ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.19, No.9, September 2008, pp.2265−2275
DOI: 10.3724/SP.J.1001.2008.02265

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

基于语义域语言模型的中文话题关联检测*

洪 宇+, 张 宇, 范基礼, 刘 挺, 李 生

(哈尔滨工业大学 计算机科学与技术学院 信息检索研究室,黑龙江 哈尔滨 150001)

Chinese Topic Link Detection Based on Semantic Domain Language Model

HONG Yu+, ZHANG Yu, FAN Ji-Li, LIU Ting, LI Sheng

(Information Retrieval Laboratory, School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

+ Corresponding author: E-mail: hy@ir.hit.edu.cn

Hong Y, Zhang Y, Fan JL, Liu T, Li S. Chinese topic link detection based on semantic domain language model. *Journal of Software*, 2008,19(9):2265−2275. http://www.jos.org.cn/1000-9825/19/2265.htm

Abstract: Topic link detection is a foundational research in the field of topic detection and tracking, which detects whether two random stories talk about the same topic. This paper proposes a method of applying semantic domain language model to link detection, based on the structure relation among contents and the semantic distribution in a story, and also verifies the influence of the strategy incorporating dependency parsing into semantic description. Evaluation on Chinese Corpus of TDT4 show that the semantic domain language model substantially improved the performance of current detection system, whose minimum DET cost is reduced by about 3 percent.

Key words: link detection; topic detection and tracking; semantic domain; language model; dependent parsing

摘 要: 关联检测是话题检测与跟踪领域的基础性研究,其任务是检测任意新闻报道对是否论述同一话题.通过分析报道内容的结构关系和语义的分布规律,提出基于语义域语言模型的关联性检测方法,并在此基础上检验融入依存分析的语义描述策略对该模型性能的影响.实验采用 TDT4 中文语料进行评测,结果显示语义域语言模型显著改进了现有检测系统的性能,其最小 DET 代价降低了约 3 个百分点.

关键词: 关联检测;话题检测与跟踪;语义域;语言模型;依存分析

中图法分类号: TP391 **文献标识码**: A

话题检测与跟踪(topic detection and tracking,简称 TDT)是一项针对新闻报道进行信息识别、挖掘和组织的研究.新闻报道是各传媒机构针对真实新闻事件的客观评论;新闻话题[1]则由种子事件及其直接相关的事件共同组成,其往往表现为某一新闻事件引发的所有相关报道;事件[2]则定义为发生于特定时间和特定地点的事情.例如,针对“2001 年 9 月 11 日美国世贸和五角大楼遭到恐怖袭击”的报道描述了话题“911”的种子事件,它与“灾后处理”、“嫌疑犯调查”和“国际社会援助”等后续相关事件构成了完整的“911”话题.

TDT 主要涉及两项任务:话题检测和话题跟踪.话题检测用于识别未知新闻事件,并将其作为种子事件挖掘新闻资源中的相关报道.话题跟踪则基于已知新闻话题,识别和收集实时新闻流中的后续相关报道.检测与跟踪

* Supported by the National Natural Science Foundation of China under Grant Nos.60435020, 60503072, 60736044 (国家自然科学基金); the National High-Tech Research and Development of China under Grant No.2006AA01Z145 (国家高技术研究发展计划(863))
Received 2007-07-14; Accepted 2007-11-20

在实际应用中往往相互依存:检测可以在新闻事件发生的初期建立初始话题模型,从而作为跟踪的目标和指南;跟踪则通过不断采集后续新闻流中的相关报道,对初始话题模型进行补充与完善.TDT 研究体系以检测与跟踪为基本框架,还包含报道切分、关联检测和新事件检测等辅助性任务.

关联检测(link detection task,简称 LDT)是 TDT 中的一项重要子任务,其定义是:检测随机选择的两篇报道是否论述同一话题[3].LDT 并不针对特定应用,而是服务于 TDT 领域的一项辅助性研究.以话题跟踪为例,一篇报道是否与当前话题相关,取决于该报道是否与构成该话题的所有报道足够相关,其中,每对报道的匹配都是一次关联性检测过程;而某些研究往往将话题的先验知识融合为话题模型,然后进行后续报道的相关性检测,其本质也是一对文本间的相关性匹配问题.此外,LDT 的核心问题是篇章理解和语义分析,往往需要剖析文本内部的层次结构、语法规则和语义关联性等等[4].因此,LDT 也是检验信息抽取、数据挖掘和自然语言处理技术的理想平台.

本文第 1 节分析研究现状.第 2 节介绍语义域的定义、构造方法及语义描述.第 3 节介绍语义域语言模型的建模方法以及基于相对熵的匹配策略.第 4 节介绍语料、评测以及实验设计.第 5 节报告实验结果并进行分析.第 6 节总结全文.

1 研究现状

传统基于统计策略的 LDT 研究将报道描述成高频特征集合进行匹配,报道之间的相关性取决于共有特征的数量及其权重.其中,Allan[5]和 Schultz[6]采用向量空间模型(vector space model,简称 VSM)描述报道的特征空间,根据特征在文本中的概率分布估计权重,利用余弦夹角衡量报道之间的相关性.此外,Yamron[7]将参与检测的两篇报道分别看作一个话题和一篇报道,采用一元语言模型(unigram language model,简称 ULM)描述报道产生于话题的概率,并通过调换两篇报道的角色分别从两个方向估计其产生概率,最终采用相对熵综合评估报道间的相关性.Lavrenko[8]针对 ULM 中数据稀疏问题以及平滑后特征权重被泛化的现象,提出相关性模型(relevance model,简称为 RM)并应用于 LDT 领域.RM 是基于 ULM 的扩展技术,它将报道作为 query 进行检索,通过相关度较高的伪相关反馈建立该报道的话题模型,并基于 ULM 评估报道之间的相关性.统计模型的缺陷在于将报道描述为基于独立假设的“词包”,忽视词义以及词与词之间的联系,报道间的匹配结果往往只能反映相似性,而非相关性.而相似性与相关性并不存在等价关系,假设两篇报道“911 恐怖袭击”和“巴厘岛惨案”,虽然报道中都频繁出现“恐怖分子”、“爆炸”、“袭击”和“伤亡”等高频特征,但这两篇报道并不相关,讨论的话题也不一致,只存在特征空间的相似性.

