文章编号：1003-0077（2011）00-0000-00

基于 Word Embedding 语义相似度的字母缩略术语消歧*

于东[1,2]，荀恩东[1,2]

（1. 北京语言大学汉语国际教育技术研发中心，北京市 10083；

2. 北京语言大学信息科学学院，北京市 10083）

摘要：提出基于 Word Embedding 的歧义词多个义项语义表示方法，实现基于知识库的无监督字母缩略术语消歧。方法分两步聚类，首先采用显著相似聚类获得高置信度类簇，构造带有语义标签的文档集作为训练数据。利用该数据训练多份 Word Embedding 模型，以余弦相似度均值表示两个词之间的语义关系。在第二步聚类时，提出使用特征词扩展和语义线性加权来提高歧义分辨能力，提高消歧性能。该方法根据语义相似度扩展待消歧文档的特征词集合，挖掘聚类文档中缺失的语义信息，并使用语义相似度对特征词权重进行线性加权。针对 25 个多义缩略术语的消歧实验显示，特征词扩展使系统 F 值提高约 4%，使用语义线性加权后 F 值再提高约 2%，达到 89.40%。

关键词：字母缩略术语；术语消歧；Word Embedding；语义相似度

中图分类号：TP391　　**文献标识码：**A

Acronym Term Disambiguation Based on Semantic Similarity Calculated by Word Embedding

Dong YU[1,2], Endong XUN[1,2]

(1. Inter. R&D Center for Chinese Education, Beijing Language and Cultural University, Beijing 10083, China; 2. College of Information Science, Beijing Language and Cultural University, Beijing 10083, China)

Abstract: This paper introduces a knowledge based unsupervised method for acronym term disambiguation. Word embedding is used for acronym term semantic representation. In the first stage of disambiguation, significantly similar documents are clustered and used as training data. Each cluster corresponds to an interpretation of an acronym term, so it can be seen as a semantic tag. On the basis, word embedding is trained for several times and semantic relation between two words can be calculated by average of cosine similarity of their vectors. In the seconded stage, the paper proposes to use feature word expansion and linear weighted semantic similarity to improve system performance. By calculating semantic similarities between documents and interpretations, implicit semantics can be mined as new feature words; the weight of a feature words is linear weighted by their semantic similarities with specific interpretation. Experimental results on 25 acronym terms show that, feature word expansion improves system F score by 4% and semantic weight gains higher performance by 2%. The final system F score is 89.40%.

Key words: acronym term; term disambiguation; word embedding; semantic similarity

1 引言

随着科技进步，各领域专业术语数量快速增长。中文文献中，许多源于国外文献的专业术语直接以字母缩略词形式使用，如 “IBM”、“NBA”等。字母缩略术语表示空间有限，一词多义现象非常普遍。如“UPS”至少包含“UPS 电源”和“UPS 物流公司”两种义项。在中国知网文献数据库中检索二者，分别得到 15541 条、8192 条结果，说明两个义项在各自领域均为常用术语。类似现象还有“防抱死制动系统(ABS)”和“ABS 树脂”。多义缩略术语专

* **收稿日期：**　　**定稿日期：**

基金项目：国家自然科学基金项目（61300081，61170162）；国家科技支撑项目（2012BAH16F00）；北京语言大学中央高校基本科研业务专项资金（14YJ030005）

业性强、更新快，随着新术语不断涌现，字母缩略术语的歧义性不断增加，不仅会增加阅读者理解难度，也会对现有的信息检索、机器翻译等应用造成许多障碍，研究字母缩略术语的消歧具有实际应用价值。

字母缩略术语的语料资源稀少，义项专业性强，因此本文选择基于知识库的无监督方法实现消歧。在这方面，传统语义消歧(WSD)方法常选取歧义词上下文语境作为特征，用向量空间模型（VSM）表示文档[11]。其实质上是根据领域特征划分歧义词所在的文档，缺乏对歧义词语义信息的挖掘和利用。使用词义网络如 WordNet、HowNet 中的语义知识辅助词义消歧能够取得较好的效果[1,12]。然而对于缩略术语而言，词义网络更新慢、覆盖度低，无法满足使用要求。

