ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 32,No. 1,Jan. 2017,pp. 141－148
DOI:10.16337/j.1004-9037.2017.01.017

http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:＋86-025-84892742

基于改进 DE-Tri-Training 算法的汉语多词表达抽取

梁颖红[1]　谭红叶[2]　鲜学丰[3]　黄丹丹[1]　钱海忠[1]　沈春泽[1]

(1.金陵科技学院软件工程学院,南京,211169; 2.山西大学计算机与信息技术学院,太原,030006; 3.苏州市职业大学计算机工程学院,苏州,215104)

摘　要：多词表达的识别错误会对很多自然语言处理任务造成不利影响。DE-Tri-Training 半指导聚类算法在聚类初期使用有指导的标注信息,取得了较好的抽取结果。本文采用基于中心词扩展的初始聚类中心确定方法和基于有指导信息的一致性协同学习数据净化方法,提出了半指导策略抽取汉语多词表达,聚类算法的中后期也加入有指导的信息,使分类器能使用正确的标注信息进行训练。通过与 DE-Tri-Training 算法的对比实验,改进的 DE-Tri-Training 算法得到的汉语多词表达抽取结果优于原来的算法,验证了改进 DE-Tri-Training 算法的有效性。

关键词：多词表达;半指导;协同训练

中图分类号：TP391　　**文献标志码**:A

Chinese Multi-word Expression Extraction Based Improved DE-Tri-Training Algorithm

Liang Yinghong[1], Tan Hongye[2], Xian Xuefeng[3], Huang Dandan[1], Qian Haizhong[1], Shen Chunze[1]

(1. Software Engineering Department, Jingling Institute of Technology, Nanjing, 211169, China; 2. School of Computer and Information Technology, Shanxi University, Taiyuan, 030006, China; 3. Computer Engineering Department, Suzhou Vocational University, Suzhou, 215104, China)

Abstract: Failing to identify multiword expression (MWE) may cause serious problems for many natural language processing (NLP) tasks. Because of lacking of Chinese MWE tagging corpus, a semi supervised method is used to extract Chinese MWE. DE-Tri-Training semi-supervised clustering algorithm uses supervised information in the beginning of the cluster, and obtains good results. The selection method of original cluster center based head word expansion and the consistency collaborative learning data depuration method based supervised information are proposed, which adds the supervised information into the mid and late steps of clustering, so that classifiers can use correct label information to train it. The contrast experiment show that the extraction results of Chinese multi-word expression using the improved DE-Tri-Training algorithm are better than that of using unimproved one. The effectiveness of the improved DE-Tri-Training algorithm is thus verified.

Key words: multi-word expression; semi-supervised; tri-training

---

基金项目:国家自然科学基金(61100138,61402134,11601202)资助项目;江苏省“333”工程高层次人才培养(BRA2015108)资助项目;金陵科技学院高层次人才工作启动费(40620022)资助项目;江苏省高校自然科学研究面上(16KJB520013,14KJB520013)资助项目;山西省自然科学基金(2011011016-2)资助项目;山西省回国留学人员科研(2013-022)资助项目;山西省 2012 年度留学回国人员科技活动择优资助项目。
收稿日期:2015-06-05;**修订日期**:2015-06-30

