

基于双语主题和因子图模型的汉语-越南语双语事件关联分析

唐莫鸣¹ 朱明玮¹ 余正涛¹ 王吉地¹ 高盛祥¹

(1. 昆明理工大学信息工程与自动化学院, 云南昆明 650500)

摘要: 随着一带一路国家战略实施, 我国与越南的交流与合作日益密切, 及时掌握两国新闻事件动态意义重大。该文针对汉越双语新闻事件关联分析所面临的跨语言关联问题, 研究汉越双语新闻事件关联分析方法。汉越双语新闻事件分析其实是多语言多文本的理解问题。其主要难点是要解决多语言多文本下的新闻事件理解问题。该文提出了基于因子图模型的局部密切度传播算法。首先使用双语主题概率模型, 从双语文档中获得双语主题以及主题概率分布。然后基于新闻事件的文本相似度构建事件因子图模型, 在因子图上对相互关联的事件使用局部密切度传播算法计算某一主题下所有相互关联的事件间的影响力。最后得到不同主题下事件间的影响力拓扑图。实验结果表明该文提出的方法相比相似度计算和词语共现的方法取得了不错效果。

关键词: 汉越双语新闻事件; 事件关联; 多语言文本;

中图分类号: TP391

Chinese-Vietnamese Bilingual Event Correlation Analysis Based on Bilingual Topic and Factor Graph

TANG Moming¹, ZHU Mingwei¹, YU Zhengtao¹, WANG Jidi¹ and GAO Shengxiang¹

(1. The School of Information Engineering and Automation, Kunming University of Science and Technology Kunming, Yunnan 650051, China)

Abstract: With the implementation of National One Belt, One Road Strategy, exchange and cooperation of China and Vietnam have obtained the positive result. In view of the cross-lingual correlation problem faced by the analysis of Chinese and Vietnamese bilingual news events, this paper studies the correlation analysis method of Chinese and Vietnamese bilingual news events. The essence of Chinese and Vietnamese bilingual news events correlation analysis is multi-lingual and multi-text understanding. The major difficulty is to solve multi-lingual and multi-text news events understanding problem. In this paper, we proposed a local intimacy propagation algorithm based on factor graph. First, we use bilingual topic model to get the bilingual topics and topic probabilistic distributions from bilingual document. Then we built events' factor graph based on event text similarity. Using local intimacy propagation algorithm to compute influence the for interrelated events on the factor graph under the same topic. Finally we got the influence topology of events under different topics. Experiments results show that the method we propose have achieved better effect compared to the traditional method.

Keywords: Chinese and Vietnamese bilingual news events; events correlation; multi-lingual text;

1. 引言

互联网技术的快速发展使得信息的采集和传播速度达到了空前的水平, 网络舆情分析已经逐渐成为网络信息监测、监控及预警分析的重要手段, 互联网上每天都有大量的新闻事件报道, 事件通过互联网进行快速的传播, 如何快速的掌握互联网新闻事件动态, 把握事件间的关联, 分析其事件间内在关系, 已经逐渐成为政府、企业和社会关注的问题。越南与我国毗邻, 国家一带一路战略大环境下, 越南与国内交流密切, 相关的新闻事

收稿日期: 2017-07-10

定稿日期: 2017-07-25

基金项目: 国家自然科学基金(61472168、61175068); 云南省自然科学基金重点项目 (No. 2013FA130), 云南省科技创新人才基金项目(No. 2014HE001)

作者简介: 唐莫鸣 (1993—), 男, 硕士研究生, 自然语言处理; 朱明玮 (1992—), 女, 硕士研究生, 自然语言处理; 余正涛 (1975—), 男, 博士生导师, 昆明理工大学信息工程与自动化学院教授, 自然语言处理与信息检索, 机器翻译。