针对统计策略的缺陷,LDT 领域的相关研究尝试将语言信息与机器学习方法进行融合.Ponte[9]选择报道中高频特征作为扩展对象,采集围绕特征频繁出现的上下文信息对其进行扩展,特征空间由原始和扩展的特征项融合而成.扩展技术不仅有助于解决数据稀疏问题,还可以辅助 LDC(linguistic data consortium)系统削弱特征的歧义性.但词扩展并不能有效挖掘和描述报道的主题及其语义,同时引入过多噪声,因此其性能往往提高不大,甚至略低于统计模型.此外,Nallapati[10]通过命名实体识别和词性标注将特征划分为不同类别,通过估计特征产生于不同语法类别的概率标记其权重.语法类别参与特征空间的描述,有助于 LDT 把握构成报道主体事件的特征,但该方法并不能涵盖所有语言现象,对报道的描述仍然是相互独立并离散分布的特征集合,无法有效反映主题的语义,因此取得的改进并不明显.

综上所述,LDT 的核心问题在于如何挖掘报道的主题并对其语义进行描述,报道间的相关性也必须依赖主题语义的一致性进行判断.基于这一思想,本文将报道划分为趋向于不同语义的结构,通过语义域语言模型(semantic domain language model,简称 SDLM)描述主题的语义空间在各结构中的概率分布,并采用相对熵评估报道之间主题语义的一致性.

2 语义域

2.1 语义域定义

语义域是一组语义趋近一致的语言结构的集合,对语义进行描述的特征集合称为语义空间.语义域通过语义片断维护语义的一致性,并通过语境增强语义的全面性.语义片断是描述某一语义的最小语言结构,语境则是同时包含语义片断及其上下文的语言结构.以图 1 中的报道“金大中获得 2000 年诺贝尔和平奖”为例,其中,句 *a* 包括一项“金大中获得和平奖”的语义片断,它与句 *c* 和句 *e* 的语义片断趋近一致,该语义对应的语境集合(语境 *a*,*c* 及 *e*)构成语义域 *a*-*c*-*e*,其语义空间如图 1 所示.

报道是以一系列凝聚于主题的语义片断为框架,并基于因果、推导、包含和发展等关系连接而成的有机整体.主题的语义片断分布于不同层次结构,并引导各部分围绕文章主线进行论述.例如图 1 中的报道,其内容主要包括“授奖”事件;“获奖”原因以及“奖项”介绍 3 个子话题.该报道以“授奖”事件为主题,以另两项子话题作为相关外延,通过分布于不同子话题但凝聚于主题的语义片断(如图 1 中语义片断 a_1,*c* 及 *e*)建立文章框架,并以此为媒介连接和组织各子话题围绕主题进行论述.因此,主题挖掘在于合理切分报道,将语义趋向于一致的片断及其相关上下文(如图 1 中语境 *a*,*c* 及 *e*)融合为新的结构-语义域,并根据语境在全文中的分布规律,估计每个语义域生成主题的概率.

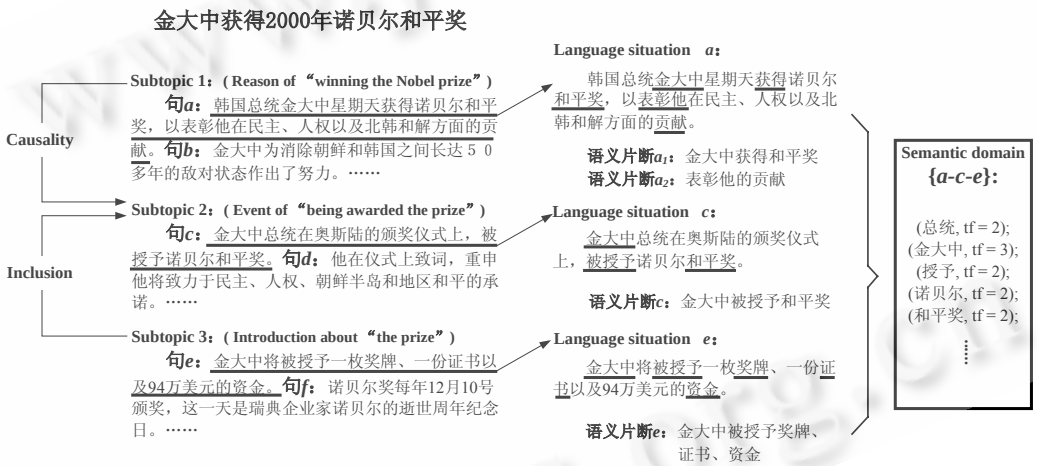

Fig.1 An example of structural analysis for a story

图 1 报道结构化分析样例

语义域并不等价于子话题,前者更注重结构内部语义的统一性,各组成部分之间不必存在必然的联系;后者则依赖话题的定义,以某一核心事件为主轴建立统一而独立的内容,可以通过凝聚于主题的语义片断与全文建立联系,如图 1 中子话题 1 的主要内容是“金大中生平在人权、民主与和解方面做出的贡献”,通过句 *a* 中包含的语义片断 a_1“金大中获得和平奖”与全文主题建立联系.因此,孤立地看待子话题,它可以包含多重语义.当子话题作为一个独立整体参与全文匹配时,内部各语义之间的相互关系对系统并不透明;同时,子话题主体内容反映的语义并不一定统一于全文主题,与其他报道的匹配往往误导主题一致性的判断.语义域因其内部语义的统一性,可以屏蔽多义性对主题理解的负面影响,只要两篇报道间语义趋向于一致的语义域在文中都足够重要,则其话题相互关联的可信性相应地会很高.

