基于 WordNet 的本体澄清*)

郭 雷 方 俊 王晓东
(西北工业大学自动化学院 西安 710072)

摘 要 由于本体能够消除概念的混淆和重用知识,因此它的质量对于语义网技术的应用非常重要。为了提高本体的质量,很多的工作集中在概念建模,但是本体表示这个非常重要的方面一直被忽视。目前本体的表示使用的是词(term),但同一个词可能有很多不同的意思,这样在基于本体的应用时将导致不清楚或错误的理解。为了解决这个问题,使用定义在 WordNet 中的词义(sense)而不是词来作为本体的表示,其原因是词义只有唯一的意思。本体澄清的定义为利用目标词周围的本体元素和被它标注的文档附近的词,对目标词进行自动消歧的过程。通过计算目标词义和它的邻居词的语义相似度,语义相关度最大的词义将选为正确的词义。实验表明,我们的算法有很好的性能。与最好的消歧算法相比,概念(Concept)精度差不多是名词精度的 2 倍,关系(Property)精度差不多是动词精度的 3 倍。实验证明了我们的算法在半自动的本体净化过程中也是非常有效的。

关键词 本体澄清,语义相关度,消歧

Ontology Clarification by Using WordNet

GUO Lei FANG Jun WANG Xiao-dong
(College of Automation, Northwestern Polytechnical University, Xi'an 710072, China)

Abstract Semantic Web technology highly depends on the quality of ontology as it reduces or eliminates conceptual confusion and reuses knowledge. In order to enhance quality of ontology, a vast amount of research has focused on concept modeling task, but there is one major problem with lexical representation of ontology. Current lexical representation is term which may have different meanings; this can result in frustrating misunderstanding and ambiguity during the management and application of ontology. To solve this problem, sense is used to replace term as the lexical representation of concepts and properties for its unique meaning. We call ontology clarification the process of automatically disambiguating terms in ontology by using its surrounding ontology elements and its nearby terms in annotated documents using this ontology. The right sense is assigned to a target term by maximizing the relatedness between the target and its neighbors for semantic relatedness between them. Experiments show our ontology clarification method has good performance. Comparing with the best word sense disambiguation method, the concept precision is almost 2 times than the precision of noun, and the property precision is almost 3 times than the precision of verb. The last experiment proves that our method is also effective in a semi-automatic ontology clarification process.

Keywords Ontology clarification, Semantic relatedness, Word sense disambiguation

1 引言

本体包括概念的集合和概念的定义以及它们的关系,它对于语义网的实现有着非常重要的作用,因为本体能够实现智能 agent 之间在语义层次无歧义地分享信息。正是由于本体的关键性,使得语义网技术的应用对本体质量要求非常高。为了提高本体的质量,目前大部分的工作集中在怎样对知识进行建模[1,2],而忽视了本体词法表示这个同样很重要的问题。在当前的表示方法中,本体中的概念和关系通常使用词来表示它们的意义。但在很多情况下,词法层次(表示概念或关系的词)和意思层次(概念或关系的意义)的差别通常是非常明显的,如"mouse"能够表示老鼠或者是鼠标的意义。这种现象使得基于本体的应用导致错误。

为了解决这个问题,我们使用词义(sense)而不是词作为本体中概念和关系的词法表示,因为词义只有唯一的意思。在这篇文章中,用定义在词法数据库 WordNet[3] 中的词义作为本体的词法表示。WordNet 数据库将一组词义与一个词关联起来,我们把这组词义叫做同义词组(synets),同义词组中的词义通过在词上添加序号来表示。本体澄清定义为利用本体中的目标词的上下文确定其正确的语义解释(Semantic Interpretation)的过程。词义消歧(Word Sense Disambiguation)是指确定文档中的词语真实意义的过程[4]。这与我们对本体元素的消歧非常相似。它们最大的不同在于用来进行消歧的上下文是不一样的。词义消歧使用目标词附近的词来确定该词的意义。在我们的消歧算法中,使用两种类型的上下文对目标词进行消歧:一种是本体中目标词周围的概念和关系,另一种是本体标注的文档中该目标词标注的词附近的词。当使用第一种类型的上下文时,概念和关系之间的语义联系同样被考虑了。使用这两种上下文来确定词的语义解释的原因在于下面的假设。