件越来越多，而这些报道分布在国内及越南相关网站及媒体上，表现为中文或者越南文，如何能够及时有效了解国内及越南国家的新闻事件动态，掌握事件间的关系，正确做出有效应对措施，处理好与越南的国际关系对区域经济发展、政治稳定、文化交流及商务合作等方面有着重要的作用。新闻对事件的描述中包含了时间、地点、人物等。事件关联是事件之间的逻辑关系，是事件之间固有的一种客观存在。新闻事件关联分析通常被看作融合上下文信息和知识库的相似度计算问题，其中事件可以用词、句子或者文本进行表征。事件间关联分析涉及很多层次。现有研究对事件关系进行了初步的定义和类别划分。其中事件关系识别是一种针对“事件间逻辑关系存在与否”进行自动判定的浅层事件关系检测任务。文献^[1]对两个包含对事件关系描述的句子进行依存句法分析，根据子句间的依存关系判断两个事件是否有联系。事件关联分析中的因果关系的研究由来已久，科学家在哲学与逻辑学上均围绕着这个问题进行了更深入的探讨，但直到近代因果关系才在计算机领域中有了定性和定量的研究。文献^[2]利用上下文中的词语对事件进行表征，根据词语所在的上下文对事件间的因果关系进行判断。将事件之间的因果关系识别视为序列标注问题，利用 CRF 进行识别。文献^[3]解决的问题同^[1]，将因果关系的识别看成二分类问题，根据上下文中的词采用有监督学习的方式对因果关系进行识别。文献^[4]将对事件进行描述的句子看成一个事件，根据不同句子中包含的不同事件要素之间的 PMI 值来计算不同事件之间因果关系的关联强度。但计算事件间的 PMI 值非常耗时，而且事件对可能是通过主事件关联的，单独计算这两个事件的 PMI 值意义不大。文献^[5]将因果关系识别形式化为约束优化问题。将描述不同事件的句子之间事件触发词的语义联系和因果关系指示词融入整数线性规划（ILP）框架，对两个事件间存在因果关系的可能性进行计算。文献^[6]将名词作为事件指示词，并将之看作事件。采用关系抽取的方法对文本中词语之间因果关系进行抽取。文献^[7]将新闻中所包含的事件分为粗粒度事件和细粒度事件，结合浅层语义分析方法和知识库对细粒度事件的因果关系进行识别，并以此推断粗粒度事件的因果关系。但事件间的关联关系不仅仅是因果关系。为了更准确的量化事件间的关联程度，文献^[8]用词语来表征一个事件，通过词语共现的方式衡量事件间的关联强度。但这种方法非常耗时，并且容易引入一些琐碎的没有价值的事件对。文献^[9]借助互信息分析不同新闻事件中所描述的参与者、时间、地点等事件要素之间的关联强度，在此基础上融合新闻的时序关系以及新闻文本内容的相似度计算新闻事件间的相互影响力。文献^[10]利用两个对事件进行描述的句子中谓词在因果语料库中的共现次数来判断事件间的因果关系。文献^[11]根据对不同事件进行描述的句子中包含的谓词和共同实体以及谓词与实体之间的关系，利用语义角色标注的方法对事件因果关系进行识别。

现有事件关联关系分析方法大多是将文章中的词和句看作事件，用机器学习的方法对同一篇文章中词与词、句与句之间的事件关联进行分析，对不同文档中事件关联关系分析涉猎较少。而且很多方法基于词法、句法、篇章结构、文本线索提取事件的因果关系。但仅仅提取事件因果关系是不够的，因为事件关联不仅局限于因果关系。而且基于文本的方法也很难使发现事件间隐含的关联。事件的发生和演化具有关联性，新闻事件的影响力具有传播特性。两个事件是否关联往往受到与这两个事件相关联的其他事件的影响。在社交网络的影响力分析中，某一用户的行为可能导致他的朋友以类似的方式表现某种现象，这与事件关联分析类似。文献^[12]提出的 Measure Influence via Reachability，利用用户之间的关注关系，借助随机游走算法构建用户关联图。通过图上的路径计算用户之间的相互影响力。但该方法只考虑了社交网络的结构。文献^[13]将用户看成节点使用 simRank 的算法用同样的方法构建图模型，根据用户的近邻节点计算任意两个节点的相似度。即任意两个节点的相似度都跟他们两个任意两个邻居的相似度成正比。但该方法同样只考虑了社交网络的结构。文献^[14]提出基于因子图的社交网络影响力分析方法，将用户之间的关注关系和用户文本内容的主题以及不同主题下微博内容之间的联系同时融入到因子图模型中进行影响力计算。取得了非常好的效果。新闻事件关联分析与社交网络类影响力分析类似。事件同样不是孤立存在的，总是与其他事件存在逻辑关联。一个事件的发生会在一定程度影响另一事件的发生或者说以某种形式控制另一件事的发生。但是很多情况下新闻事件之间的关联没有显式表现出来，需要我们根据事件上下文分析。因此我们在双语事件关联分析中借用社交网络影响力分析方法，参考文献^[14]提出的社交网络影响力传播分析方法解决事件间的关联分析。抽取双语事件的共享主题，利用描述事件的文本间相似度构建因子图模型，求解不同主题下事件间的相互影响。