2.2 语义域凝聚策略

如第 2.1 节所述,语义域是语义片断趋向一致的语境集合,因此语义域的建立需要预先设置语境的抽取原则,篇幅过大的语境将涵盖过多语义,影响语义域的语义一致性,例如将子话题作为语境;相应地,篇幅过小的语境无法完整描述语义,例如以逗号划分的子句(如图 1 中句 *c* 的子句“金大中总统在奥斯陆的颁奖仪式上”).因此,

本文将句号、叹号和问号划分的句子作为语境.一个理想的语境形如图 1 中的句 c,整句的语义统一于“金大中被授予和平奖”.但是,并不是所有句子都具备上述理想结构,比如,句 a 包含“获奖”事件和“获奖”原因两种语义,前者与句 c 语义一致,后者则与句 d 的语义强相关,使句 a 可以同时隶属于不同语义域.因此,通过句子构造语境并进行语义域凝聚的过程中,句子内部语义倾向性的衡量是不可忽视的.针对一篇待测报道 D,语义域凝聚策略以下面 4 步所示.

(1) 将 D 切分为句子集合 $S=\{s_1,s_2,...,s_n\}$,并建立每个句子的语义空间;

(2) 针对 S 中所有句对 $\{(s_i,_j)|s_i,s_j\in S\}$,基于语言模型计算每个句对的相关度 $P(s_i|s_j)$;

(3) 训练阈值 θ,将相关度大于 θ 的句对作为候选语义域嵌入集合 E;检验所有相关度低于 θ 的句对,如果存在与其他句子相关度都低于阈值 θ 的句子,则直接将其设置为语义域.

(4) 根据非传递性假设重组集合 E 中的候选项,形成各种新的语义域.该过程遵循两个必要条件:i) 合并为一个语义域的任何句对,其语义相关度都大于阈值 θ;ii) 语义域之间没有交集,即如果一个句子同时属于多个语义域,则嵌入最相关的语义域.

语义域凝聚策略的核心问题在于集合 E 中候选语义域的重组过程,其基本条件是语义域之间是否满足相关度的传递性.假设集合 E 中两个候选语义域分别为 $e'=\{s_j,s_k\}$ 和 $e''=\{s_j,s_i\}$,即 s_j 同时相关于 s_k 和 s_i,则传递性认为 s_k 和 s_i 也相关.但以句子作为语境并进行语义域凝聚的过程中,该假设并不恒定成立,原因在于某些句子并不具备理想的单义结构.如图 1 中的句 a 包含两个语义片断 a_1 和 a_2,a_1 使句 a 相关于句 c,a_2 所在的子句使句 a 相关于句 d,但句 c 和句 d 的语义并不相关,传递性不成立.因此,凝聚过程必须基于非传递性进行语义域的合并与拆分.如图 2 中左图所示,s_j 同时相关于 s_k 和 s_i,但 s_k 和 s_i 的相关度很低,不适合嵌入同一语义域.因此,凝聚策略将相关度更高的 $e''=\{s_j,s_i\}$ 作为一个语义域,而将 $e'=\{s_j,s_k\}$ 中的 s_j 撤销,将 s_k 独立作为语义域 $e'=\{s_k\}$.在该例中,只有 s_i 和 s_k 的相关度也高于阈值 θ,才可以将三者视为语义一致,并融合成语义域 $e'''=\{s_i,s_j,s_k\}$,如图 2 中右图所示.语义域凝聚的效果依赖于阈值 θ 的估计,θ 过高,则语义趋近一致的句子有可能被错误地拆分,而过低,则将语义不一致的句子凝聚为一个语义域.

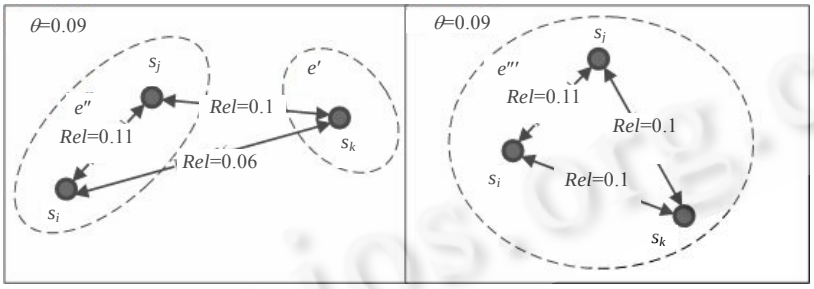

Fig.2 Cohesion of semantic domain based on non-transitive assumption

图 2 基于非传递性假设的语义域凝聚

2.3 语义域描述

语义域采用权重较高的特征集合建立语义空间,权重借鉴 TFIDF 进行计算,区别在于,TF 基于特征在语义域内出现的频率进行统计,本文将这一特征频率表示为 TF_{SD}.非传递性凝聚过程使语义域内的特征频率(TF_{SD})分布相对均匀,从而削弱少量高频特征引起的语义描述偏见性.图 3 是容易引起误判的两篇不相关报道样例,其中,TF 是基于全文统计出的特征频率,图 3 仅列举了 TF 最高的 5 个特征;TF_{SD} 的特征选自最可能成为主题的语义域(语义域作为主题的概率估计见第 3.1 节).如图 3 所示,特征“诺贝尔”在两篇报道中的 TF 值都过高,由它引起的语义偏见增强了报道间的相关性,因此增大了 LDT 系统将两者误判为相关的概率.相对地,TF_{SD} 使所有重要特征在语义描述中发挥相对均衡的作用,有助于屏蔽个别高频特征对语义倾向的误导.

