利用句法模式从术语词典中抽取语义关系

张希府[1], 戴云徽[2], 高志强[1]

(1. 东南大学 计算机科学与工程学院, 江苏 南京 210096; 2. 南京理工大学 经济管理学院, 江苏 南京 210094)

[摘要] 提出了一种基于句法模式的语义关系抽取方法, 用于从术语词典中抽取语义关系. 该方法以句法模式为中心, 结合了自然语言处理技术和统计的思想, 充分利用术语词典文档中的句法信息, 通过抽取包含着语义关系信息的句法模式, 并将其与词典文本进行近似匹配以达到抽取语义关系的目的. 实验结果表明, 该方法可以有效地从术语词典中抽取多种语义关系.

[关键词] 句法模式, 语义关系, 术语词典

[中图分类号] TP 391 [文献标识码] A [文章编号] 1672-1292(2008)04-0043-03

Applying Syntactic Patterns to Semantic Relation Extraction From a Terminology Dictionary

Zhang Xifu[1], Dai Yunhui[2], Gao Zhiqiang[1]

(1. School of Computer Science and Engineering, Southeast University, Nanjing 210096, China;
2. College of Economy, Nanjing University of Science and Technology, Nanjing 210094, China)

Abstract In this paper, we present a syntactic pattern based method for extracting semantic relations from a terminology dictionary. Focusing on syntactic patterns, combined with NLP techniques and the idea of statistics, this method, making full use of syntactic information in the dictionary text, extracts syntactic patterns including semantic relations and applies them to approximate matching with sentences from a dictionary so as to extract semantic relations. Results from the experiment show that this method can be used effectively in obtaining various semantic relation samples from a terminology dictionary.

Key words syntactic pattern, semantic relation, terminology dictionary

手工构建领域本体耗时费力, 这促使人们实现半自动、自动化构建本体, 本体学习就是这样一种过程. 按照 Staab的定义[1], 本体学习包括学习术语、同义词、关系等. 从关系学习的方法来看, Hearst[2]提出的利用句法模式的方法, 逐渐被广泛采用. 这类方法的特点是利用了文本中词之间的顺序信息, 适用于多种语义关系的学习. 从术语词典中抽取语义关系, 前人做了相关研究. Maria[3]等人提出利用模式方法, 从百科全书中自动抽取语义关系, 他们借助 WordNet识别两个术语的语义关系, 抽取句法模式, 利用编辑距离泛化相似模式, 最后用于关系的抽取; Finkelstein-Landau等人[4], 基于自扩展思想, 人工给定一些具有某种关系的术语对, 通过程序发现包含术语对的句子, 从中抽取句法模式, 由专家确认后用于发现更多的术语对, 继而抽取更多的句法模式. 除句法模式外, 学习语义关系的方法还有: 基于统计的方法, 如 Heyer等人[5]为了从语料库中学习关系, 定义搭配的概念; 基于图模型的方法, 如 Jannink[6]提出了一个模型 ArcRank用于抽取词典中词之间的层次关系, Blondel[7]也是利用基于词典的图模型来抽取同义关系.

本文工作包括: 词典文档的标注、模式的提取、文本的近似匹配、关系实例的定位和抽取、结果的评估. 实验结果表明, 本文的方法可以有效抽取多种语义关系实例.

1 方法

相关模块的处理包括预处理、模式抽取、模式匹配和关系抽取.

收稿日期: 2008-06-18
基金项目: 国家科技支撑计划(2006BAK10B02)资助项目.
通讯联系人: 高志强, 副教授, 博士, 研究方向: 人工智能、虚拟现实和计算材料学. E-mail: zqgao@seu.edu.cn

1.1　预处理

预处理包括原始词典文档的格式化和各类标注的添加. 首先将 PDF格式的词典文档转换成 XML格式,通过 XML解析器就可以方便地将术语和定义分开. 预处理包括分词、分句和各类标注的添加. 词性标注、语态标注和 NP标注包含文档原有信息, 术语和关系信息是基于上面标注添加的. 本文利用 GATE实现这些信息的添加. 在本文中, 关系抽取是一个有监督过程, 需要人工添加关系标注, 它也是模式抽取的基础.