2. 跨语言主题提取

文献^[15]通过双语词典为传统 PLSA 算法的似然函数添加软约束，从未对齐的双语文本中抽取双语文档共享的主题。以及文档在主题上的概率分布。PLSA 算法的似然函数如(1)所示。

$$L(C) = \sum_{i=1}^s \sum_{d \in c_i} \sum_w c(w, d) \log \sum_{j=1}^k P(\theta_j | d) P(w | \theta_j) \quad (1)$$

其中 d 表示文档， θ_j 表示文档 d 的主题。 $P(\theta_j | d)$ 表示以概率 $P(\theta_j | d)$ 选中文档 d 的主题 θ_j 。 w 表示文档 d 的主题词。 $P(w | \theta_j)$ 表示以概率 $P(w | \theta_j)$ 产生主题 θ_j 的主题词 w 。 $c(w | d)$ 表示单词 w 在文档 d 中出现的概率。我们利用双语词典构建汉语-越南语的词汇二部图 $G_{cv} = (V_{cv}, E_{cv})$ ，其中 c 表示汉语， v 表示越南语。如果汉语词汇 V_c 和越南语词汇 V_v 语义相关，则将这两个单词以边 e_{cv} ，边上的权重为 $w(c, v)$ 。我们通常用最大似然估计来估计 $L(C)$ 的参数以及获取主题。我们为 PLSA 的似然函数添加双语约束 $R(C)$ ，如公式(2)所示。

$$R(C) = \frac{1}{2} \sum_{\langle u, v \rangle} w(u, v) \sum_{j=1}^k \left(\frac{p(w_u | \theta_j)}{\text{Deg}(u)} - \frac{p(w_v | \theta_j)}{\text{Deg}(v)} \right)^2 \quad (2)$$

$L(C)$ 表示汉语词和越南词的语义差异，其中 $\text{Deg}(u)$ 代表单词 u 的所有入度边的权重和。将之看作损失函数，来优化 PLSA 特征函数的求解，PLSA 获得的双语主题是语义相关的。

$$O(c, G) = (1 - \lambda)L(C) - \lambda R(C), \lambda \in (0, 1) \quad (3)$$

我们使用 EM 算法求上述公式(3)所示目标函数的参数极大似然估计。在 EM 算法 E 步的时候我们使用如公式(4)估计事件文本的主题以及主题概率分布。

$$z(w, d, j) = \frac{p(\theta_j | d) p(w | \theta_j)}{\sum_j p(\theta_j | d) p(w | \theta_j)} \quad (4)$$

在 EM 算法的 M 步，重新估计分布参数，使数据的似然性最大。从而在非对齐的双语文档中挖掘潜在主题。我们输入非对齐的双语文档集合和双语词典。能获得双语文档共享的主题 $Z = \{z_1, z_2, \dots, z_T\}$ 以及双语文档在共享主题上的概率分布 $\{\theta_i^z\}_{z=1}^T$ ，即事件 i 在不同主题下的概率分布。

3. 汉语-越南语双语事件关联关系分析

事件关联分析就是分析描述不同事件的文本以获取事件之间的关系。其中事件可以用词，句子或者文本进行表征。本文针对汉语越南语双语新闻事件关系识别是对“事件间逻辑关系存在与否以及关联程序强弱”进行自动判定和计算的关系检测任务。

双语新闻事件之间的关联关系可以通过新闻文本的文本相似性度量。但是由于相互关联的双语事件间往往通过某种主题进行关联，而且在该主题下两个事件是否关联往往受到与这两个事件相关联的其他新闻事件的影响。因此我们计算相同主题下两个双语新闻事件关联时不仅要考虑新闻文本之间的相似度还要考虑这两个事件受到与之关联的其他新闻事件的影响。同一主题下双语事件关联计算与社交网络中用户影响力传播的问题相似。在社交网络影响力传播的计算中，认为用户的影响力通过用户之间的关注关系传播，利用社交网络中用户之间的关注关系构建基于用户的因子图模型，将用户视为因子图中的节点，用户之间存在关注关系表示为两个节点之间存在一条边。在因子图模型上通过用户之间的关注关系计算社交网络中节点之间影响力传播的问题。相比于传统图模型，因子图模型考虑了节点自身的属性，节点之间的关联信息以及全局约束信息，利用因子图模型的这些优势可以使我们计算事件之间关联强度时充分利用新闻事件影响力传播特性。因此我们基于因子图模型构建事件间的影响力传播模型。通过新闻事件文本的相似性构建事件因子图，即文本相似度大于一定阈值