引 言

自然语言处理领域中，多词表达(Multi-word expression，MWE)的准确抽取和翻译会影响机器翻译、信息检索和词义消歧等研究的性能提升。Sag 等对多词表达的定义[1]是：把两个或多个词组合在一起形成的具有单一意义的单元叫做多词表达。多词表达被认为是自然语言处理领域的难点和性能提升的瓶颈问题。为了规避汉语多词表达语料的构建，研究者大多使用英汉双语平行语料来进行研究[2-6]，有少数的研究者先标注小规模语料再进行多词表达抽取[7]。有指导学习方法和无指导学习方法是常用的语料建设方法。无指导学习方法不用手工标注语料，但学习性能不太理想。半指导学习方法能在人工标注和系统性能之间取得折中的效果，因此得到越来越多研究人员的青睐。半指导聚类方法使用较少的指导性信息，一般分为基于约束和基于距离两大方法[8]。其中，基于约束的方法使用约束条件使得聚类的数据能够更加快速地聚到合适的类中，受到研究者的关注[9]。Wagstaff 等使用增强的限制条件来分派聚类的数据[10]。文献[11]也使用少量的标注数据来优化初始的聚类中心。基于聚类的方法对初始聚类中心的确定非常重要，同时初始种子的选择会对结果造成很大影响。基于分类的方法先学习标注语料，再对未标注语料进行标注，然后再把标注的语料放到已标注的语料中，如此反复。如果放回到标注语料中不准确的结果，会使错误进一步扩散到下一步的标注中。聚类和分类方法各有优缺点，把两者结合，进行优势互补是研究者们努力的方向。本文提出基于半指导的聚类算法进行多词表达抽取，解决人工标注语料的费时费力问题。在半指导聚类算法中，DE-Tri-training 算法是比较受关注的算法。从 2003 年开始，自然语言处理领域的计算机语言联合会(Association for computational linguistics，ACL)年会为多词表达设立 workshop 主题以供全世界研究者交流。目前大多研究者把多词表达抽取看成分类问题[2-7]，从预先标注好的语料中进行学习。近两年，有少数研究者为了解决语料短缺问题使用了聚类方法进行多词表达抽取[12]。因为缺少大规模的专用汉语多词表达标注语料，以往从事汉语多词表达抽取的研究者往往运用已有的汉语标注语料，采用统计上下文词搭配信息、互信息和对数似然函数来衡量汉语词之间的结合强度，把结合紧密的词当作汉语多词表达[5]；另外一些研究者采用英汉双语对齐语料，采用统计方法和错误驱动规则来筛选候选多词表达[11]；还有少数研究者自己构建特定领域的语料资源，如 Wang 等[7]构建了汉语习语库，为多词表达中习语类型抽取提供了资源。对于缺少专门多词表达标注语料的汉语多词表达抽取，聚类方法无疑是比较好的选择。然而，为了提高聚类的准确率，寻求比较好的半指导聚类算法是研究者努力的目标。

文献[8]提出 DE-Tri-training 半指导 K 近邻聚类算法，是有指导和无指导方法有机结合的典型代表，而且也取得了较好的结果。DE-Tri-training 的半指导 K 近邻聚类算法的基本思想是：首先利用事先人工标注的小部分语料，采用 Tri-training 进行学习并对未标注语料进行标注，再把标注好的结果添加到已标注语料中，这样就扩大了标注语料的规模，然后把扩大规模后的标注语料作为下一步聚类的种子，而且为了尽可能把正确的标注结果放回到原来的标注语料中，对经过 Tri-training 标注的结果采用了数据编辑技术，以去除不正确的标注结果。周志华等[13-16]也证明了该算法的有效性。但是，该方法存在两个缺陷：(1)它在聚类过程中还是采用随机确定初始中心的方法，这对聚类结果会产生不利的影响；(2)采用数据编辑技术去除不正确的标注结果时，需选择聚类结果中的 3 个最近邻，至少两个最近邻和此结果一致，才认为该结果为真，再放回已标注语料中。因为聚类方法缺少有指导的标注语料信息，而以上方法采用聚类结果来确认是否正确，这种去除不正确结果的方法缺乏可信度。

1 改进的 DE-Tri-Training 算法

在语料构建过程中，如果把不正确的聚类结果放回到标注语料中，在后续的循环过程中会把这个错误放大，致使聚类的结果更加不准确，因此对放回标注语料库的结果进行数据净化非常关键。本文对

DE-Tri-training 的半指导 K 近邻聚类算法进行改进:(1)采用中心词驱动的方法来确定聚类的中心,以提高聚类的准确率;(2)采用一致性协同学习原则来去除不正确的标注结果。利用 Tri-training 的 3 个分类器对拟放回已标注语料的聚类结果分别用 3 个分类器来识别,如果至少两个分类器的结果与原来聚类的结果一致,才认为该聚类结果正确,再放入到已标注的语料中。改进的一致性协同学习方法与原来的数据编辑技术区别主要在于:分类器由正确的标注结果训练得到,而原来的数据编辑技术使用无指导的聚类结果,因此改进的一致性协同学习方法在理论上更具优越性。