1.2　模式抽取

通过预处理, 词典文档包含许多标注信息, 这使得模式抽取成为可能. 不同于其它工作的做法[3,4], 本文只在模式抽取时进行了一定的泛化处理.

模式的抽取包括 5个步骤: (1)按被 NP覆盖和未被 NP覆盖, 将句子变换为一个由 NP节点和 token节点组成的序列. (2)同一术语跨越的多个节点合并为一个节点. (3)相邻的 NP节点合并为一个节点. (4)考察一个句子所包含的各 NP节点, 统计包含术语的 NP节点(以下称术语 NP节点)的数目, 忽略不为 2的句子. (5)按照预定义的宽度, 裁剪掉术语 NP节点两侧宽度范围之外的节点. 对一个关系实例, 通过处理, 可得到一条简化的句法模式.

1.3　模式匹配

本文的匹配策略基于最长公共子序列(LCS)的思想. 动态规划法是求取 LCS的典型算法. 下面给出一个例子. S表示"NP to NP(NP needed). ", P表示" be connected to NP (NP). ". 用 $S[i]$和 $P[j]$表示模式 S的第 i个节点和模式 P的第 j个节点, 且当 $i=0$或 $j=0$时, $\mathrm{LCS}[i,j]=0$,否则按下面公式计算 LCS矩阵各元素的值:

$$\mathrm{LCS}[i,j]=\begin{cases}\mathrm{LCS}[i-1,j-1]+1, & S[i]=P[j]\\ \max\{\mathrm{LCS}[i,j-1],\mathrm{LCS}[i-1,j]\}, & S[i]\neq P[j]\end{cases}$$

根据计算得 S和 P的最长公共子序列为" to NP (NP needed). ", 长度为 6得到 LCS的长度及其在原模式中的索引, 利用公式 $\mathrm{Match}=\dfrac{\alpha N_{innp}+\beta N_{inp}+\chi N_{onnp}+\delta N_{onp}}{L}\times\dfrac{2M}{S_{span}+P_{span}}$计算出两条模式的匹配度.

本文将公共子序列中的节点分为 4类: 位于两个术语 NP节点之间的节点称为内部节点, 其余的称为外部节点. α、β、χ、δ 4个权值, 反映 4类节点的重要性. 针对公共子序列, N_{innp}表示内部非 NP节点的个数, N_{inp}表示内部 NP节点的个数, N_{onnp}表示外部非 NP节点的个数, N_{onp}表示外部 NP节点的个数. L表示模式 S的长度. M表示公共子序列的长度. S_{span}表示 S模式中公共子序列所跨越的节点个数. P_{span}对应 P模式. S、P代表不同的角色, S代表句子, 即转为模式格式的句子, 而 P代表抽取到的句法模式.

1.4　关系抽取

本部分利用匹配的结果定位术语, 进而抽取关系实例. 首先将句子转换为模式格式, 然后与某种关系所对应的模式逐一匹配. 如果匹配度超过设定的阈值, 则近似地认为该句子与该模式匹配, 最后根据术语 NP在句子中的位置提取关系实例.

2　实验

实验采用 k- fold策略, 使用一个电信领域的小规模术语词典, 包含 910条术语. 本文随机地选择 400条术语用于训练和测试. 为了先验地评估语义关系抽取结果, 本文采用了精度、召回率和 F1-Measure的量化标准. 实验涉及几个可调的参数: 划分组数 k、模式裁剪的宽度 w、候选模式的数目 t、匹配度的精度 d和匹配度的阈值 θ

2.1　实验结果

图 1显示了这 3种关系所对应的 F1-Measure随匹配度阈值的变化情况.

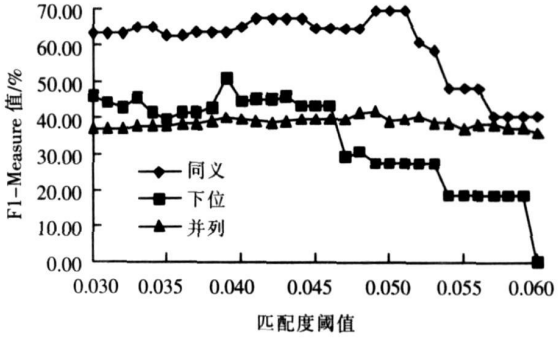

图 1　F1-Measure 值随匹配度阈值变化曲线图

Fig.1　F1-Measure curve along with match-degree threshold

表 1是 3类关系的 F1-Measure取最大值时的实验数据. 训练和测试两列分别存储各关系在各组实验中用作训练和测试的句子数目. 模式列对应一组实验所抽取的句法模式数目. 学习和存在两列显示的分别是程序和人工抽取的关系实例数目. 匹配一列存放的是学习和存在两列重叠部分的数目. 最后 3列均为百分比.

