

基于特征组合的中文实体关系抽取

黄 鑫^{1,2}, 朱巧明², 钱龙华², 刘梅梅^{1,2}

(1 江苏省计算机信息处理技术重点实验室, 江苏 苏州 215006; 2 苏州大学 计算机科学与技术学院, 江苏 苏州 215006)

摘 要: 结合中文关系抽取的要求, 以 ACE2005 的中文语料为数据进行关系抽取实验. 在抽取中文词法、实体、句法、语法基本特征后, 提出采用特征组合方法, 使用支持向量机的机器学习 (SVM) 方法, 在上关系探测和关系大类上 F 值分别提高了 1.36% 和 3.97%, 达到 72.77 和 61.03, 并分析出各部分组合特征的贡献. 实验数据表明词语和实体组合特征对中文关系抽取的作用较大.

关键词: 中文关系抽取; 信息抽取; ACE2005; 支持向量机; 特征组合

中图分类号: TP393

文献标识码: A

文章编号: 1000-7180(2010)04-0198-03

Chinese Entity Relation Extraction Based on Features Combination

HUANG Xin^{1,2}, ZHU Qiao-ming², QIAN Long-hua², LIU Mei-mei^{1,2}

(1 Key Lab of Computer Information Processing Technology of Jiangsu Province, Suzhou 215006, China;

2 School of Computer Science & Technology, Soochow University, Suzhou 215006, China)

Abstract: This paper carried out a series of experiments on Chinese relation extraction classification based on standard and training corpus of ACE2005 (Automatic Content Extraction 2005). It explores word, entity, syntax, gram features in Chinese at first, and then present a method which combines these basic features. The F-score of Chinese relation extraction for Relation Detection and six major types in ACE2005 Chinese corpora improves 1.36% and 3.97% and achieves 72.77 and 61.03 respectively in SVM, then give the contribution of different combined features. It illustrates that the combined features of words and entities are very effective for Chinese Relation Extraction.

Key words: Chinese relation extraction; information extraction; ACE2005; SVM; feature combination

1 引言

实体间关系抽取是从文本中抽取特定的事实信息, 称之为实体, 然后判断出这些实体的关系. 美国国家标准技术研究院 (NIST) 组织的自动内容抽取 (Automatic Content Extraction, ACE) 评测针对中文语言的特征, 在 ACE2005^[1] 的中文部分给出了 7 种实体类别, 并在此基础上规定了各个实体类别之间存在的 6 种实体关系类别^[2,4]. 文献[5-6]分别通过抽取词语特征和抽取实体覆盖特征两个方法作了研究. 文中以 ACE2005 中文部分的语料为数据, 对

实体间的语义关系进行研究. 提出在抽取词语、实体、句法、语法特征的基础上, 通过词语特征、实体特征、句法特征内部的组合的方法, 有效提高关系抽取性能.

2 基于特征向量的关系抽取

2.1 特征向量

基于特征向量的中文实体关系^[7]抽取研究重点在于实例特征向量构造. 特征应该具有如下特点: 首先, 应能够最大限度地包含文档的重要信息; 其次, 特征抽取应该是可行且相对简单, 最简单的特征

收稿日期: 2009-03-12; 修回日期: 2009-06-12

基金项目: 国家“八六三”计划项目 (2006AA01Z147); 江苏省自然科学基金 (60673041)

选择方法是基于特征项出现频率的;最后恰当的特征是与分类相关的特征,具有较强的区分度。

中文的实体间语义关系抽取的成败在于能否有提取各种有效的语言特征和语法结构,所以在特征方面主要选择词法、实体、句法、语义以及结构化信息。词法特征的提取相对简单,实体特征依赖于原始语料标注信息的处理,而句法、语义以及结构化特征的提取需要对原始语料进行句法分析、短语块标注等预处理工作,受制于这些预处理工作的性能,目前无法提取出更多准确而有效的新特征。