则认为两事件之间存在关联。将新闻事件表征为因子图上的节点，相互关联的两事件的节点之间存在一条边关联。利用事件因子图中节点的特征函数，使用社交网络中的局部密切度传播算法，计算图中相互关联的节点之间影响力。最终得到不同主题下事件间影响力传播拓扑图。

主题相似度越高的新闻所报道和关注事件越接近，这些事件的发生和发展往往是由主题相关其他事件导致的。因此主题相近的新闻所报道的事件的关联程度比主题不同的新闻报道的事件高。因此本文是利用主题模型分析出来的主题词来表征新闻事件。利用新闻文本间的主题相似性来度量两个新闻事件的文本相似性。两个新闻的主题相似性越接近，新闻文本相似性也就越高。如果新闻主题相似性越疏远，新闻文本相似性也就越低。本文使用 JSD(Jensen-Shannon Divergence)距离的倒数来度量事件间的主题相似性。如果两个新闻文本主题分布的 sim 越大，主题之间的相似度也就越高则新闻报道的事件存在关联性的可能性越大。我们选取阈值 ε ，通过阈值来判断不同新闻事件是否存在关联关系。即两新闻文本的主题相似度大于 ε ，则两新闻事件存在关联；若两新闻文本的主题相似度小于 ε 则两新闻事件之间不存在关联。文本主题相似度计算如公式 (5) ~ (8)：

$$\text{sim}(\theta_i, \theta_j) = \frac{1}{\text{JS}(\theta_i, \theta_j)} \quad (5)$$

$$\text{JS}(\theta_i || \theta_j) = \frac{1}{2} D_{\text{KL}}(\theta_i, \theta_k) + \frac{1}{2} D_{\text{KL}}(\theta_j, \theta_k) \quad (6)$$

$$D_{\text{KL}}(\theta_i || \theta_j) = \sum_{k=1}^k p(z_k | \theta_i) \log_2 \frac{p(z_k | \theta_i)}{p(z_k | \theta_j)} \quad (7)$$

$$\theta_k = \frac{1}{2} (\theta_i + \theta_j) \quad (8)$$

θ_i, θ_j 表示事件 i, j 的主题概率分布。 $\text{JS}(\theta_i || \theta_j)$ 是事件 i, j 的 JS 距离。 $D_{\text{KL}}(\theta_i || \theta_j)$ 是事件 i, j 的 KL 距离。 θ_k 是 θ_i, θ_j 的均值。

我们通过事件之间的影响力传播方法来进一步判断通过事件文本相似度得到的关联事件是否真正关联。本文在计算同一主题下双语事件影响力大小时参考社交网络影响力传播分析的方法。使用基于因子图模型的局部影响力传播算法。自动识别和量化相同主题下，因子图模型中的事件间关联关系。因子图模型将节点属性信息和节点之间的信息结合起来，在识别节点关联时能提供更多信息。因子图由以下变量组成：一组观察值 $\{v_i\}_{i=1}^N$ ，每个观测值都有隐藏变量 $\{y_i\}_{i=1}^N$ 以及特征函数：局部因子图模型有三个特征函数节点的特征函数，边的特征函数和全局特征函数。节点特征函数描述了节点的基本信息，边的特征函数描述了因子图模型中节点的相互依赖关系，全局特征函数确定了因子图上的全局约束。我们设定事件节点对应因子图中的观测值（即因子图中的节点），事件之间的关系则对应因子图中的边，表 1 总结了局部密切度传播算法用到的符号。

我们对输入的文档集合两两求共享主题，最终将所有求得的共享主题汇总，找出所有文档集合都共有的主题作为主题集合 Z 。将每个输入文档集合（每个集合描述相同事件）视为事件节点，事件节点的集合 $V = \{v_1, v_2, \dots, v_n\}$ 。如果在主题 z 下，任意两个事件节点 i, j 的相似度 $w_{ij}^z = \theta_j^z \text{sim}(\theta_i, \theta_j)$ 大于给定的阈值 ε ，则认为在主题 z 下节点 i, j 相互关联，即节点 v_i 和 v_j 所对应的隐含变量 y_1 和 y_2 之间存在一条边 e_{ij} 将这两个节点关联起来，并且定义边上的权重为 w_{ij}^z 。通过计算主题 z 下节点集合 V 中任意两个节点间的相似度，我们将获得主题 z 下边的集合 $E = \{e_{ij}\}_{i, j \in V}$ 。通过上述计算我们将获得事件的拓扑图 $G(V^z, E^z)$ 。如图 1，构建主题 T_1 下，事件的因子图模型。