表 1 3类关系的抽取结果

Table 1 Extraction results for three relations

关系	组	训练	测试	模式	学习	匹配	存在	精度	召回率	F1
同义	1	16	5	17	5	4	6	80.00	66.67	73.34
	2	17	4	18	3	2	5	66.67	40.00	53.34
	3	17	4	18	4	2	5	50.00	40.00	45.00
	4	17	4	19	3	3	4	100.00	75.00	87.50
	5	17	4	20	3	3	4	100.00	75.00	87.50
下位	1	21	6	22	4	1	6	25.00	16.67	20.84
	2	21	6	22	6	2	6	33.33	33.33	33.33
	3	22	5	23	5	3	6	60.00	50.00	55.00
	4	22	5	22	3	3	6	100.00	50.00	75.00
	5	22	5	23	2	2	5	100.00	40.00	70.00
并列	1	72	18	65	28	7	23	25.00	30.43	27.72
	2	72	18	64	17	6	24	35.29	25.00	30.14
	3	72	18	70	16	8	22	50.00	36.36	43.18
	4	72	18	70	22	11	20	50.00	55.00	52.50
	5	72	18	71	9	7	21	77.78	33.33	55.56

2.2 结果分析

由实验可见, 本文的基于句法模式的语义关系抽取方法可以有效地从词典文档中抽取同义、下位和并列 3种语义关系. 对同义关系的抽取效果最好, 原因在于同义关系在词典文档中的句法模式相对稳定, 数目有限. 对于并列关系的抽取结果不是很好, 究其原因, 并列关系比较多, 模式比较平凡, 与并不包含此关系的句子也能较好地匹配, 导致抽取了一些错误的关系实例.

3 结语

实验表明, 本文的方法可以有效地从术语词典中抽取同义、下位和并列等语义关系实例. 为了改善抽取性能, 本文设置了一些可以调节的参数, 如样例集合的划分组数 k、模式裁剪宽度 w、候选模式数目 t、匹配度精度 d 和匹配度阈值 θ. 从实验结果可能看出, 通过合理调节这些参数, 可以有效地提高关系抽取的性能.

[参考文献] (References)

[1] Staab S, Hotho A. Machine learning and the semantic web[C] // Tutorial at the 22nd International Conference on Machine Learning. Germany: Bonn, 2005.

[2] Hearst M A. Automatic acquisition of hyponyms from large text corpora[C] // Proceedings of the 14th Conference on Computational Linguistics. USA: Association for Computational Linguistics Morristown, 1992: 232-28.

[3] Ruiz-Casado M, Alfonseca E, Castells P. Automatic extraction of semantic relationships for wordnet by means of pattern learning from wikipedia[C] // NLDB. Heidelberg: Springer, 2006: 672-79.

[4] Finkelstein-Landau M, Morin E. Extracting semantic relationships between terms: supervised vs. unsupervised methods[C] // Proceedings of International Workshop on Ontological Engineering on the Global Information Infrastructure. Germany: Dagstuhl Castle, 1999: 71-80.

[5] Heyer G, Läuter M, Quasthoff U, et al. Learning relations using collocations[C] // Proceedings of the IJCAI Workshop on Ontology Learning. Seattle, USA: IJCAI, 2001: 2-4.

[6] Jannink J, Wiederhold G. Thesaurus entry extraction from an on-line dictionary[C] // Proceedings of Fusion 99. Sunnyvale, CA: Omni Press, 1999.

[7] Blondel V D, Senellart P. Automatic extraction of synonyms in a dictionary[C] // Proceedings of the SIAM Workshop on Text Mining. Arlington: Virginia, 2002.

[责任编辑: 严海琳]