*)本文获得国家自然科学基金资助项目(60675015)资助。**郭 雷** 博士生导师,主要从事神经网络、模式识别和知识管理等;**方 俊** 博士生,主要从事语义网和本体管理方面研究;**王晓东** 博士生,主要从事语义网和智能检索。

假设 本体中元素的语义解释能通过它附近的本体元素和标注的文档周围的词共同确定。

这个假设和我们的直觉一致，因为在同一本体或句子中一起出现的词具有某种语义联系。对于本体元素的消歧应考虑本体语义信息和本体应用信息。由该假设，正确的语义解释是和邻居词的语义相似度最大的目标词的词义。实验表明，我们的算法精度比现在最好的词义消歧算法要高。

下一节我们将介绍相关的工作，本体澄清算法将在第3节详细介绍，第4节是证明算法有效性的实验，最后是全文的总结和未来的工作。

2 相关工作

根据我们的知识，目前没有对本体中的词来进行消歧的工作。与我们的方法最相近的工作是用于领域本体创建的OntoLearn[5]。在OntoLearn中，为了发现本体候选词的真实意义，使用了消歧的技术，而我们的工作是对于已经存在的本体进行消歧的处理。

3 本体澄清

本体澄清的目标是自动确定本体元素在WordNet中的正确词义。这个过程中，将考虑本体中的语义信息和标注信息。在这篇文章中，本体澄清的对象是使用RDFS和OWL语言描述的本体，只经过稍微的改动。该方法能处理其他表示语言描述的本体。

WordNet包括名词、动词、形容词和副词的定义，以及它们之间的关系信息，这些关系将这些词组成一个网络层次结构。这篇文章使用的WordNet的版本为2.1，共包括155327个词和117597个词义。

3.1 算法

假设本体O和使用该本体标注的文档D。本体澄清算法为本体元素确定正确的在WordNet中定义的词义。存在着两种类型的词：单个词和组合词。单个词仅包含一个能在WordNet中找到的词，它的词义$S(t)$定义如下：

$$S(t)=s_k, where s_k \in Synset(t)$$

组合词本身不能在WordNet找到，它包含几个在WordNet中定义的词，$t=w_1w_2\cdots w_n$，它的词义是组成它的词的词义并集：

$$S(t)=\bigcup_1^n s_k, where s_k \in Synset(w_k), w_k \in t$$

通过消歧可以得到词的词义解释[4]。在消歧的处理过程中，选择目标词附近一定范围的上下文。在本体中，上下文通过目标词路径长度L来决定。概念和它直接相连的关系的路径长度为0.5。L_c和L_p分别表示概念和关系的路径长度。被标注的文档的上下文通过窗口W来确定，包括目标词在最中间的W个词。

本体澄清算法只对本体中的概念和关系进行消歧，实例(individual)并不进行处理。因为实例一般都只有一个意义，没有在WordNet中出现的词也被忽视。已在RDFS和OWL中定义的关系，比如说“rdfs:subClassOf”和“owl:sameAs”，将不会进行消歧处理。已定义的关系将用来判断不同的上下文。本体澄清包括三步。

步骤1 预处理

为了更好地进行消歧，会先对本体中概念和关系的词进行一些预处理。首先词被划分成单词，这可以根据单词之间特殊的连接符号或者是开头字母的大写来划分。然后在这些单词中，去除掉一些连接词、冠词、形容词、副词等的停用词。