表 1 符号说明

符号	描述
N	因子图中事件节点的个数
M	因子图中节点间边的个数
T	主题个数
V	因子图中节点的集合
E	因子图中边的集合
v_i	单个事件节点
y_i^z	主题 Z 下对节点 v_i 有较高影响力的邻接节点
θ_i^z	主题 Z 是由节点 v_i 生成的概率
e_{st}	节点 v_s 和节点 v_t 之间的边
w_{st}^z	主题 Z 下, 边 e_{st} 上的权重
μ_{st}^z	主题 Z 下节点 v_s 对节点 v_t 的影响力

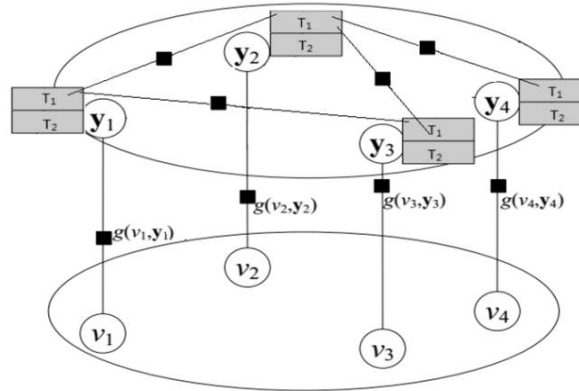


图 1 主题 T1 下事件的因子图模型

图 1 主题 1 下事件因子图 $\{v_1, v_2, v_3, v_4\}$ 节点代表事件, $\{y_1, y_2, y_3, y_4\}$ 是分别对应节点 $\{v_1, v_2, v_3, v_4\}$ 的隐含变量, 集合 $\{y_1, y_2, y_3, y_4\}$ 中的每个元素对应从双语文档中抽取的双语主题的主题概率分布。

$g(v_i, y_i, z)$ 对应节点 v_i 在主题 z 下的特征函数。 $g(v_i, y_i, z)$ 是在主题 z 下定义在节点 v_i 上的特征函数, 描述节点的基本信息。这里我们将节点定义为 $g(v_i, y_{ik}, z)$ 如公式 (9)

$$g(v_i, y_{ik}, z) = \frac{w_{iy_i^z}^z}{\sum_{j \in \text{NB}(i)} (w_{ij}^z + w_{ji}^z)}, y_i^z \neq i \quad (9)$$

$\text{NB}(i)$ 表示与节点 v_i 存在边相连的相邻节点。 w_{ij}^z 表示在主题 z 下, 节点 v_i 和 v_j 的相似性。如果节点 v_i 和节

点 v_j 在主题 z 下有着较高的相似度那么可以说 v_j 对节点 v_i 有着较高的影响力。我们提出的密切度传播算法将因子图中信息传递的规则转化成等价的更新规则，信息将会在节点之间传递而不是在因子图上传递。为了解决因子图模型时间复杂度高的问题，我们将节点的特征函数归一化如公式 (10)

$$b_{ij}^z = \log \frac{g(v_i, y_i, z) \Big|_{y_i^z=j}}{\sum_{k \in \text{NB}(i) \cup \{i\}} g(v_i, y_i, z) \Big|_{y_i^z=k}} \quad (10)$$

我们提出的算法引了两组变量 $\{r_{ij}^z\}_{z=1}^T$ 和 $\{\alpha_{ij}^z\}_{z=1}^T$ 代表事件节点之间的影响力。更新的规则如公式 (11) (12)

(13)

$$r_{ij}^z = b_{ij}^z - \max_{k \in \text{NB}(j)} \{b_{ik}^z + \alpha_{ik}^z\} \quad (11)$$

$$\alpha_{jj}^z = \max_{k \in \text{NB}(j)} \min\{r_{kj}^z, 0\} \quad (12)$$

$$\alpha_{ij}^z = \min(\max\{r_{ij}^z, 0\}, -\min\{r_{kj}^z, 0\} - \max_{k \in \text{NB}(j) \setminus \{i\}} \min\{r_{kj}^z, 0\}), i \in \text{NB}(j) \quad (13)$$