1.1 基于中心词扩展的初始聚类中心确定方法

改进 K-均值聚类算法初始时不随机确定每个分类的中心,而是采用少量特征数据(依据从手工标注的语料库中统计而来)来确定每个类的初始中心。有指导的统计方法和无指导的聚类方法有机结合,弥补了原来随机确定初始中心的缺陷。改进的 K-均值聚类算法步骤如下:(1)从少量标注语料库中统计数据信息,采用有指导的策略把句子中的单词先分到不同的类中。(2)运用聚类算法调整中心,进行聚类。(3)最后根据单词在句子中的位置确定多词表达的边界。

(1) 基于中心词扩展的方法

短语被看成是单词词性按一定规则构成的聚簇[17],在一个短语中,各单词词性间存在相互依赖关系;同时每个词性与它周围邻近词性的关系也可能比较紧密,即如果一个词性出现在某种短语中,则这个词性周围出现的词性在这类短语里同时出现的几率也很大,这种可能性就用关联度进行表示。对于一种短语,有一种或几种词性的出现频率高于其他词性。因此在一个句子中,如果某个词的词性与某种短语中经常出现的词性相吻合,则可以假定这个词性为中心词。本文为每个短语选择两个中心词。有了中心词后,即可以利用词语间的关联度计算在中心词被设定的这类短语中,中心词与其周围的词性之间的关联度,如果中心词与其邻近的词性之间的关联度大于阈值,则将这一邻近的词也划入这一类型的短语,使短语的边界扩大;然后再计算新划入词的词性与它邻近的词性的关联度,如果这两个词性之间的关联度仍大于阈值,则继续将新词划入这一类型的短语,并继续计算新的短语边界上的词性和边界外邻近的词性之间的关联度,以此类推,直到边界上的词性和边界外相邻的词性之间的关联度小于阈值,则边界不再扩大。

(2) K-均值聚类算法

根据统计数据得到的中心词难免会出现偏差,因此运用聚类算法调整中心,并计算多词表达内的每个词与每个中心的距离,如某词和其他类中心的距离小于该词与当前类中心的距离,则把该词移动到距离中心最小的类中。改进 K-均值聚类算法的详细情况如下。使用的属性:当前词的"词"和"词性"以及前一个词的"词性"信息作为属性;中心的确定方法:当某类中的其他词距离某个词的距离几乎相等时,把这个词作为中心。距离函数为

$$d(x_i,x_j)=\sqrt{\sum_{r=1}^{n}(a_r(x_i)-a_r(x_j))^2} \tag{1}$$

式中:x_i 为某个词;x_j 为另一个词;$d(x_i,x_j)$为这两词之间的距离;$a_r(x_i)$为 x_i 的第 r 个属性值;n 为属性的个数。改进后的 K-均值聚类算法流程图见图 1。

1.2 基于有指导信息的一致性协同学习数据净化

在原来的 DE-Tri-training 算法中,对新标注的数据在聚类中寻找它的 3 个最近邻,如果这 3 个近邻中至少两个和它本身的标注结果一致,就认为该标注结果是对的,则放入到已标注语料中。原算法中,以上过程在聚类过程中进行,缺少有指导标注信息的借鉴。改进的 DE-Tri-training 算法将充分利用已标注信息,使用事先确定的 3 个分类器,把经过聚类的标注结果分别放入 3 个分类器中,如果至少两个的标注结果与原来一致,才认为该结果为真,再放入到已标注语料中。本文的改进 DE-Tri-training 半指