“Kim Daejung was awarded the Nobel Peace Prize”		“Gao Xingjian was awarded the Nobel Prize for literature”	
TF_{SD}	TF	TF_{SD}	TF
（金大中，tf=3） （总统，tf=2） （授予，tf=2） （诺贝尔，tf=2） （和平奖，tf=2）	（诺贝尔，tf=12） （民主，tf=6） （金大中，tf=6） （人权，tf=5） （韩国，tf=4）	（高行健，tf=3） （诺贝尔，tf=3） （文学奖，tf=3） （获得，tf=2） （中文，tf=2）	（诺贝尔，tf=11） （高行健，tf=5） （文学奖，tf=5） （人权，tf=4） （中文，tf=4）

Fig.3 An example of comparison between TF_{SD} and TF

图 3 TF_{SD} 与 TF 对照样例

基于相互独立并离散分布的一元词特征描述语义的方法,往往忽视词与词之间语义上的联系,如“韩国总统金大中获得诺贝尔和平奖”与“美国前副总统戈尔获得诺贝尔和平奖”中,词特征的组成非常类似,检测系统难以敏感地识别两者语义上的差别.针对这一不足,本文尝试通过依存分析技术挖掘词汇之间的关联性,将语义上存在依存关系的二元词对(以下简称依存对)作为语义空间的基本特征,如依存对(金大中,获得).在基于依存分析的语义描述过程中,系统通过分析依存对在依存树中所处的层次衡量其描述语义的能力,并通过依存对中的词特征在语义域中出现的频率 TF_{SD} 评价它的重要性.假设某一语义域中,词 x 与词 y 存在依存关系,则该依存对的权重计算公式如下:

$$W(x,y)=\frac{tf_{SD}(x)}{l}\frac{tf_{SD}(y)}{l}\sum_i \log\left(\alpha+\frac{n_i}{\eta_i}\right) \tag{1}$$

其中,$tf_{SD}(x)$与 $tf_{SD}(y)$分别表示 x 与 y 的特征频率 TF_{SD};l 表示语义域的词特征总数;n_i 表示依存对(x,y)在语义域内的语境 i 中出现,且语境 i 的依存树共有 n_i 层;η_i 为(x,y)在依存树中所处的层次,层次越高,则依存对描述语义的能力越强,如图 4 所示.α表示不同层次间依存对价值差异的平滑因子,α越大,则差异越小.

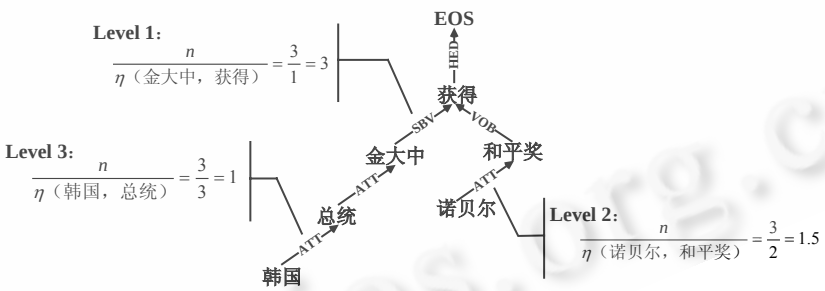

Fig.4 An example of dependency parsing

图 4 依存分析样例

3 语义域语言模型及关联性匹配

3.1 语义域语言模型

语义域凝聚将报道组织为多个语义结构,从而为描述主题语义在报道中的概率分布提供了平台.针对任意报道对,SDLM 以语义域为媒介并基于语言模型描述主题语义的相关性.关联检测旨在判断两篇报道的主题是否一致,而主题语义的相关性是对其一致性的最佳描述.假设两篇待测报道 D_1 与 D_2 的主题分别为 T_1 和 T_2,SDLM 通过统计 T_1 的语义空间在 D_2 各语义域内的概率分布,描述了 T_1 与 T_2 的语义相关性,其公式如下:

$$P(T_1|T_2)=\prod_{t\in T_1}P(t|T_2),\ P(t|T_2)=\sum_{r\in R_2}P(t|r)P(r|D_2) \tag{2}$$

公式(2)的含义是:一篇报道主题的语义空间频繁分布于另一篇报道最可能成为主题的语义域中,则这两篇报道主题相关的可能性很高.该式中,t 为主题 T_1 语义空间中的某一特征;r 为 D_2 中某一语义域;R_2 为 D_2 所有语

义域的集合.$P(t|r)$表示 t 在语义域 r 中的概率分布,一元特征的 $P(t|r)$利用 TF_{SD} 和反文档频率 IDF 进行计算,依存对的 $P(t|r)$如公式(1).$P(t|D_2)$描述 D_2 将语义域 r 作为主题的概率,公式如下:

$$P(r \mid D_2) = \frac{|r|}{\sum_{s \in r} loc(s)} \cdot \log \frac{size(r)}{coll.size} \tag{3}$$

公式(3)是一种经验性的估计,前半部分是语义域 r 中的句子在 D_2 中平均位置的倒数,平均位置越靠前,则该指标越大.其中,$|r|$是语义域 r 中句子的总数,$loc(s)$表示句子 s 在 D_2 中的位置.事实上,由于新闻报道的特性,即希望以最快捷且精确的方式传递信息,往往在开篇便点明文章的主题;同时,凝聚于主题的语义片断既是文章的框架,也是连接各子话题的媒介,往往作为子话题的引导对相关内容进行拓展,因此,在主题对应的语义域中,句子在报道中的平均位置相对靠前.该公式的后半部分描述语义域 r 的篇幅对其作为主题可能性的影响,其中,$size(r)$是 r 包含的特征数量;$coll.size$ 是报道 D_2 包含的特征总数.

3.2 主题语义关联性匹配

在 SDLM 进行主题相关性匹配之前,首先依靠公式(3)对报道 D_1 中的各语义域进行排序;然后,选择排序最靠前的语义域作为主题 T_1 的描述,并根据公式(2)建立主题相关性的语义域语言模型;最后,选择相对熵(Kullback-Leibler divergence,简称 K-L 距离)评估主题间语义的相关性.