近几年，基于神经网络的 Word Embedding 方法在词语语义表示方面表现出很好的性能，受到广泛关注[2,3,4]。Word Embedding 的任务是将语料库中的每个词表示为一个低维实数向量，建立离散词汇与实数域特征向量之间的映射，能够使语义类似的词语，其向量表示也较为接近，任意两个词语的语义相关程度可以由两者向量的余弦相似度表示。利用该特点，本文在消歧过程中计算缩略术语多个义项 Word Embedding，利用义项语义特征对基本 VSM 模型进行扩展，提出针对缩略术语的消歧方法。

本文主要工作包括三个方面：(1)采用多步聚类思想，使用显著相似性聚类，从原始数据中抽取可靠知识；(2)利用第一步聚类结果进行义项反标注，进而训练每个义项的 Word Embedding，挖掘每个义项的语义信息；(3)提出特征词权重的语义线性加权方法，进行二步聚类，有效提高系统整体消歧性能。与已有工作相比，本研究能够提取并充分利用高置信数据，结合 Word Embedding 表示方法，无监督地获取歧义义项的语义表示，实现特征词领域权重和语义权重的融合，最终实现语义消歧。

2 相关研究

2.1 统计词义消歧

语义消歧解决同一词汇在不同语境下的义项识别和标注问题。1990 年后，基于统计的多义词语义消歧技术成为研究主流。Schütze[5]将语义消歧问题转化为聚类问题，成为该领域的主流方法。鲁松[13]使用向量空间模型计算相似度实现消歧；何径舟[14]使用最大熵选择特征计算聚类相似度，有效提升了中文词义消歧性能。多义词的词义消歧任务一般针对通用词汇，重点是区分词语在不同语境下所代表的语义，即语言本身的歧义性，难度较大。本文所讨论的问题则限于实体词的消歧，不涉及语言本身的歧义性。

2.2 中文实体词消歧

实体词的语义消歧是语义消歧中的一个重要分支，可分为两个子问题：(1)实体词边界划分歧义消解；(2)多义实体词概念消歧。前者主要解决语言本身歧义，后者则根据实体词上下文语境，实现实体概念的区分。该领域有代表性的研究问题是人名消歧，Mann[6]将该问题看成基于人物属性的无监督聚类问题。在中文人名消歧方面，丁海波[15]使用多阶段的消歧聚类策略，李广一[16]、Z Peng[7]均采用多步聚类方法解决该问题。此外，J Liu[8]、杨欣欣[17]利用外部知识源进行知识扩展，也有效提高了消歧性能。目前，国际 WePS 评测和国内评测 CLP2010、CLP2012 均设有人名消歧的任务。

字母缩略词语也属于实体词范畴，且具有较强的专业性，因此需要更广泛的知识以覆盖相关领域；混杂在中文中的字母缩略词提供的词汇特征很少，也与传统问题有所区别。

2.3 字母缩略词语义消歧

国外也已有学者关注字母缩略语带来的歧义问题。如 Liu[9]，Stevenson[10]在医学缩略词消歧领域的工作，更多地考虑了上下文的词汇特征，这是因为在英文文献中，缩略字母往往来源于上下文词串，而中文文档中类似信息很少，因此更需要语义信息辅助消歧。

3 语料库构建

本文利用百科网站建立多义术语知识库，利用通用搜索引擎自动获取术语在各种语境中的使用数据作为测试集，经后处理和部分人工校对后，建立具有一定规模的多义术语数据库。该数据库包括两部分：(1)由字母缩略术语、中文译文、以及多种释义文本构成的知识库；(2)包含多义术语的测试文档集，其中每个测试文档仅指向一个多义术语。知识库中的每行包含多义术语的一个释义，提供义项标签(id)、译文(def)、以及义项释义文档。测试库中每行对应一个测试文档，通过“答案标签(ans)”指示文档对应的义项。如图 1 所示。