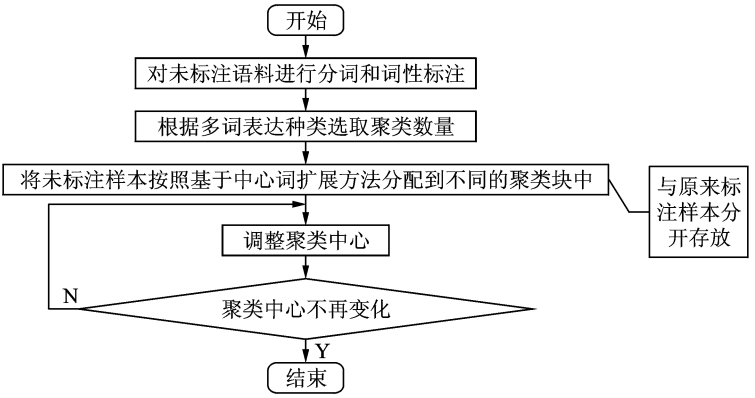


图 1 基于中心词扩展方法的 K-均值聚类算法流程图

Fig. 1 Flow chart of K-means clustering algorithm based on extended method of center word

导聚类算法的流程见图 2,加底纹部分是改进内容。改进 DE-Tri-training 半指导聚类算法形式化描述如下(原算法见文献[10])。其中(1c),(2),(3),(4)为改进部分。

算法 1 改进的 DE-Tri-training 半指导聚类算法

输入:数据集 $X=\{x_i\}_{i=1}^{n}$,$x_i\in\mathbf{R}^d$,聚类的个数为 K,初始标注的种子集 $S=U_{K=1}^{K}S_k$,$S_k\neq\varnothing$,3 个没有训练的分类器为 H_1,H_2 和 H_3。

输出:对 X 的 k 个划分 $\{X_h\}_{h=1}^{k}$ 和局部优化函数。

步骤如下:

(1) 执行 DE-Tri-training 过程来扩大和编辑初始种子集 S:

(1a) $L\leftarrow S$,$U\leftarrow X\text{-}L$;通过 Bootstrap 从 L 中产生训练集 S'_1,S'_2 和 S'_3,并分别训练分类器 H_1,H_2 和 H_3;

(1b) for each $H_i(i=1,2,3)$:

让 $H_j\&H_k(j,k\neq i)$从 U 中选择并标注数据子集 $L_i=\{x\mid x\in U,H_j(x)=H_k(x)\}$,并组成新的数据集 $S'_i=L\cup L_i$;

(1c) 对每一个 S'_i 中新标注的数据子集 L_i 执行数据净化(借鉴有指导信息的一致性协同学习方法):

$S'\leftarrow S'_i$;对 L_i 中的每个 x,分别放入分类器 $H_i(i=1,2,3)$中,如果至少两个分类器的标注结果为 c,则把 S'中的 x 标为 c;否则,从 S'中移除 x;$S'\leftarrow S'_i$;

(1d) 对每一个 $H_i(i=1,2,3)$:如果$|S'_i|>|S|$,用 S'_i 重新训练 H_i;

(1e) 如果 $H_i(i=1,2,3)$中的任何一个发生改变,转(1b);

(1f) $S\leftarrow S'_1\cup S'_2\cup S'_3$;对 $S\sim L$ 中的每个数据使用分类器 $H_i(i=1,2,3)$通过 Weighted voting 准则进行重新标注。

(2) 采用基于中心词扩展的策略初始化聚类中心。根据种子集中标注信息把种子集分为 k 个类,$S=\bigcup_{h=1}^{k}S_h$,$S_h\neq\varnothing$,将未标注样本按照基于中心词扩展方法分配到不同的聚类块中与原来标注样本分开存放。

(3) 重新计算聚类的中心:$d(x_i,x_j)=\sqrt{\sum_{r=1}^{n}(a_r(x_i)-a_r(x_j))^2}$,$x_i$ 为某个词,x_j 为另一个词,$d(x_i,x_j)$为这两词之间的距离,$a_r(x_i)$为 x_i 的第 r 个属性值,n 为属性的个数。