概率统计将 K-L 距离用于度量两个正函数是否相似,自然语言处理常将其用于衡量词的同义性或文章内容的相似性.假设两篇报道 D_1 和 D_2 主题的语义域语言模型分别为 M_1 和 M_2,则它们的 K-L 距离计算如下:

$$D_{K-L}(M_1 \| M_2) = \sum_t P(t \mid M_1) \log \frac{P(t \mid M_1)}{P(t \mid M_2)} \tag{4}$$

其中,$P(t|M_1)$与 $P(t|M_2)$分别是特征 t 在模型 M_1 和 M_2 内的概率分布.公式(4)首先针对报道 D_1 抽取主题 T_1;然后,针对 T_1 中的任意特征 t,采用公式(2)统计它在 D_1 和 D_2 各语义域中的概率分布,此时,$P(t|M_1)$与 $P(t|M_2)$即为 $P(t|T_1)$与 $P(t|T_2)$;最后,采用 K-L 距离衡量报道间的主题一致性.因此,该匹配过程存在不均衡现象,通常弥补这一缺陷的方法是双向 K-L 距离的求和运算,即 $D(M_1|M_2)+D(M_2|M_1)$.此外,K-L 距离本质上计算的是两种概率分布的非近似度,因此,模型 M_1 和 M_2 之间的近似性通过原指标的负值进行描述.此外,本文还通过嵌入概率区分度 CP[8](clarity probability),提高 K-L 距离计算的可区分性.SDLM 在主题语义相关度匹配过程中涉及两个参数,一个是主题语义空间的特征数量;另一个是主题语义相关度的阈值,参数估计将在实验部分加以介绍.

4 实验设计

4.1 语 料

实验采用语言数据协会(Linguistic Data Consortium,简称 LDC)提供的 TDT4 语料进行评测.TDT4 由路透社和 CNN 等媒体机构于 2000 年 8 月~2001 年 1 月间播报的新闻报道组成,包含中文、英文和阿拉伯文 3 种语言形式,并涉及文本、音频广播及其翻录 3 种信息表述方式.该实验基于 TDT4 文本形式的中文语料进行评测,其待测索引列表包含 26 066 个新闻报道对,LDC 人工标注 3 075 对报道为相关,其他为不相关.在此基础上,实验抽取 10 000 对新闻报道作为训练语料,其中包含相关报道 1 200 对;其他作为测试语料.

4.2 评测体系

实验基于美国国家标准与技术研究院(NIST)针对 TDT 发布的评测指南,采用检测错误代价 C_{Det} 分别从漏检和误检两个角度进行评测,公式如下:

$$C_{Det} = C_{Miss} P_{Miss} P_{target} + C_{FA} P_{FA} P_{non-target} \tag{5}$$

其中,P_{Miss} 和 P_{FA} 分别表示系统的漏检率和误检率,漏检即为系统未识别出新话题,误检则是系统将旧话题的后续相关报道误判为新话题;C_{Miss} 和 C_{FA} 分别代表漏检和误检的代价系数(C_{Miss}=1,C_{FA}=0.1);P_{target} 和 $P_{non-target}$ 是先验目标概率(P_{target}=0.02,$P_{non-target}$=1−P_{target}).检测错误代价 C_{Det} 的规范化形式 $Norm(C_{Det})$见公式(6).此外,NIST 面

向 TDT 研究提供了可视化的评测工具,即检测错误权衡图(detection error tradeoff,简称 DET).DET 利用二维坐标系的纵轴表示漏检率;横轴表示误检率,并根据漏检与误检随阈值θ的变化趋势绘制系统的性能曲线.由于系统漏检与误检的概率越低,其性能越好,因此,DET 曲线越靠近坐标系的左下角越代表系统性能更优越.DET 曲线上的最小 $Norm(C_{Det})$指标代表检测系统的最佳性能,简写为 Min $Norm(C_{Det})$.

$$Norm(C_{Det})=\frac{C_{Det}}{\min(C_{Miss}\cdot P_{target},C_{FA}\cdot P_{non-target})} \tag{6}$$

4.3 实验流程

基于SDLM建模策略的LDT系统在语言模型的基础上,嵌入基于语义域的主题描述模块.因此,实验主要针对如下 3 个方面的问题进行设计:① SDLM 基于主题语义匹配相关性,相比于采用全文较高权重描述语义的方法是否性能更优?② 与前人工作比较,SDLM 是否改进了 LDT 系统性能?③ 相比于单一词特征,基于依存对描述语义是否提高了 SDLM 的性能?

针对上述问题,实验分别基于一元(unigram)特征和依存对(dependency pair)设计 SDLM:U-SDLM 和 D-SDLM,其共同点在于都需要将报道划分不同语义域,并采用公式(3)抽取主题,最终基于语义域语言模型评估主题相关性.不同点在于,U-SDLM 的语义域凝聚和主题建模过程都以词为基本特征;D-SDLM 则以依存对为特征.此外,实验建立基于一元语言模型(unigram language model,简称 ULM)、一元相关性模型(unigram relevance model,简称 URM)以及二元语言模型(bigram language model,简称 BLM)的 LDT 系统,并与 SDLM 进行比较,其流程如下:

① 基于训练语料对 U-SDLM 和 D-SDLM 的相应参数进行估计;训练 URM 和 BLM 的平滑系数**,分别为 λ[8]和折扣系数 d_r[11],相应指标见第 5.2 节的表 1;

② 测试阶段,实验针对一元特征对比 U-SDLM,ULM 以及 URM 的性能;针对二元特征,实验选择 D-SDLM 和 BLM 进行比较.该测试侧重检验基于语义域的主题挖掘及其匹配对 LDT 性能的影响.在此基础上,实验选择 U-SDLM 和 D-SDLM 进行比较,检验基于依存关系的语义描述是否可以改进 U-SDLM 的性能.

5 实验结果与分析

5.1 SDLM参数估计及分析

SDLM 共涉及如下 3 个参数:依存对权重计算中的平滑因子α;语义域凝聚过程中的阈值θ;主题语义空间的特征个数 N.实验系统 U-SDLM 以词为特征,因此无须训练α,而 D-SDLM 则需要估计所有参数.本节以 D-SDLM 为例论述其参数估计的方法、原理及结果,U-SDLM 中θ与 N 的训练方法与此相同.

(1) 估计平滑因子α

公式(1)中的对数表达式根据依存对在依存树中的层次关系描述其语义表述能力,其中,平滑因子α用于调整不同层次表述语义的能力差异.随着α取值的增大,对数表达式区分层次间语义表述能力的敏感性逐渐减弱,同时,公式(1)更依赖于 TF_{SD}进行语义描述.该实验分别人工标注相关与不相关句对各 1 000 组,根据公式(1)描述句子的语义空间并计算相关度,在此基础上,通过观察相关度均值随α的变化趋势对α进行估计.如图 5 所示,相关和不相关句对的相关度随α的增大逐渐趋近于一致,说明完全依据 TF_{SD} 描述语义的效果并不理想,其区分相关性的能力不强;而α较小时,即当依存树层次关系极大地影响权重描述时,相关性的可区分性较强.实验将α设置为 0.2,该指标在有效区分相关性的同时,相关句对的相关度均值也达到最大值.