2.2 词语特征

词语特征是实体最基本也是最简单的特征,它主要包含实体中心词,扩展词。除了实体本身,考虑到上下文的联系,把实体前后两个词也加入到词语特征之中,如图1所示。

BWM1	BWM1	MW1	BWF	BWO	BWL	MW2	BWM2	BWM2
F	L						F	L

图1 词语特征的抽取

其中, MW1: 实体1词语; MW2: 实体2词语; BWM1F: 实体1前倒数第二个词语; BWM1L: 实体1前最后一个单词; BWF: 实体间第一个词语; BWO: 实体间第一个和第二个的其他的单词; BWL: 实体间前最后一个单词; BWM2F: 实体后的第一个单词; BWM2L: 实体2后的第二个单词。

2.3 实体特征

实体信息是由ACE标注的,它包含了许多丰富的信息,在对标注文件处理后形成实体信息文件,如图2是对一个实体的标注信息及其说明。

```
Entity7-9GPE|Population-Center|SPC|NAM|GPE|NAM
|null|台北22234 实体 实体编号 实体大类(ET)|实体子
类(EST)|实体参考类别(EC)|实体引用类型(|GPE 角色|
LDC 引用类型|LDC 属性|实体中心词 实体所在行号 实
体中心词开始号 实体中心词结束号 实体扩展词开始号 实
体扩展词结束号
```

图2 实体特征的抽取及其说明

2.4 句法语法特征

在句法和语法特征中主要是有各个句子的基本短语块组成的。在对话料做完句法语法分析的基础上,获取各个基本短语信息。如图3所示。

MW1(ET1)	BCHF	BCHB	BCHL	MW2(ET2)
----------	------	------	------	----------

图3 基本短语块特征的抽取

其中, MW1: 实体1词语; MW2: 实体2词语; BCHF: 实体间第一个基本短语; BCHO: 实体间第一个和第二个的其他的的基本短语; BCHL: 实体间前最

后一个基本短语。

2.5 组合特征

组合特征是在现有基本特征的基础上,通过内部组合,形成新的特征。组合特征可以充分利用上下文信息,增加各个关系之间的区分度。文中主要从以下3个方面来增加组合特征:

- 词语组合特征: MH1MH2: WM1 的核心词和 WM2 的核心词组合; BWM1: WM1 前两个词的组合; BMM2: WM2 后的两个词的组合。

- 实体组合: ETIET2: 实体1 实体2 大类组合; ESTIEST2: 实体1 实体2 子类组合; ECIEC2: 实体1 实体2 参考类别组合; MLIML2: 实体1 实体2 引用类型的组合; MSLIMSL2: 实体1 实体2 参照方式的组合。

- 句法组合: BCH1: 实体1 前两个基本短语的组合; BCH2: 实体2 后两个基本短语的组合。

3 实验结果及其分析

3.1 实验数据

实验使用 ACE2005 中的中文标注语料库作为数据。该语料库的数据来源于广播、新闻、报纸等。其中共使用了 578 篇文章,其中 458 篇作为训练语料, 120 篇作为测试语料。表1列出了训练语料和测试语料中实体间关系的统计数据。

表1 训练和测试数据中关系类别的统计

Physical	1115	420
Part - whole	1647	436
Personal - Social	470	130
ORG - Affiliation	1541	405
Agent - Artifact	444	106
General - Affiliation	1376	331

实验首先对话料进行分词,词性标注,句法分析预处理,然后结合实体标注文件,按照上述特征抽取方法产生特征文件。图4是“然而他们最终还是选择回房间用最实惠的方式来庆祝。”抽取出的特征。其中他们和房间是两个实体,它们之间是 Physical 关系。这句话按照特征抽取的方法得出的文件,在此基础上形成特征向量文件。利用 svm_light 对特征文件进行训练测试,最后产生结果。对得出的性能进行评测,使用准确率(Precision)、召回率(Recall)、F 值(F-Score)作为标准。

3.2 实验结果及分析

实验结果是以关系探测和关系识别性能值为指标来评价,同时也分析出各部分特征对关系抽取的

贡献.从表 2 可以看出,加入组合特征后,系统在关系探测和关系识别分别提高了 1.36 和 3.97.图 5 给出加各个特征在 6 大类关系的贡献统计图,结果表明 6 大类关系都随着各个组合特征的加入 F 值不断提高,其中在词语的基础上,实体及其组合特征的加入对 6 大类的关系抽取贡献是最大.