(4) 如果 k 个聚类中心不再变化,结束;否则,转(2)。

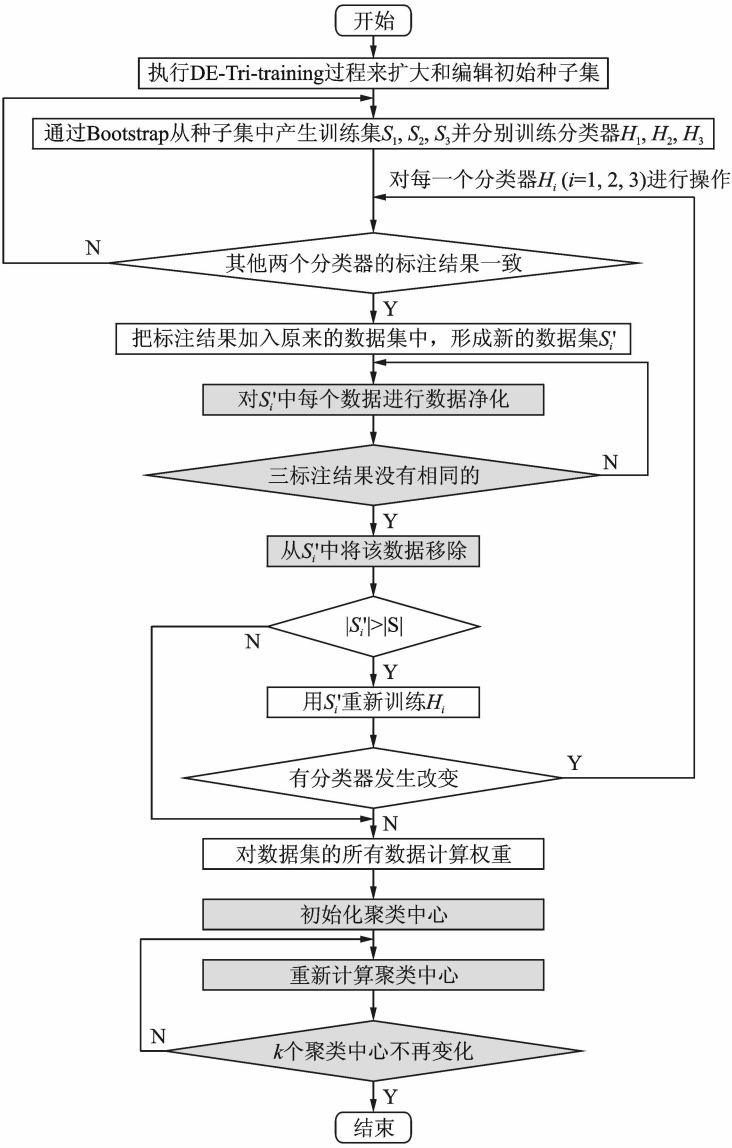


图 2 改进的 DE-Tri-training 半指导聚类算法流程图

Fig. 2 Flow chart of improved DE-Tri-training semi-supervised clustering algorithm

2 实验数据及方法

因为没有公开的汉语多词表达语料,先人工标注小部分汉语多词表达语料(8 000 句)作为聚类的种子,人工标注的多词表达语料的领域包含科技(3 000 句)、新闻(3 000 句)和医学(2 000 句)3 个领域,对汉语的复合名词、动词结构和习语进行了人工标注。把标注好的语料作为种子,又从互联网上下载了 1.5 GB 的汉语多领域语料作为聚类使用的生语料。实验设计如下。

(1)采用原来的 DE-Tri-training 算法和本文改进的 DE-Tri-training 算法,对复合名词、动词结构和习语进行抽取。把标注语料分为 10 份,其中 10%作为测试集,其余 90%作为训练集,采用 10 重交叉验证,取平均值作为最后结果。聚类算法使用了整个数据集。3 个分类器分别采用支持向量机(Support

vector machine, SVM)、K 近邻(K-nearest neighbor, KNN)和条件随机场(Conditional random field, CRF),以上 3 个分类器代码改写自 Python 源代码。聚类的个数 $K=4$($K=1$ 为复合名词;$K=2$ 为动词结构;$K=3$ 为习语;$K=4$ 为不属于任何类)。为了防止聚类不收敛,把聚类的最大循环次数设为 200。为了选取合适的聚类错误率阈值,分别对不同阈值进行比较,结果见表 1。