(2) 估计语义域凝聚阈值θ

在语义域凝聚过程中,报道随阈值θ的增加被划分为颗粒度更小、数量更多的语义域,与此同时,语义域内句对的平均相关度逐步递增.但阈值θ的持续提高将造成不包含语义域的报道增多,因而报道内语义域的平均数

** ULM 采用 $TF_{SD}.IDF$ 计算 $P(t|M)$,无平滑参数,BLM 基于 MAL 算法统计二元组共现概率.

量在特定阶段开始逐渐减少***,直到阈值过大导致数量及相关度的平均值同时衰减,其衰减起始点分别对应α=0.2 时不相关句对与相关句对的平均相关度(0.065 和 0.15),如图 5 及图 6 所示.凝聚过程获得的语义域越少,单句越多,则规模恒定的主题空间涵盖的语义越多****,把握核心语义的能力越差.此外,语义域内句对相关度偏低,同样也将造成主题语义描述的偏差.因此,该实验为了保证语义域内句对足够相关,同时能为报道维护更多的语义域,将区间[0.068,0.155]作为阈值θ训练的候选集合.在此基础上,该实验采用正态分布拟合数量及相关度随阈值的变化趋势,并基于其联合分布估计阈值θ,其指标为 0.09.

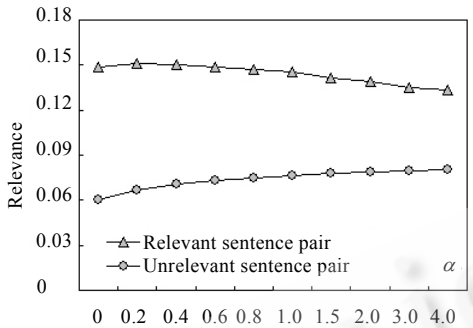

Fig.5 The change trend of correlation between sentence pair with smoothing factor α

图 5 句对相关度随平滑因子α变化趋势

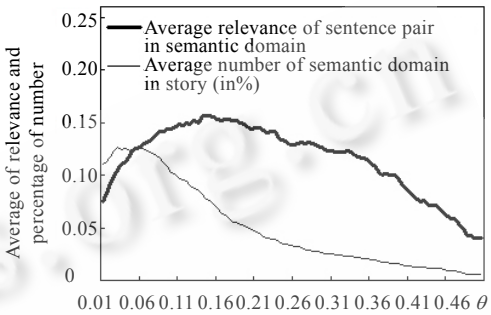

Fig.6 The trend of correlation between sentence pair and semantic domain scale changes with factor θ

图 6 句对相关度及语义域数量随θ变化趋势

(3) 估计主题特征数 N

SDLM 基于语义域的概率分布挖掘主题并描述其语义,通过语言模型衡量报道之间的相关性.语言模型的公式化表述是一系列条件概率的连乘,这就要求参与匹配的特征空间需要保持一致的规模,即特征数量相同,从而使相关性指标可以在统一标准下进行比较.但 SDLM 抽取得到的主题规模并不相同,以 D-SDLM 为例,图 7 是其α=0.2 和θ=0.09 时主题规模的概率分布情况,横轴为主题包含的特征数 N;纵轴为对应某一 N 值的报道数在语料中的百分比,大部分报道的 N 值分布于[20,60]区间内,实验将其作为训练 N 值的候选集.

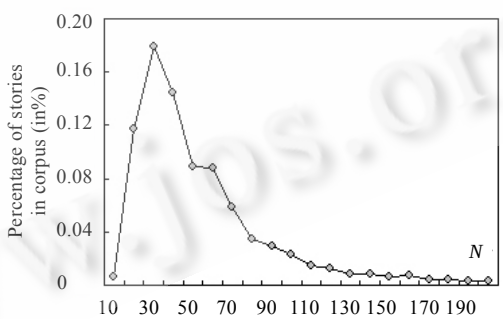

Fig.7 The distribution trend of dependency pair in topic mined by SDLM

图 7 SDLM 主题挖掘包含依存对的分布趋势

训练过程以 5 为颗粒度逐步调整 N 值,如果待测报道的主题规模大于 N,则将描述主题的语义域中权重略低的特征忽略,反之则利用全篇报道中权重较高的特征进行补充,图 8 是 D-SDLM 系统在区间[20,60]内的性能

*** 图 5 是针对包含两个以上句子的语义域获得的训练数据,不涉及单句组成的语义域,因此语义域平均数量存在衰减过程.

**** 单句形成的语义域往往包含较少特征,如果将其作为规模恒定的主题,则需要从全文中抽取权重较高的特征进行补充(详情参考下文"估计主题特征数 N"),从而破坏了主题空间的单义特性.

变化趋势(*N* 值为图中曲线标识内的阿拉伯数字)*****.如图 8 所示,系统性能在 *N* 值偏低和偏高时都不理想,原因在于 *N* 值偏低造成特征空间的数据稀疏,偏高则嵌入更多主题以外的高权重特征,削弱了主题挖掘的作用,这一现象验证了高权重特征并不总是更善于描述主题的假设.实验将 D-SDLM 中的 *N* 值设置为 25.