$\text{NB}(j)$ 是与节点 j 存在边相连的邻居节点， r_{ij}^z 表示事件 i 发生对事件 j 发生的影响（即节点 i 对节点 j 的影响）， α_{ij}^z 表示事件 j 发生对事件 i 发生的影响（即节点 j 对节点 i 的影响），这两个值初始为 0。最终我们基于变量 \mathbf{y} 和 α 使用 Sigmoid 函数定义事件间影响力分数如公式 (14)

$$\mu_{st}^z = \frac{1}{1 + e^{-(r_{st}^z + \alpha_{st}^z)}} \quad (14)$$

算法如表 2 所示。对于图中的每个节点 v_t 计算其每个邻接节点 $s \in \text{NB}(t) \cup \{t\}$ 与其自身的影响力分数 μ_{st}^z 。

对于每个主题 z 我们过滤掉那些相关度低的节点（即该节点的主题分布概率低于阈值 ε ）。然后对于图 $G(\mathbf{V}, \mathbf{E})$ 中任意一对有边相连的节点对 (v_s, v_t) ，将节点 v_s 对节点 v_t 的影响力 μ_{st}^z 和节点 v_t 对节点 v_s 的影响力 μ_{ts}^z 相加求平均值，即得到节点 v_s, v_t 之间的关联强度。最终我们得到主题 z 事件影响矩阵 $G_{N \times N}^z$ 。对所有双语主题使用局部密切度传播算法得到所有主题下的事件影响矩阵，将求得的所有主题下事件之间的关联强度相加，求平均值就得到事件之间的影响力矩阵。

表 2 相同主题新闻事件关联算法

算法 1. 相同主题新闻事件关联算法.	
输入：	主题 Z 下的事件因子图 (V^z, E^z) 以及事件的主题概率分布 $\{\theta_v\}_{v \in V}$
输出：	主题 Z 下的事件间影响矩阵 $G_{N \times N}$
①	构建每个事件节点的特征函数 $g(v_i, y_i, z)$
②	计算主题 Z 下事件之间的相似度，构建主题 Z 下事件因子图
③	根据公式 $b_{ij}^z = \log \frac{g(v_i, y_i, z) \Big _{y_i^z=j}}{\sum_{k \in \text{NB}(i) \cup \{i\}} g(v_i, y_i, z) \Big _{y_i^z=k}}$ 将事件因子图中每个事件节点特征函数归一化成 b_{ij}^z

④对于主题 Z 下的事件因子图中每一条边, 利用公式 $r_{ij}^z = b_{ij}^z - \max_{k \in NB(j)} \{b_{ik}^z + \alpha_{ik}^z\}$ 更新 r_{ij}^z

⑤对于主题 Z 下的事件因子图中每一个节点利用公式更新 α_{jj}^z

⑥对于主题 Z 下的事件因子图中每一条边利用公式更新 α_{ij}^z

⑦迭代步骤③-⑥直到结果收敛

4. 实验与分析

4.1 实验数据获取与预处理

由于目前还未曾见到汉语-越南语事件影响力分析的相关语料。因此本文通过人工方式构建了一定量的语料。该数据集中包含了 600 个显式相关新闻事件对和 600 个无关的新闻事件对。每个新闻对包含两篇新闻 A 和 B, 分别报道了两个不同的事件。本文就在这 1200 个新闻对进行了实验。其中中文-中文、越南越-越南语、中文-越南语的相关和无关新闻对各 200 个。本文采用识别结果的准确率 P, 召回率 R, 以及 F 值作为评价指标。F 值越高效果越好。它们的计算方式如下:

$$P = \frac{\text{识别结果中正确的相关事件个数}}{\text{识别结果中相关事件个数}} \times 100\%$$

$$R = \frac{\text{识别结果中正确的相关事件个数}}{\text{测试数据中相关事件个数}} \times 100\%$$

$$F = \frac{2PR}{P + R}$$

4.2 实验环境

本文的实验环境为: Intel G620 2.6GHz 的 CPU, 4G 的内存, 320G 的硬盘, WindowsXP 的操作系统。开发工具为 My Eclipse 10。