(2)在本文的数据集上实现 K-均值聚类算法,代码改写自 Python 源代码,对复合名词、动词结构和习语的多词表达进行抽取,其结果与改进的 DE-Tri-training 算法进行比较。因为多词表达大多采用分类方法进行抽取,而且使用语料与本文不同,因此不具备可比性。为了与本文方法比较,选取了典型的 K-均值聚类算法在本文采用的数据集上抽取相同类型的多词表达。K-均值聚类中的 $K=4$($K=1$ 为复合名词;$K=2$ 为动词结构;$K=3$ 为习语;$K=4$ 为不属于任何类),采用 log-likelihood 函数计算距离,经过 87 次迭代之后收敛。从表 1 和图 3 可以看出,当错误率阈值在 0.000 05 时抽取结果最好,之后阈值再增大结果反而下降,因此把 0.000 05 作为最后的错误率阈值。

表 1 采用不同聚类错误率阈值的结果比较

Tab. 1 Comparison results based on different clustering error rate threshold

阈值	多词表达类别	精确率/%	召回率/%	$F_{\beta=1}$/%
0.000 01	复合名词	80.64	83.58	82.08
	动词短语	74.57	79.67	77.04
	习语	69.12	74.36	71.64
0.000 05	复合名词	81.32	87.86	84.46
	动词短语	77.57	84.65	80.96
	习语	70.36	78.67	74.28
0.000 08	复合名词	79.18	82.65	80.88
	动词短语	74.32	80.24	77.17
	习语	68.36	73.28	70.73

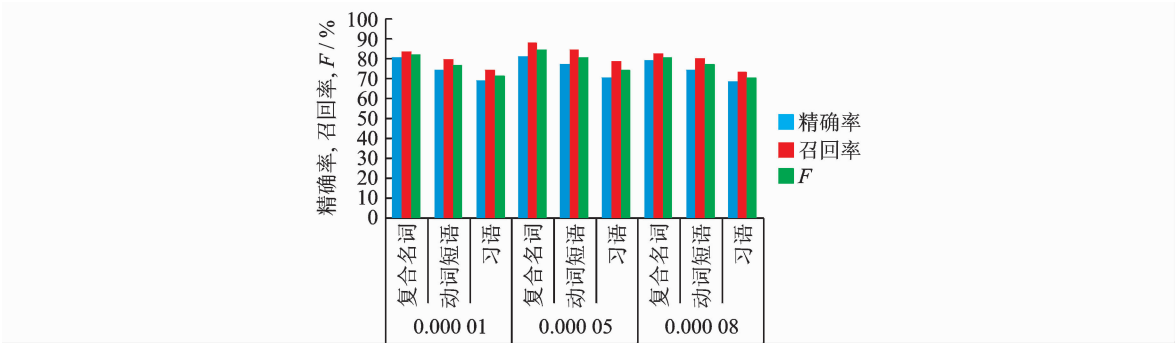


图 3 采用不同聚类错误率阈值的结果比较

Fig. 3 Comparison results based on different clustering error rate threshold

3 实验结果及分析

为了验证本文改进 DE-Tri-training 方法的有效性,在相同的语料上分别采用原来未改进的 DE-Tri-training 算法以及 K-均值聚类算法对汉语的 3 种多词表达(复合名词、动词短语和习语)进行了抽取,3 种方法的结果见表 2,比较结果见图 4。