BWF_最终 BWL_回 BW0_还是 BW0_选择 BWM1F_然而 BWM1L_然而 BWM1_然而 BWM2F_用 BWM2L_最 BWM2_用最 MW1_他们 MW2_房间 MH1MH2_他们 # 84-40+房间 # 82-41 ET1ET2 _ PER. NONE + FAC. NONE EST1EST2 _ PER. NONE | PER. GROUP + FAC. NONE | FAC. SUBFAC EC1EC2 _ SPCSPC ML1ML2 _ PRONOM MSL1MSL2 _ PRO. PRONOM. NOM # MB _ 0 # WB _ 4 BCHF_最终 BCHB_还是 BCHB_选择 BCHL_回 BCHL2_选择回 CONTEXT_(PER. None, FAC. None)然而 _<1 他们_1>_最终_还是_选择_回_<2 房间_2>_用_最_实惠_的方式_来_庆祝_. YES. PHYS.

图 4 特征文件抽取实例

表 2 组合特征对性能的贡献

特征	关系探测			关系识别		
	P	R	F	P	R	F
未加组合特征	80.05	64.45	71.41	70.51	48.06	57.16
加入组合特征	80.44	66.44	72.77	74.1	51.88	61.03

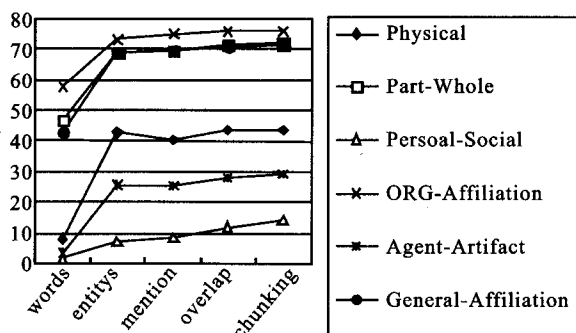


图 5 各关系在 6 大类关系上的贡献

表 3 则在整体上,从各个特征及其组合对整个信息抽取的贡献做了统计.词语及其组合和实体及其组合信息对最终结果的提高贡献较大,分别为 37.19 和 20.02,基本短语对系统的性能贡献不是很明显,只有 0.39,这与和图 5 中对 6 大类的统计基本一致.

组合特征有效结合基本特征,提高了各个实体关系的区分度,提高了关系抽取的性能.其中实体及其组合特征一方面由于来源 ACE 的实体手工标注信息,所以实体特征丰富、准确、完整.另一方面通过实体特征的组合,形成新的特征,可以更好的把两个

实体联系起来,增加各个关系之间的区分度,准确描述关系.

表 3 各特征在 ACE2005 中文关系抽取中贡献的统计

特征	Precision	Recall	F1
词语及其组合特征	73.54	24.89	37.19
+ 实体类别及其组合	68.45	49.14	57.21
+ 实体引用类别组合	71.65	49.94	58.86
+ 实体覆盖组合	72.08	52.34	60.64
+ 基本短语及其组合	74.10	51.88	61.03

4 结束语

文中介绍了以 ACE2005 的中文语料库为数据,以 SVM-Light 为分类器,使用基于特征向量的方法来完成的中文实体间语义关系抽取实验.文中重点在对中文语料基本特征抽取后,充分考虑中文语料的特点,通过对语料处理产生的实体标注信息,结合多种实体信息,提出对词语、实体、句法采用组合的方法.结果表明组合特征对性能提高起到了作用.

寻找有效特征是整个基于特征方法的核心,下一步工作就是寻找更好的特征来提高整个关系抽取的性能,可以通过引入中文的知网(HowNet)中更多的语义信息,来提高系统的区分度,从而实现更好的中文实体语义关系抽取.

参考文献:

- [1] ACE 2005. The automatic content extraction (ACE) projects[EB/OL]. (2007-10-11)[2009-03-10]. <http://www.ldc.upenn.edu/Projects/ACE/>.
- [2] Nanda Kambhatla. Combining lexical, syntactic and semantic features with maximum entropy models for extracting relations[C]//Proc. Of ACL'2004. Barcelona, Spain: [s. n.], 2004.
- [3] Wang Ting, Li Yaoyong, Kalina Bontcheva, et al. Automatic extraction of hierarchical relations from text[C]//Proc. of the Third European Semantic Web Conference. Jeju Island, South Korea: [s. n.], 2005.
- [4] 孙晓玲,林鸿飞. 人际网络关系抽取和结构挖掘[J]. 微电子学与计算机, 2008, 25(9): 233-236.
- [5] 车万翔,刘挺,李生. 实体关系自动抽取[J]. 中文信息学报, 2005, 19(2): 1-6.
- [6] 董静,孙乐,冯元勇,等. 中文实体关系抽取中的特征选择研究[J]. 中文信息学报, 2007, 21(4): 80-85.