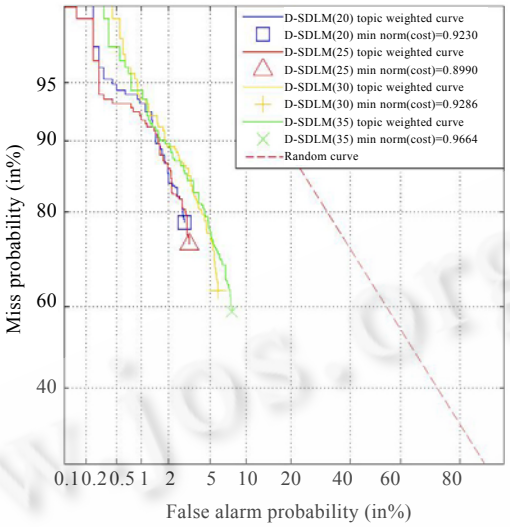

Fig.8 The performance trend of D-SDLM system changing in scale sector [20,60] of terms

图 8 D-SDLM 系统在特征数区间[20,60]内的性能变化趋势

5.2 DET测试结果及分析

第 4.3 节所述,实验首先面向一元特征测试 U-SDLM,ULM 以及 URM 的性能,如图 9 中空心几何标识及其所在曲线,相应参数和最小 DET 代价见表 1.该测试验证了 SDLM 有益于报道间的相关性检测,原因在于 U-SDLM 面向报道的主题建立话题模型,其在描述核心语义的同时,屏蔽了作为相关外延的其他内容对话题模型的误导.与此相对,ULM 中特征的语言模型 $P(t|M)$以全文为参照系,忽视了报道内部语义结构对篇章理解的作用.以图 1 中的报道为例,ULM 建立的话题模型中包含特征“诺贝尔”和“金大中”,它们的共现关系蕴含了重要的语义信息,有助于相关报道的匹配;但由于上述特征在报道中出现的频率约为 12:6,当反文档频率(IDF)指标近似时,ULM 赋予特征“诺贝尔”的 $P(t|M)$指标远大于“金大中”,其对于区分诸如题为“高行健获得诺贝尔奖”等不相关报道将起到负面作用.这一现象源于图 1 的子话题 3 中频繁出现特征“诺贝尔”,但子话题 3 并不是主题,从而误导话题模型侧重点的偏移.U-SDLM 则根据语义域的相关性以及主题挖掘的概率模型(如公式(3)),提取图 1 中的语义域 *a-c-e* 作为主题,并在此基础上建立报道的话题模型,特征“诺贝尔”和“金大中”在话题模型中的 $P(t|M)$趋近于 2:3,从而在保证主题语义的情况下,降低了相关外延的误导作用.因此,U-SDLM 的性能提高主要来自误检率的降低.URM 通过检索相关报道,并基于各报道共有的高权重特征建立话题模型,其本质上也是一种主题语义的挖掘过程,有助于 LDT 系统降低误检率.但 URM 过分依赖语料的规模和当前报道的时序,较小的语料规模和较早的时序都将降低检索质量,从而造成话题模型融入更多噪声,降低 LDT 系统的性能.此外,检索过程造成的时间消耗极大地降低了 URM 的系统效率.

实验面向二元特征对 D-SDLM 与 BLM 进行比较,其性能如图 10 所示,相应参数和最小 DET 代价见表 1.实验结果显示,BLM 的误检率较低,而漏检率很高;相对地,D-SDLM 在误检率损失不大的情况下,显著降低了漏检率.影响 BLM 性能的主要因素更多地源于主题语义的描述偏差,而不是二元特征对语义的表述能力.为验证

***** 图 8 中 DET 曲线存在不完整现象,其原因在于大量报道对相关度为 0,当阈值低于某一非零最小相关度时,漏检率和误检率提前达到极值,即漏检率为 0,误检率为 1,并且在阈值下调的过程中保持恒定,从而造成 DET 曲线直指坐标系右下角.由于此时系统的 DET 代价最大,不影响最小代价的标注,为了便于实验结果的观测,本文略去其在图中的轨迹.

这一推论,图 10 附加了基于依存关系的二元语言模型(dependency parsing BLM,简称 D-BLM)在 LDT 系统中的性能.D-BLM 与 BLM 的区别在于,前者基于依存关系抽取二元组,后者则将文本中连续的词对作为二元组.相比于 BLM,D-BLM 降低了二元组的随机性,有助于语义表述的精确性.但结果显示,D-BLM 的性能提高得并不明显,同样存在高漏检现象,只有基于语义域挖掘主题语义形成 D-SDLM 后才会扭转这一缺陷.

Table 1 Parameter and minimum normal DET cost

表 1 参数及最小规范化 DET 代价

LDT system	Smoothing parameters	θ	N	Min *Norm* (*Cost*)
U-SDLM	–	0.35	55	0.682 2
ULM	–	–	55	0.711 7
URM	λ=0.99	–	65	0.726 2
D-SDLM	α=0.2	0.09	25	0.898 1
BLM	d_t=0.9	–	30	0.920 9