表 2 本文方法与 Baseline 和 K-均值聚类算法的结果比较

Tab. 2 Comparison results among the improved method, Baseline and K-means

方法	多词表达类别	精确率/%	召回率/%	$F_{\beta=1}$/%
DE-Tri-training 算法(Baseline)	复合名词	79.82	82.47	81.12
	动词短语	75.45	78.84	77.10
	习语	68.63	72.15	70.35
K-均值聚类算法	复合名词	75.83	80.26	77.98
	动词短语	72.15	73.67	72.90
	习语	53.64	68.72	60.25
改进的 DE-Tri-training 算法	复合名词	81.32	87.86	84.46
	动词短语	77.57	84.65	80.96
	习语	70.36	78.67	74.28

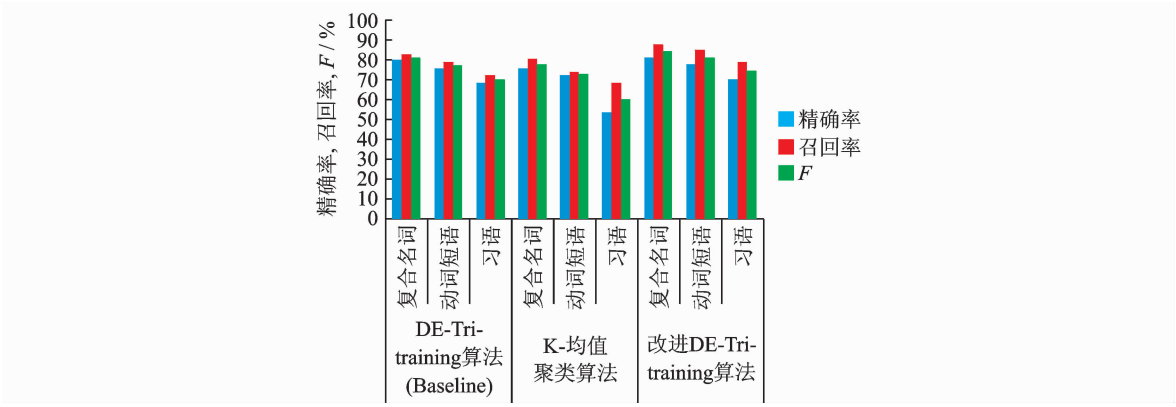


图 4 本文改进方法与其他方法的结果比较

Fig. 4 Comparison results based on improved method and other methods

从表 2 和图 4 可以看出,改进后的 DE-Tri-training 算法对复合名词、动词短语和习语的抽取结果均好于原来的 DE-Tri-training 算法和 K-均值聚类算法,DE-Tri-training 算法的结果优于 K-均值聚类算法,说明本文采用的基于中心词扩展的初始聚类中心确定方法和基于有指导信息的一致性协同学习数据净化方法是有效的。从复合名词、动词短语和习语各自的抽取结果分析,复合名词的结果要好于动词短语,动词短语要好于习语。实验表明,习语是一些约定俗成的短语,单单依赖上下文的信息进行抽取不能达到理想的效果。

4 结束语

本文对有指导和无指导方法相结合的 DE-Tri-training 半指导聚类算法进行了分析,指出了它在聚类过程中的两个缺陷。针对此缺陷,采用基于中心词扩展的初始聚类中心确定方法和基于有指导信息的一致性协同学习数据净化方法,加强了有指导信息对聚类的影响。实验表明,改进后的 DE-Tri-training 方法在抽取汉语复合名词、动词短语和习语过程中优于原来的方法和 K-均值聚类算法。在以后的研究中,考虑加入习语词典、动词词典等知识提高习语和动词短语抽取的准确率。本文方法对汉语多词表达的抽取和语料的构建提供了一个新的方法和思路。

参考文献：

[1] Sag I A, Baldwin T, Bond F, et al. Multiword expressions: A pain in the neck for NLR[J]. Computational Linguistics and Intelligent Text Processing, Lecture Notes in Computer Science, 2002,2276: 189-206.

[2] Duan Jianyong, Lu ruanzhan, Wu Weilin, et al. A bio-inspired approach for multiword expression extraction[C]// Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics. Sydney, Australia: BPA Digital, 2006:4876-4883.