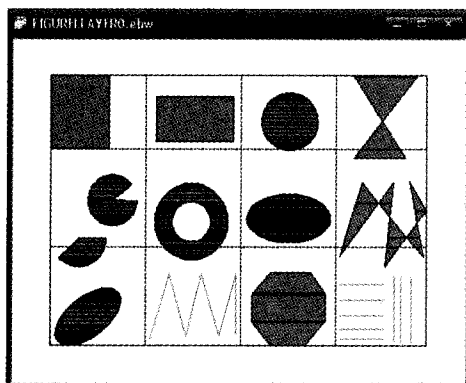
(下转第 204 页)

可适用于任意曲线图形,可实现不同尺寸图形离散的疏密程度不同,保证曝光效率.

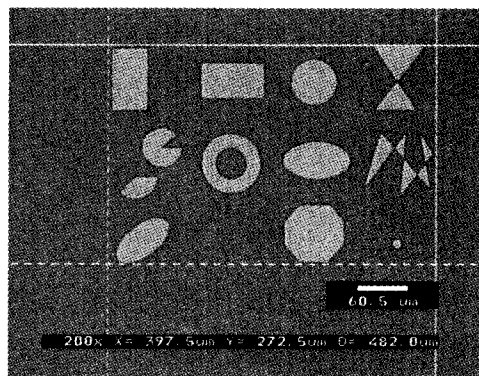
4 实验结果

文中研究的圆环弧类图形离散算法实现于自主研制开发的电子束曝光软件系统中.利用此软件系

统在电工所研制的基于 SEM 的电子束曝光系统进行曝光实验.曝光图形包括圆、圆环、椭圆、扇形等圆环弧类图形.图 5(a)为利用软件系统设计的实验曝光版图分割后结果,(b)为曝光结果.可看出所有图形曝光正常,特别是圆环弧类图形曝光后边缘光滑无图形缺失情况.



(a) 图形分割结果



(b) 曝光结果

图 5 图形分割及曝光实例

5 结束语

文中提出了一种精确、通用的圆环弧类图形离散算法,以图形的曲率半径作为离散精度依据,既保证了高的离散精度又减小了数据量,提高了数据传输效率,该算法为曝光版图中不断出现的圆环弧类图形进行精确曝光提供了保障.

参考文献:

- [1] 顾文琪. 电子束曝光微纳加工技术[M]. 北京:北京工业大学出版社,2004.
- [2] 吴石增,陈军. 电子束曝光机特殊图形辅助发生器[J]. 微电子学与计算机, 1995,12(4):14-17.

- [3] 崔铮. 微纳加工技术及其应用[M]. 北京:高等教育出版社,2005.
- [4] 孙家广,杨长贵. 计算机图形学[M]. 3版. 北京:清华大学出版社,1998.
- [5] Feito F, Torres J C, Urena A. Orientation, simplicity and inclusion test for planar polygons [J]. Computer & Graphics, 1995, 14(4):595-600.
- [6] 王志强,肖立谨,洪嘉振. 多边形的简单性、方向及内外点的判别算法[J]. 计算机学报,1998,21(2):183-187.

作者简介:

魏淑华 女,(1981-),博士.研究方向为微纳加工技术.

韩立男,(1970-),博士生导师.研究方向为微纳加工技术.

(上接第 200 页)

- [7] 于江德,肖新峰,樊孝忠. 基于隐马尔可夫模型的中文文本事件信息抽取[J]. 微电子学与计算机,2007,24(10):92-98.

作者简介:

黄鑫男,(1984-),硕士.研究方向为中文信息处理.

朱巧明 教授,博士生导师.研究方向为中文信息处理、网络语义处理、嵌入式等.

钱龙华 博士,副教授.研究方向为中文信息处理.

刘梅梅 硕士研究生.研究方向为语义网络处理.