步骤2 概念词的语义消歧

在语义消歧方法中，和上下文语义相关度最大的词义就是目标词的正确意思。概念词的语义消歧分为三步。首先，对于本体中概念词的上下文，将其分为4个集合：*superClass*集合、*subClass*集合、*equivalentClass*集合和*other*集合，分别使用CC_1，CC_2，CC_3和CC_4来表示。*superClass*和*subClass*集合包目标词的父类和子类概念，它们有“rdfs:subClassOf”的关系；*equivalentClass*集合包含相同类，它们之间有“owl:equivalentClass”或“owl:sameAs”的关系；*other*集合中包含与目标词有其他关系的概念以及被标注文档中给定窗口内的词。接着，获取目标候选词义附近的词义集合，词义集合分为三类：*Hyponymy*集合，*Hypernymy*集合和*Synonyms*集合，它们分别包括与候选词义上、下位词和同义词。我们使用S_1，S_2和S_3来表示这三类集合。最后，上下文集合和候选词义之间的相关度可以使用下面的式子来计算：

$$Rel(C,s_i)=R(CC_1,S_1)+R(CC_2,S_2)+R(CC_3,S_3)+R(CC_4,\{s_i\})$$

语义相关度函数R将在3.2节介绍。通过上面的处理，选取相关度最大的词义作为目标概念词的语义解释。接下来，就可以使用相同的方法处理其他的概念词了，在后续处理时，将使用已经确定的词义。

步骤3 关系词的语义消歧

对关系词的处理跟步骤2很相似。在这里，上下文集合分为五类：*superProperty*集合、*subProperty*集合、*equivalentProperty*集合、*inverseProperty*集合和*other*集合，分别用PC_1，PC_2，PC_3，PC_4和PC_5来表示。*superProperty*和*subProperty*集合包括与目标词有“rdfs:subPropertyOf”关系的词；equivalentProperty集合包括与该词有“owl:equivalentProperty”和“owl:sameAs”关系的词；*inerseProperty*集合包括与该词有“owl:inverseOf”关系的词；*other*集合包括与该词有其他关系的词和被标注文档中目标词附近的词。WordNet中目标候选词义附近的词义集合分为四类：*Hyponymy*集合、*Hypernymy*集合、*Synonyms*集合和*Antonyms*集合。前三类和步骤2中的解释一样，*Antonyms*集合包含候选词义的反义词，使用S_4来表示。然后，我们使用下面的公式来计算相关度：

$$Rel(PC,S_j)=R(PC_1,S_1)+R(PC_2,S_2)+R(PC_3,S_3)+R(PC_4,S_4)+R(PC_5,\{s_j\})$$

相关度最大的候选词义作为目标词的语义解释。先对概念进行消歧，后对关系进行消歧的原因在于动词的消歧比名词的消歧要困难[6]。在WordNet中，名词平均有5个词义，而动词平均有16个词义，这使得对动词消歧选择的范围要大得多，从而导致精度降低。为了提高关系消歧的精度，我们使用已经确定语义解释的概念词。

在上面的算法中，使用Link Grammar Parse[8]来分解处理组合词。我们将分别计算这些分解词的相关度，然后再加权比较。由于组合词分解的单词也许包含不同的词性，如果概念组合词的分解词不是名词或关系组合词的分解词不是动词，则采用下面的式子来计算相关度：

$$Rel(C,s_i)=R(C,\{s_i\})$$

3.2 语义相关度的计算

词或词义之间语义相关度方法可以分为三类：基于路径

的方法、基于信息量的方法和基于释义的方法[6]。本文选择基于释义的方法来作为语义相关度计算的基本方法，其原因主要是有两点：首先，基于释义的方法能够计算不同词性的词之间的相关度，而其他两种方法只能计算相同词性的词之间的相关度；其次，通过词义消歧实验[7]，基于释义的方法是最优的方法。基于释义的方法通过计算两个词释义相互覆盖的程度来衡量相关度。本文使用的基于释义的方法是 extended gloss based measure[6]。