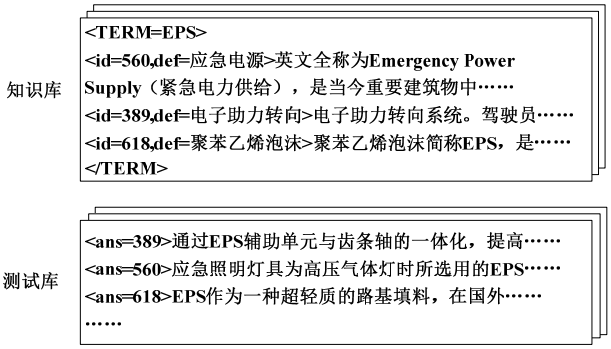

图 1 多义缩略术语知识库和测试库格式

针对消歧问题，多义术语数据库要求选用常用术语词条为对象；词条的每个义项均有明确、清晰的释义文本；词条的每个义项均有一定规模的测试文本量。数据库建设分两步：

首先建立术语知识库。根据术语词表获取百度百科中对应的多义词条页面，以及对应的各个义项页面内容，采用文献[18]中提出的描述式定义语言模式，自动抽取释义语句，经人工筛选后得到每个义项定义和释义描述文本，构成知识库。

然后根据知识库构建测试集。以术语义项为检索词，如“EPS 电子助力转向”，利用搜索引擎返回与术语最相关的文档，保留包含目标术语词、不重复且长度在一定范围内的句子作为测试文档。最后经人工校对和标注，得到带有义项标签的测试文档集。

本文最终建立包含 25 个多义缩略术语的数据库，共包含 98 个义项，2384 条测试数据。平均每个词条有约 4 个义项，“测试/义项”数量比超过 10，保证数据具有多样性、丰富性。详见表 1。

表 1 多义术语语料库数据统计

术语	定义数量	测试数量	术语	定义数量	测试数量
ABC	7	90	MBA	2	73
ABS	5	133	MIMO	3	60
AFP	4	69	NP	2	45
AP	2	54	OA	2	102
APC	3	48	PCI	4	56
BOM	3	73	PCT	3	100
CAD	3	81	PPA	5	64
CR	5	76	SAP	4	87
CVT	4	113	SAS	9	183
DMA	4	118	SIP	2	112
EPS	6	107	TT	8	136
FCC	3	83	UC	3	78
			UPS	2	243
总计				**98**	**2384**

4 研究方法

4.1 整体框架

本文研究问题可描述为：多义术语 w 有 h 个义项，每个义项一个标签(id)标记，得到的义项集合记为：$C_w = \{w\#1, w\#2, ..., w\#id, ..., w\#h\}$。在测试文档 d 中出现 w，则文档 d 与 w 的任意义项间存在关系 $R(w\#id|d)$，其中有且只有 $w\#id^*$是其正确义项。消歧任务是通过分析计算关系 $R(w\#id|d)$，寻找与 d 最接近的义项，即：

$$w\#id^* = \arg\max_{w\#id \in C_w}(R(w\#id \mid d)) \tag{1}$$

本文采用无监督方法，将多义缩略术语消歧看作两步聚类问题。聚类过程使用对特征词加权的向量空间模型，以释义文档和测试文档两者间的相似度作为聚类依据，思路如下。

无监督聚类性能很大程度上取决于特征选取和聚类策略。实体消歧问题中，多步聚类能有效提高系统性能，为减少错误传递，第一步聚类的准确性尤其重要。本文使用显著相似聚类策略，建立具有高置信度的初始义项类簇。此外，传统的实体消歧方法一般通过抽取歧义词的不同属性或上下文关键词作为特征进行聚类。而在科技文献中，术语上下文词汇能够体现文档领域，但与术语的语义并无直接解释关系。针对该问题，本文利用第一步聚类得到类簇的义项标签对歧义术语进行义项反标注，然后训练 Word Embedding 模型得到各个义项的语义向量，在此基础上实施第二步聚类。在第二步聚类计算特征词权重时，将 Word Embedding 语义相似度与 TFIDF 权重进行线性加权，作为新的特征权重，有效综合了领域特征和义项的语义特征，提高消歧性能。系统结构如图 2 所示。

