概念,其下就是继承的核心概念。这个片段中显示的全部中文核心术语都正确的映射到了对应的概念,但如果使用基准算法就会有错误。比如"例程"(routine)会被对应到例行公事,而不是例行的计算机程序。

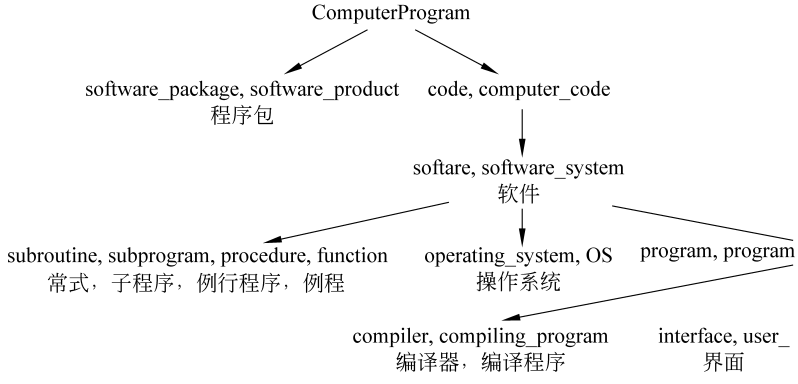

图2 自动创建的中文核心本体的一个片段

6 结束语

本文提出一种利用共享后缀词来改进核心本体自动创建的方法(COCA_SE)。它利用共享后缀词很可能也继承自相同的上位概念这一观察现象,找到最佳的上位概念,并利用上位概念调整原来的 COCA 算法从中文到英文最后到同义词词义的映射权重,更好地找到最恰当的同义词词义和对应的上位本体。实验证明,基于共享后缀词集的算法取得了最好的概念映射性能,提高了自动创建的核心本体的质量。本文中假定共享后缀词集的特性和其他特性不相关,但实际上关联是存在的,未来可以采用有指导(supervised)的方法来学习特性之间的融合参数来达到更高的性能。另外一方面,还可以集成更多的信息,如同义词词义的定义,互联网上的词汇语义资源(如维基百科等),进一步提高性能。

参考文献

[1] Studer, R., Benjamins, V. R. and Fensel, D. Knowledge engineering: Principles and methods [J]. Data & Knowledge Engineering, 1998, 25(1-2): 161-197.

[2] Navigli, R., Velardi, P. Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites [J]. Computational Linguistics 2004, 30: 151-179.

[3] Pease, A., Niles, I., Li, J. The Suggested Upper Merged Ontology: A Large Ontology for the Semantic Web and its Applications [C]// Working Notes of the AAAI-2002 Workshop on Ontologies and the Semantic Web. Palo Alto, California, USA: 2002, 28.

[4] Niles, I., Pease, A. Towards a standard upper ontology [C]// Proceedings of the international conference on Formal Ontology in Information Systems. Ogunquit, Maine, USA: 2001: 2-9.

[5] Miller, G., Beckwith, R., Fellbaum, C., Gross, D., Miller, K. Introduction to WordNet: An On-line Lexical Database * [J]. International Journal of Lexicography 1990, 3: 235-244.

[6] Fellbaum, C., NetLibrary, I. WordNet: an electronic lexical database [M]. USA: MIT Press, 1998.

[7] Dong, Z., Dong, Q. HowNet and the Computation of Meaning [M]. Singapore: World Scientific Publishing Co., 2006.

[8] Huang, C., Chang, R., Lee, S. Sinica BOW (Bilingual Ontological Wordnet): Integration of Bilingual WordNet and SUMO [C]// Proceedings of the 4th International Conference on Language Resources and Evaluation. Lisbon, Portugal: 2004: 26-28.

[9] Hirst, G. Ontology and the Lexicon [M]. Handbook on Ontologies, 2004.

[10] Doerr, M., Hunter, J., Lagoze, C. Towards a Core Ontology for Information Integration [J]. Journal of Digital Information 2003, 4: 169.

[11] Tang, A., Zhen, Z., Fan, J. Thesaurus-based Approach to Build Domain Ontology [J]. New Technology of Library and Information Service (in Chinese) 2005: 1-5.

[12] Chen, Y., Lu, Q., Li, W., Li, W., Ji, L., Cui, G. Automatic Construction of a Chinese Core Ontology from an English-Chinese Term Bank [C]// Proceedings of Workshop OntoLex07 From Text to Knowledge: The Lexicon/Ontology Interface, the 6th International Semantic Web Conference. Busan, Korea: 2007.

[13] Chen, Y., Lu, Q., Li, W., Cui, G. Chinese Core Ontology Construction from a Bilingual Term Bank [C]// Proceedings of the 6th Language Resources and Evaluation Conference (LREC2008). Marrakech, Morocco: 2008.

[14] Cui, G., Lu, Q., Li, W. Preliminary Chinese Term Classification for Ontology Construction [C]// Proceedings of the 6th Workshop on Asian Language Resources, in the Third International Joint Conference on Natural Lanugrage Processing (IJCNLP). Hyderabad, India: 2008.

[15] Ji, L. N., Lu, Q., L., Chen, Y.: Automatic Construction of a Core Lexicon for Specific Domain [C]// Proceeding of the 6th International Conference on Advanced Language Processing and Web Information Technology. Luoyang, China: 2007.

[16] Kang, W. and Sui, Z. F. Research on Automatic Chinese Multi-word Term Extraction Based on Term Component [C]// Proceedings of the 22nd International Conference of Computer Processing of Oriental Languages. Hong Kong: 2009: 57-67.

[17] Scheuren, F., Association, A. S. What is a Survey? [EB/OL]. 1997. http://www.amstat.org/sections/srms/whatsurvey.html. Published by American Statistical Association.

[18] Niles, I., Pease, A. Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology [C]// Proceedings of the IEEE International Conference on Information and Knowledge Engineering. Las Vegas, Nevada, USA, 2003: 412-416.