集合间的语义相关度由集合中元素的相关度来决定。假设 $A=\{a_1,a_2,\cdots,a_{|A|}\}$，$B=\{b_1,b_2,\cdots,b_{|B|}\}$，那么集合 A 和 B 之间的语义相关度采用下面的公式来计算：

$$R(A,B)=\frac{\sum relateness(a_i,b_j)}{|A||B|},a_i\in A,b_j\in B$$

函数 *relateness* 使用基于释义的方法来计算词或词义之间的语义相关度，采用下面的公式对该相关度进行归一化处理：

$$relatedness(s_i,s_j)=\frac{number\_of\_overlaps}{(wordNumInGlossOfs_i+wordNumInGlossOfs_j)/2}$$

4 实验与评估

我们通过采用式(1)比较算法自动消歧的结果和手动处理的标准结果来计算本体澄清的精度。概念精度和关系精度分别采用式(2)和(3)来计算。

$$Precision=\frac{correct\ disambiguated\ terms}{all\ disambiguated\ terms} \quad (1)$$

$$Concept\ Precision=\frac{correct\ disambiguated\ concept\ terms}{all\ disambiguated\ concept\ terms} \quad (2)$$

$$Property\ Precision=\frac{correct\ disambiguated\ property\ terms}{all\ disambiguated\ property\ terms} \quad (3)$$

4.1 实验数据

表 1 实验中使用的本体

Num	Ontology	Library	概念定义
1	ATO_Mission_Model. owl	DAML	Simple
2	ATO_Ontology. owl	DAML	Simple
3	Comunication. owl	DAML	Simple
4	Government. owl	DAML	Simple
5	Contact-ont. owl	DAML	Complex
6	SUMO. owl	DAML	Simple
7	Camera. owl	Protégé	Complex
8	Countries. owl	Protégé	Complex
9	Delegation. owl	Protégé	Complex
10	ka. owl	Protégé	Complex
11	koala. owl	Protégé	Complex
12	people+pets. owl	Protégé	Complex
13	PNOntology. owl	Protégé	Complex
14	travel. owl	Protégé	Complex
15	tambis-full. owl	Protégé	Complex

实验数据使用从 DAML[1)] 和 Protégé[2)] 本体库中收集的本体，这些本体如表 1 所示。在分析这些本体的过程中，我们发现本体中的概念基本上都采用“rdfs:subClassOf”定义，这些定义可分为两类：一种“rdfs:subClassOf”的对象为一个简单的概念，我们将这种本体叫做简单本体；另一种其关系的对象是非常复杂的概念，我们将其叫做复杂本体。复杂本体和简单本体的不同之处在于，复杂本体概念的上下文集合的元素个数平均是简单本体的 3.5 倍，关系的上下文集合的元素个数平均是简单本体的 2 倍。我们使用 google 搜索 300 篇文本文档，使用上表中的本体对每 20 篇进行手动标注。

4.2 算法评估

第一个实验：调查路径长度对于本体澄清的影响。从图 1 和 2 中，可以看出 $L_c=2$ 时，平均概念精度有最大的值；当 $L_P=2.5$ 和 $L_c=2$，平均关系精度有最大的值。其原因在于路径长度越大，上下文集合中将包括更多的元素，从而使得消歧越加准确；另一方面，路径长度如果太大的话，将使得上下文集合中包含很多无关的噪声元素，从而导致消歧错误。