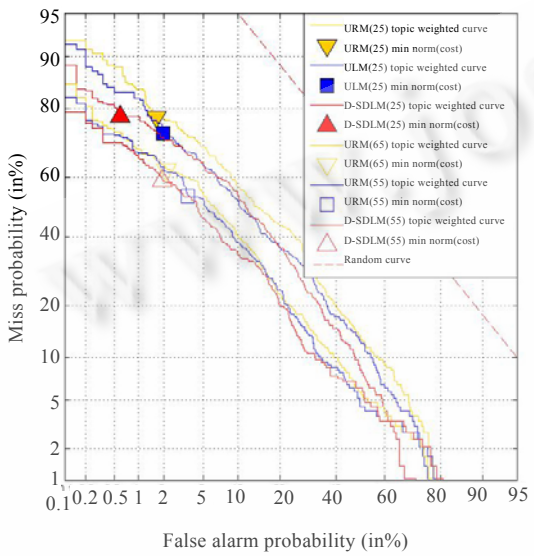

Fig.9 Performance comparison among the testing unigram models

图 9 基于一元特征各测试模型的性能对比

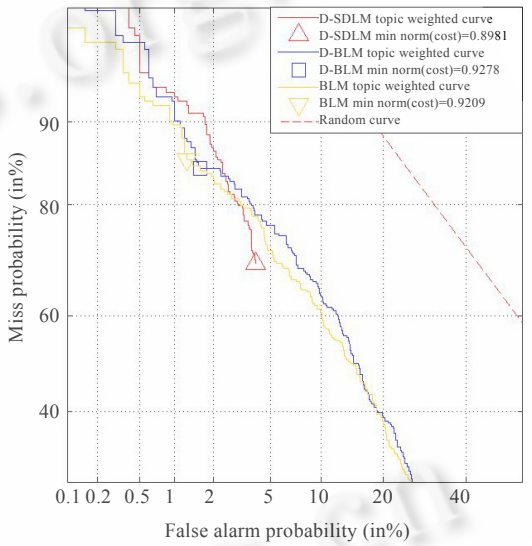

Fig.10 Performance comparison among the testing bigram models

图 10 基于二元特征各测试模型的性能对比

虽然 D-SDLM 提高了二元语言模型的性能,但相比于 U-SDLM 等一元模型,其性能明显偏低,见表 1.其原因在于语义表达形式多样化,同一语义的不同依存对无法进行匹配,如依存对(授予,和平奖)与(荣获,和平奖),从而遗漏了部分语义信息,这也是上述 D-BLM 对 BLM 改进不大的原因.但实验发现,D-SDLM 最佳性能对应的特征数(25 个依存对)明显少于 U-SDLM 等一元语言模型,见表 1.U-SDLM 采用这一特征数(25 个词特征)的性能大幅降低,如图 9 中实心标志及其所在曲线所示,此时,D-SDLM 的漏检率小于一元语言模型,而误检率相对略高.因此,基于依存关系描述语义对 SDLM 性能的提高仍有很大的可拓展空间,一种改进设想是,基于词典识别依存对中同义不同形的词项,但该方法不仅增加了匹配过程的开销,也对词典的规模具备依赖性.另一种方法是,建立一元模型和二元模型的融合机制,如基于线性组合的折衷策略.

6 结 论

针对 TDT 领域中的关联性检测任务,本文基于语义域及其概率分布挖掘主题,并融合依存关系辅助其语义描述,在此基础上建立话题模型参与新闻报道相关性的匹配.实验验证基于语义域语言模型进行话题建模改进了传统一元和二元语言模型的检测性能.

关联检测领域主要涉及两方面研究,一方面是以向量空间模型、语言模型和相关性模型等为主流的统计策

略,另一方面则在此基础上融合命名实体识别、词扩展和句法分析等自然语言处理技术.结合前人工作和本文实验观察,自然语言处理对关联检测系统的性能提高有限,不仅体现了模型设计本身的缺陷,也反映了目前自然语言处理技术在处理大规模数据中仍有很大可提升的空间.因此,针对该方向的后续研究将主要包含如下几个方面:(1) 更有效地结合自然语言处理技术提高现有模型描述语义的能力;(2) 在提高篇章理解能力的基础上改进现有话题建模策略;(3) 建立模型之间的融合机制以发挥各自的优点.

References:

[1] Allan J. Topic Detection and Tracking: Event-Based Information Organization. Springer-Verlag, 2002. 1–16.

[2] Yiming Y, Ault T, Pierce T, Lattimer CW. Improving text categorization methods for event tracking. In: ACM, ed. Proc. of the SIGIR 2000. Athens: Association for Computing Machinery Press, 2000. 65–72.

[3] Luo WH, Liu Q, Chen XQ. Development and analysis of technology of topic detection and tracking. In: Sun MS, ed. Proc. of the JSCL-2003. Beijing: Tsinghua University Press, 2003. 560–566 (in Chinese with English abstract).

[4] Yu MQ, Lou WH, Xu HB, Bai S. Research on hierarchical topic detection in topic detection and tracking. Journal of Computer Research and Development, 2006,43(3):489–495 (in Chinese with English abstract).

[5] Kumaran G, Allan J. Text classification and named entities for new event detection. In: ACM, ed. Proc. of the SIGIR 2004. New York: Association for Computing Machinery Press, 2004. 297–304.

[6] Farahat A, Chen F, Brants T. Optimizing story link detection is not equivalent to optimizing new event detection. In: Isahara H, ed. Proc. of the ACL-03. Sapporo: Association for Computational Linguistics Press, 2003. 232–239.

[7] Allan J, Carbonell J, Doddington G, Yamron J, Yiming Y. Topic detection and tracking pilot study final report. In: Proc. of the Broadcast News Transcription and Understanding Workshop, Vol.2. 1998. 1–25.

[8] Lavrenko V, Allan J, Deguzman E, Laflamme D, Pollard V, Thomas S. Relevance models for topic detection and tracking. In: Proc. of the 2nd Int'l Conf. on Human Language Technology Research. San Francisco: Morgan Kaufmann Publishers, 2002. 104–110.

[9] Ponte J, Croft WB. Text segmentation by topic. In: Peters C, ed. Proc. of the European Conf. on Research and Advanced Technology for Digital Libraries. London: ECDL Press, 1997. 113–125.

[10] Nallapati R. Semantic language models for topic detection and tracking. In: Hearst M, ed. Proc. of HLT-NAACL2003 Student Research Workshop. Edmonton: Association for Computational Linguistics, 2003. 1–6.

[11] Chen SF, Goodman J. An empirical study of smoothing techniques for language modeling. Computer Speech and Language, 1999,13(4):310–318.

附中文参考文献:

[3] 骆卫华,刘群,程学旗.话题检测与跟踪技术的发展与研究.孙茂松,陈群秀.见:全国计算语言学联合学术会议(JSCL-2003)论文集.北京:清华大学出版社,2003.560–566.

[4] 于满泉,骆卫华,徐洪波,白硕.话题识别与跟踪中的层次化话题识别技术研究.计算机研究与发展,2006,43(3):489–495.

洪宇(1978－),男,黑龙江哈尔滨人,博士生,CCF 学生会员,主要研究领域为话题检测与跟踪,信息过滤,个性化信息检索.

刘挺(1972－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为自然语言处理,信息检索.

张宇(1972－),男,博士,副教授,CCF 高级会员,主要研究领域为信息过滤,自动问答,自然语言处理.

李生(1943－),男,博士,教授,博士生导师,CCF 会员,主要研究领域为自然语言处理,信息检索,机器翻译.

范基礼(1986－),男,本科生,主要研究领域为话题检测.