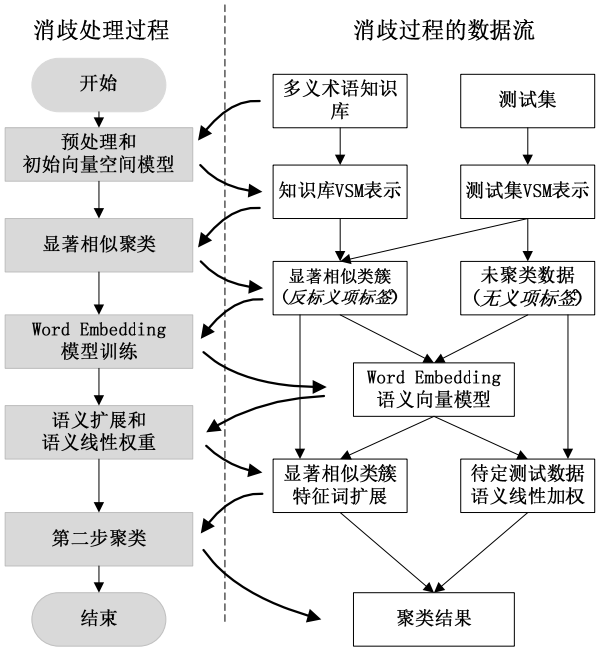

图 2 术语消歧框架

4.2 预处理和初始 VSM 模型

向量空间模型中，文档 d 可以被形式化为一个 n 维向量，其每一维表示词典中的一个词，值为该词的特征权重 s_i，文档 d 可以被形式化为 $d = \{s_1, s_2, ..., s_n\}$。考虑到文档中出现的词汇所代表的信息差异，一般需要对文档进行预处理。本文使用 ICTCLAS[1]对知识库、测试文档集进行分词，然后去掉句子中的标点、符号和停用词，其余词作为特征词。特征词权重一般选用 TFIDF 权重，可以最大程度上区分不同领域文档，在文本分类、信息检索领域得到广泛应用。在消歧问题中，特征词权重应表示该词对当前文档歧义术语各个义项的区分度。在

[1] http://www.ictclas.org/

文档中，与待消歧词语义相关的词语往往出现频率较低，而出现频率较高的实词虽然有助于区分文档，但对区分义项并无明显作用。因此本文对 TF 值进行调整，降低 TF 在权重中的作用，保证低频词信息得到有效利用：

$$tfidf(w) = \sqrt{tf(w)} \times idf(w) \tag{2}$$

预处理后，得到初始的知识库及测试数据的 VSM 模型。根据该模型，任意两个文档间相似度可以由两者向量的余弦相似度计算：

$$\mathrm{Cos}(V_1, V_2) = \frac{\sum_{i=0}^{n} v_{1i} v_{2i}}{\sqrt{\sum_{i=0}^{n} v_{1i}^2} \cdot \sqrt{\sum_{i=0}^{n} v_{2i}^2}} \tag{3}$$

4.3 显著相似聚类

第一步聚类利用初始 VSM 模型，计算义项文档和测试文档的相似度，将满足显著相似条件的测试文档聚类到对应义项中，以抽取高置信度数据。显然，两者相似度越高则越有可能属于同一个义项。文献[16]设计最高相似度与次高相似度的差值阈值，作为选择显著相似文档的准则。本文中，为进一步提高准确度，采用相似度比值阈值作为显著相似条件。

对于缩略术语 w，在知识库中包含 h 个义项 $C_w = \{c_{w1}, c_{w2}, ..., c_{wh}\}$，在测试集中有 m 个文档 $D_w = \{d_{w1}, d_{w2}, ..., d_{wm}\}$。聚类过程以 C_w 中每个义项为中心，计算 d_{wi} 每个文档与所有义项的相似度，并取最高值和次高值文档：

$$(c_{wu}, c_{wv})_{d_{wi}} = \max_{1 \le j \le k} 2(\mathrm{Cos}(d_{wi}, c_{wj})) \tag{4}$$

如果有 $\mathrm{Cos}(d_{wi}, c_{wu}) / \mathrm{Cos}(d_{wi}, c_{wv}) \ge th1$，则 $d_{wi} \in c_{wu}$，否则放弃聚类该文档。显然，阈值 $th1$ 越高，聚类条件越严格，聚类准确度越高，但放弃聚类文档也越多。阈值 $th1$ 既要保证高准确率，又要保留一定样本数量，以达到聚类目的。

由于显著相似聚类可以得到很高的准确度，因此聚类结果可视作对知识库义项文档集的扩充，并作为消歧算法的有标签样本。聚类过程中仍然会引入少量错误数据，但通过 Word Embedding 学习各个义项的语义表示向量，可以有效降低错误聚类数据带来的影响。