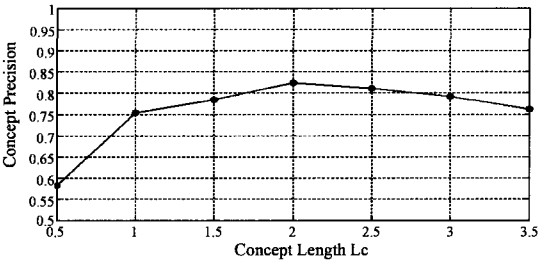

图 1 随路径长度 L_c 变化的概念精度图

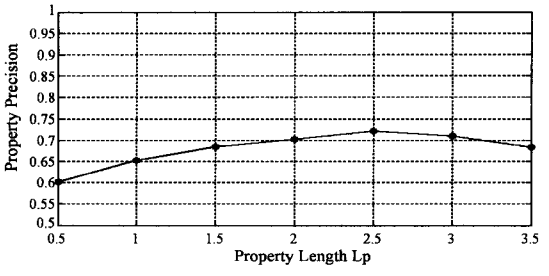

图 2 随路径长度 L_p 变化的关系精度图当 $L_c=2$

第二个实验比较只考虑本体结构的消歧算法和另外考虑标注文档的消歧算法的精度。设定 $L_P=2.5$，$L_c=2$ 和 $W=11$。W 取 11 的原因在于从文献[7]中的实验可知该值是最优的取值。通过采用本体澄清算法，实验结果如图 3 和表 2 所示。在表 2 中，符号“+”用来表示另外考虑标注文档的消歧算法。从结果可以发现 5 个现象。第一，另外考虑标注文档的消歧算法，精度大于只考虑本体结构信息的消歧算法，这是因为考虑文档将会增加上下文集合中元素的个数；第二，通过与 extended gloss overlap 算法进行的词义消歧的实验[6]相比，可发现概念精度差不多是名词精度的 2 倍，关系精度差不多是动词精度的 3 倍，这是因为词义消歧的上下文中包含太多的噪声；第三，概念精度比关系精度要高，这是因为一方面概念的上下文集合的元素个数比关系的上下文集合的元素个数要多，另一方面关系的候选词义比概念的候选词义要多；第四，简单本体的精度比复杂本体的精度要低，原因在于复杂本体中的上下文集合包含更多的元素；第五，考虑文档的概念和关系的精度差小于不考虑文档的它们之间的精度差，这是因

(下转第 185 页)

1) http://www.daml.org/ontologies/

2) http://protege.cim3.net/cgi-bin/wiki.pl?ProtegeOntologiesLibrary

ence, vol. 3100. Berlin: Springer-Verlag, 2004: 78-95

[5] Kryszkiewicz M. Rough set approach to incomplete information systems[J]. Information Sciences, 1998, 112: 39-49

[6] Stefanowski J, Tsoukias A. Incomplete information tables and rough classification[J]. Computational Intelligence, 2001, 17: 545-566

[7] Leung Y, Wu W Z, Zhang W X. Knowledge acquisition in incomplete information systems: A rough set approach[J]. European Journal of Operational Research, 2006, 168: 164-180

[8] Wu W Z, Xu Y H. On two types of generalized rough set approximations in incomplete information systems // Hu Xiaohua, Liu Qing, Skowron A, et al., eds. 2005 IEEE International Conference on Granular Computing. Beijing, China, July 2005: 303-306

[9] 杨晓平. 不完备信息系统一种新的粗糙集的性质[J]. 计算机科学, 2004, 31(10A): 64-65, 94

(上接第 147 页)

为标注文档能增加上下文元素。

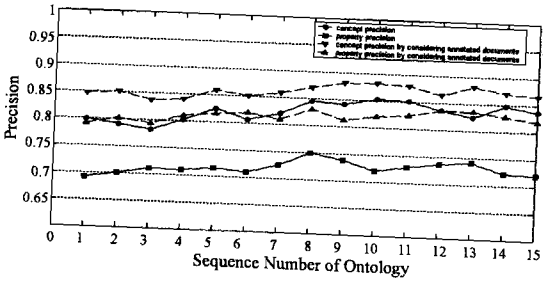

图 3 精度比较图

表 2 平均精度比较

	平均概念精度	平均关系精度	平均关系精度+	平均概念精度+
简单本体	76.9%	70.5%	84.5%	80.3%
复杂本体	83.8%	73.0%	87.1%	82.3%
Overall	82.4%	72.2%	86.3%	81.6%