4.3 实验 1 节点相似度阈值对实验结果的影响

本文在构建因子图模型时, 如果在主题 z 下, 任意两个事件节点的相似度大于给定的阈值 ε , 则认为在主题 z 下节点 i, j 相互关联。由于 ε 是判断事件节点是否存在边关联起来的阈值。 ε 设置过高可能会过拟合, ε 过低可能引入大量噪声。因此本试验考察不同的 ε 是否会对实验结果造成影响, 如果有影响则选取最佳 ε 值。分别在 $\varepsilon = 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8$ 时, 在测试数据上进行了实验。如表3所示为 ε 取不同值的结果对比。

表 3 不同判定节点相似度阈值对实验结果的影响

ε	查全率 (Recall)	查准率 (Precision)	F
0.2	0.8958	0.4988	0.6407
0.3	0.8452	0.6486	0.4341
0.4	0.8023	0.7787	0.4925
0.5	0.7754	0.8687	0.8194
0.6	0.7256	0.8526	0.7840
0.7	0.6651	0.7986	0.7258

0.8	0.6264	0.8058	0.7049
-----	--------	--------	--------

实验结果说明当 α 越低，由于事件间相似性较低导引入大量噪声导致大量不相关事件被识别为相关，影响结果准确性。当 α 越高，可能过拟合，导致原本相关的事件被判定为不相关。因此 α 过高过低都会影响实验结果，所以根据实验数据，本文所有的实验选择 $\alpha = 0.5$ 。

4. 4 实验 2 跨语言相关事件识别方法实验

本文所采用的方法是基于影响力传播模型的。因此本文在构建的测试数据集上，分别针对单语和双语新闻对，进行了相关事件识别方法的实验。单语环境下我们选取 LDA 主题抽取模型。为了达到最优的实验结果，本文经过多次实验对识别过程中参数的取值进行了调整，获得了一组的最优的参数取值，如表 4 所示。

表 4 单语和双语环境下识别方法参数选择

	α	主题个数
单语	0.62	11
双语	0.50	8

通过该参数得到在单语和跨语言环境下的识别效果。如表 5 所示

表 5 单语和双语环境下相关事件识别方法实验结果

	查全率 (Recall)	查准率 (Precision)	F
单语	0.8110	0.7725	0.7913
双语	0.5075	0.6408	0.6727

实验结果表明，本文所提出的相关事件识别方法，在单语言环境下和多语言环境下都取得了不错的识别效果。说明两种方法都能比较准确的识别出测试数据中相关事件。此外通过两种方法共同使用时的实验结果可以发现，两种方法具有一定的互补性，如果将两种方法同时使用，能够有效提高识别召回率。

4. 5 实验 3 不同事件相关度计算方法对比

本文提出的方法利用新闻事件的文本相似性和新闻事件影响力传播来判断双语事件是否相关。为了验证本文做法的有效性。本文还利用收集到的汉语和越南语新闻，对不同事件主题词之间的 PMI (point wise mutual information) 值进行统计，利用 PMI 值来度量事件之间的相关度。PMI 如公式 (15)

$$PMI(t_i, t_j) = \log \frac{n(t_i, t_j)}{n(t_i) \times n(t_j)} \quad (15)$$

其中 $n(t_i, t_j)$ 表示词 t_i 和词 t_j 共同在新闻中出现的次数， $n(t_i)$ 和 $n(t_j)$ 则分别表示词 t_i 和词 t_j 单独在新闻中出现的次数。

本文对这两种方法都进行了实现，并在测试数据上进行了对比实验。如表 6 所示为两种方法的结果对比。其中 A 为本文所采用的方法，B 则是通过计算新闻主题词 PMI 值进行相关事件识别的。

表 6 不同事件相关度计算方法对比

	查全率 (Recall)	查准率 (Precision)	F
A	0.7754	0.8687	0.8194
B	0.5075	0.6408	0.6727

实验结果表明，A 的识别效果较 B 有较大幅度的提升。说明在考虑新闻事件的影响力传播，对于相关事件识别

是十分必要的。通过分析实验数据发现，B 的效果不够理想是因为忽略了两个新闻事件是否相关，跟与这两个新闻事件关联的其他新闻事件相关，会忽略很多相关信息从而影响了识别效果。