[3] Constant M, Sigogne A. MWU-aware part-of-speech tagging with a CRF model and lexical resources [C]// Workshop at ACL 2011,From Parsing and Generation to the Real World. Portland, Oregon: USA Production and Manufacturing,2011: 49-56.

[4] Vincze V, István Nagy T, Berend G. Detecting noun compounds and light verb constructions: A contrastive study [C]// Proceeding MWE'11 Proceedings of the Workshop on Multiword Expressions: From Parsing and Generation to the Real World. Portland, Oregon: USA Production and Manufacturing, 2011: 116-121.

[5] Piao S S, Sun Guangfan, Rayson P, et al. Automatic extraction of Chinese multiword expressions with a statistical tool[C]// Proceedings of the Workshop on Multi-word expressions in a Multilingual Context. Trento:Italy: J. Weeds, 2006:17-24.

[6] Duan Jianyong, Zhang Mei, Tong Lijing, et al. A hybrid approach to improve bilingual multiword expression extraction[J]. Lecture Notes in Computer Science, 2009(5476):541-547.

[7] Wang Lei. Construction of a Chinese idiom knowledge base and its applications[C]// Proceedings of Coling 2010 Multi-word Expressions. Beijing, China: Natural Language Engineering, 2010:10-17.

[8] Deng Chao, Guo Maozu. Tri-training and data editing based semi-supervised clustering algorithm[J]. Lecture Notes in Computer Science, 2006,4293: 641-651.

[9] Bilenko M, Basu S, Mooney R J. Integrating constraints and metric learning in semi-supervised clustering[C]// 21st International Conference on Machine Learning. Banff, Canada: Schapire RE, 2004: 81-88.

[10] Wagstaff K, Cardie C, Rogers S, et al. Constrained K-means clustering with background knowledge[C]// 18th International Conference on Machine Learning (ICML-01). Williamstown, USA: Morgan Kaufmann Publishers Inc,2001: 577-584.

[11] Tomás D, Giuliano C. Exploiting unlabeled data for question classification[J]. Lecture Notes in Computer Science, Natural Language Processing and Information Systems, 2011, 6716: 137-144.

[12] Tutubalina E. Clustering-based approach to multiword expression extractionand ranking[C]// Workshop at ACL 2015. Beijing, China: ACL, 2015:39-43.

[13] Zhou Z H, Li M. Tri-training: Exploiting unlabeled data using three classifiers[J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(11): 1529-1541.

[14] Li M, Zhou Z H. Improve computer-aided diagnosis with machine learning techniques using undiagnosed samples[J]. IEEE Transactions on Systems, Man and Cybernetics-Part A: Systems and Humans, 2007, 37(6): 1088-1098.

[15] Zhang M L, Zhou Z H. CoTrade: Confident co-training with data editing[J]. IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics, 2011, 41(6): 1612-1626.

[16] Zhou Fa,Zhang Wei,Sun Ke,et al. Optimized fuzzy clustering method for health monitoring of shield tunnels[J]. Transactions of Nanjing University of Aeronautics and Astronautics, 2015, 32(3):325-334.

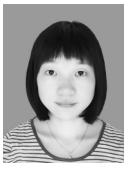
[17] 梁颖红，赵铁军，刘博，等. 基于关联度评价的中心词扩展的英文文本语块识别[J]. 计算机研究与发展，2006，43(1)：153-158.
Liang Yinghong, Zhao Tiejun, Liu bo, et al. English text chunk recognition based on relevance degree evaluation and head word extension strategy[J]. Computer Research and Development,2006, 43(1):153-158.

作者简介：

梁颖红（1970-），女，教授，研究方向：自然语言处理、网络信息挖掘，E-mail：liangyh7036@126.com。

谭红叶（1972-），女，副教授，研究方向：中文信息处理和信息抽取。

鲜学丰（1980-），男，副教授，研究方向：智能信息处理和 Deep Web 信息挖掘。



黄丹丹（1987-），女，讲师，研究方向：信息安全与密码学。

钱海忠（1977-），男，副教授，研究方向：数据语义处理。

沈春泽（1976-），男，讲师，研究方向：自然语言处理。