最后一个实验评估本体澄清算法在半自动过程中的有效性。计算精度的公式为

$$Precision_n = \frac{correct\ disambiguated\ termsintop\ nsenses}{all\ disambiguated\ terms}$$

$$Concept\ Precision_n = \frac{correct\ disambiguated\ concept\ terms\ in\ top\ n\ senses}{all\ disambiguated\ concept\ terms}$$

$$Property\ Precision_n = \frac{correctd\ isambiguated\ property\ terms\ in\ top\ n\ senses}{all\ disambiguated\ propert\ yterms}$$

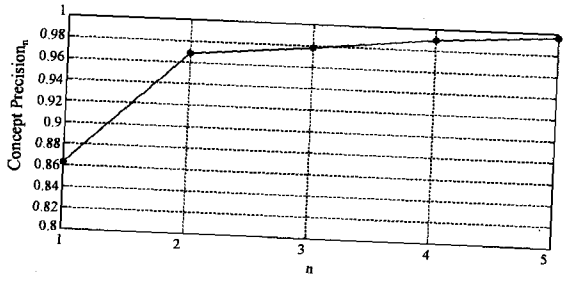

图 4 随 n 变化的 Concept Precision$_n$

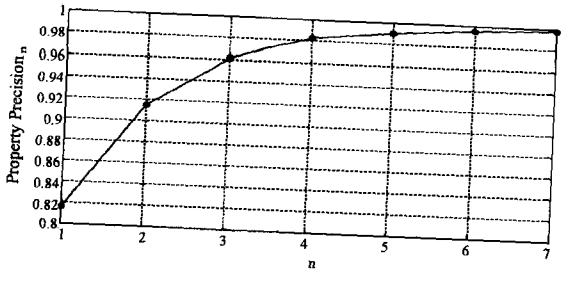

图 5 随 n 变化的 Property Precision$_n$

在上面的计算公式中，词义按照相关度来排序。若目标词的语义解释在最高的 n 个词义内，则认为是正确的。在半自动的本体澄清的过程中，用户能够从最高的 n 个候选词义中选择正确的语义解释。从实验结果图 4 和 5 可看出，concept precision$_3$ 和 property precision$_4$ 能够达到 98%的精度，从而证明了本体澄清算法在半自动过程中的有效性。

结束语 为了提高本体的质量，用 WordNet 中的词义表示本体的元素。本文陈述了考虑本体结构和被标注文档自动对本体元素进行语义消歧的本体澄清过程，实验证明了该方法的有效性。未来的工作分为两方面：首先计划开发本体澄清的工具，其次将调查澄清后的本体给基于本体的应用所能带来的好处。

参 考 文 献

[1] Ushold M, Gruninger M. Ontologies: Principles, methods and applications. The Knowledge Engineering Review, 1996, 11(2): 93-136

[2] Guarino N. Formal ontology and information systems // Proc. of the 1st Int'l Conf. on Formal Ontologies in Information Systems (FOIS98). Trento, Italy, IOS Press, 1998: 3-15

[3] Fellbaum C. Wordnet: An Electronic Lexical Database. Cambridge: MIT Press, 1998

[4] Missikoff M, Navigli R, Velardi P. Integrated approach to Web ontology learning and engineering. IEEE Computer, 2002, 35 (11): 60-63

[5] NavigliR, Velardi P, Gangemi A. Ontology learning and its application to automated terminology translation. IEEE Intelligent Systems, 2003, 18 (1): 22-31

[6] Banerjee S, Pedersen T. Extended gloss overlaps as a measure of semantic relatedness // Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence. Acapulco, 2003: 805-810

[7] Pedersen T, Banerjee S, Patwardhan S. Maximizing Semantic Relatedness top erform Word Sense Disambiguation. research report umsi 2005/25. Supercomputing Institute, University of Minnesota, 2005

[8] Sleator D, Temperley D. Parsing English with a Link Grammar. technical report. CMU-CS-91-196. Carnegie Mellon University, 1991