5. 总结

为了全面准确的识别出汉语-越南语新闻相关事件，本文针对新闻相关事件识别方法展开研究，将对事件进行报道的所有不同语言新闻作为判断事件之间相关性的依据。本文构建双语主题模型抽取从双语新闻文档中抽取双语主题，将双语文档通过主题关联起来。针对新闻事件影响力传播特性，构建因子图模型，通过因子图模型上事件影响力传播计算不同主题下双语事件关联强度。实验结果表明，本文所采用的识别方法，能够有效识别汉语越南语双语相关事件。但是本文没有考虑汉语-越南语双语新闻事件之间要素的关联对双语事件关联的影响。我们将在今后的工作中将事件要素融入到双语事件关联分析的方法中。

参考文献

- [1]. 马彬, 洪宇, 杨雪蓉, 等. 基于语义依存线索的事件关系识别方法研究[J]. 北京大学学报(自然科学版), 2013, 49(1):109-116.
- [2]. 付剑锋, 刘宗田, 刘炜, 等. 基于层叠条件随机场的事件因果关系抽取[J]. 模式识别与人工智能, 2011, 24(4):567-573.
- [3]. Bethard S, Martin J H. Learning semantic links from a corpus of parallel temporal and causal relations[C]// ACL 2008, Proceedings of the Meeting of the Association for Computational Linguistics, June 15-20, 2008, Columbus, Ohio, Usa, Short Papers. 2008:177-180.
- [4]. 杨竣辉, 刘宗田, 刘炜, 等. 基于语义事件因果关系识别[J]. 小型微型计算机系统, 2016, 37(3):433-437.
- [5]. Do Q X, Chan Y S, Roth D. Minimally supervised event causality identification[C]// Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, Uk, A Meeting of Sigdat, A Special Interest Group of the ACL. DBLP, 2011:294-303.
- [6]. Paramita Mirza, Sara Tonelli. An Analysis of Causality between Events and its Relation to Temporal Information[C]//Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 2097-2106, Dublin, Ireland, August 23-29 2014
- [7]. Mulkar-Mehta R, Welty C A, Hobbs J R, et al. Using Part-Of Relations for Discovering Causality.[C]// Twenty-Fourth International Florida Artificial Intelligence Research Society Conference, May 18-20, 2011, Palm Beach, Florida, Usa. 2011.
- [8]. 仲兆满, 刘宗田. 利用事件影响关系识别文本集合中重要事件的方法[J]. 模式识别与人工智能, 2010, 23(3):307-313.
- [9]. 孙涛. 面向市场情报分析的 Web 实体事件融合问题研究[D]. 山东大学, 2014.
- [10]. Abe S, Inui K, Matsumoto Y. Two-Phased Event Relation Acquisition: Coupling the Relation-Oriented and Argument-Oriented Approaches.[C]// COLING 2008, International Conference on Computational Linguistics, Proceedings of the Conference, 18-22 August 2008, Manchester, Uk. 2008:1-8.
- [11]. Chambers N, Dan J. Unsupervised Learning of Narrative Schemas and their Participants.[C]// ACL 2009, Proceedings of the Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing of the Afnlp, 2-7 August 2009, Singapore. 2009:602-610.
- [12]. Jeh G, Widom J. Scaling personalized web search[C]// International Conference on World Wide Web. ACM, 2003:271-279.
- [13]. Jeh G, Widom J. SimRank: a measure of structural-context similarity[C]// Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2002:538-543.
- [14]. Tang J, Sun J, Wang C, et al. Social influence analysis in large-scale networks[C]// ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2009:807-816.
- [15]. Zhang D, Mei Q, Zhai C X. Cross-Lingual Latent Topic Extraction[C]// ACL 2010, Proceedings of the Meeting of the Association for Computational Linguistics, July 11-16, 2010, Uppsala, Sweden. 2010:1128-1137.



唐莫鸣

(1993—)，昆明理工大学研究生，研究方向：自然语言处理和信息检索。

Email: 553731700@qq.com

电话: 15559676947

邮编: 650500

地址: 云南省昆明市呈贡县昆明理工大学呈贡校区



朱明玮

(1992—)，昆明理工大学研究生，研究方向：自然语言处理和信息检索。

Email: 1518164914@qq.com

电话: 18288210735

邮编: 650500

地址: 云南省昆明市呈贡县昆明理工大学呈贡校区



余正涛

(1975—)，博士，现任昆明理工大学信息工程与自动化学院教授，博士生导师，昆明理工大学智能信息处理重点实验室主任。主要研究方向包括自然语言处理，机器翻译和信息检索。

Email: ztyu@hotmail.com

电话: 13888616568

邮编: 650500

地址: 云南省昆明市呈贡县昆明理工大学呈贡校区