4.4 Word Embedding 模型训练

本文使用 Mikolov[2,3]所提出的 Word2Vec 工具实现义项语义的 Word Embedding 训练。Word2Vec 是一个无隐含层的神经网络，直接训练词的 N 维实数向量与内部节点向量的条件概率，并使用了一系列优化方法以提高训练效率。训练结果中，任意两个词的语义相关程度可以通过计算两个词对应向量的余弦相似度得到。

使用 Word Embedding 进行语义消歧，关键问题是如何表示同一术语的多个义项。多义术语每个义项的语义有很大区别，用一个向量很难统一描述。可将多义词进行义项标注，构建带有义项标签的训练语料，用不同标签区分多个义项，再训练 Word Embedding，从而得到不同义项的向量表示。根据该思路，本文利用 4.3 第一步聚类结果，用每个聚类对应的义项标签对歧义术语进行义项反标注，形成标注数据，然后连同未标注数据一同训练。

与神经网络训练类似，Word2Vec 采用随机初始权重，每次训练只得到一个局部最优解，多次训练得到的结果存在差异。当数据规模较小时，这种差异尤其突出。针对该问题，可以从两方面改进：(1)将语料适当重复若干次后训练模型，相当于增加每个样本训练机会，从而降低多次训练间的差异；(2)在同一参数下训练多份向量，在使用过程中综合多份向量结果。此外，数据的排列对神经网络权重训练也会产生影响，本文将训练数据按出现的歧义术语排序，再随机调整少量数据的顺序，使得同一个歧义术语对应的文档相对集中，又有一定

随机性，以提高寻找到最优解的可能性。模型训练过程如图 3 所示。

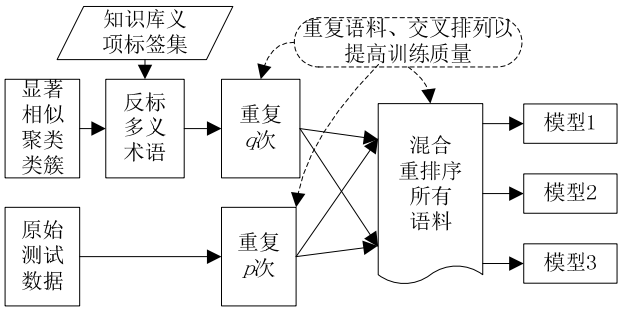

图 3 Word Embedding 训练过程

语料的重复次数对模型的影响可以通过实验进行分析。消歧方法主要利用 Word Embedding 寻找各义项的相关词，因此要求模型中与每个义项最接近的前 k 个词具有较高的一致性，将这 k 个词视为一个集合，则两个模型间的重叠情况可以由 Jaccard 相似系数评价：

$$J_{(p,q)}(V1,V2|D,k)=\frac{|V1\cap V2|}{|V1\cup V2|} \tag{5}$$

其中 $V1$ 和 $V2$ 是同一参数下两次训练得到的模型，D 为义项集合，p 为未标注数据重复次数，q 为标注数据重复次数。测试中，令 k=10，在不同的 p、q 条件下各训练 3 次，求两两 Jaccard 相似系数并取均值，结果见图 4。

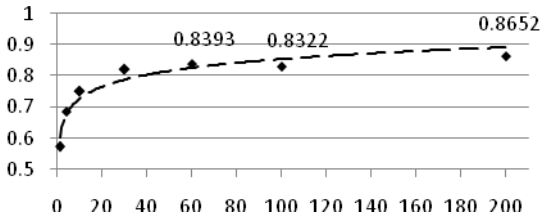

图 4 语料重复次数与 Jaccard 相似度

根据结果，在 p=q=30 之后，训练结果的平均重合度达到 80%以上，此后随着语料重复数量增加，重合度缓慢增长，考虑训练效率因素，在 p=q=60 时就能得到较好的性能。

4.5 基于语义扩展的二步聚类

本节利用 Word Embedding 语义信息实现多义术语消歧，包括两个方面内容：(1)利用语义相似度，对第一步聚类结果进行特征词扩展，弥补文档中缺失的语义信息；(2)用特征词与义项之间的相似度对特征词的 TFIDF 权重加权，提高与义项语义接近的词条的权重。过程中，为降低 Word Embedding 差异导致的误差，使用同一参数重复训练三次，以三个模型结果的交集和平均相似度来计算。

4.5.1 基于语义相似度的特征词扩展

针对第一步聚类类簇中的文档，进行特征词扩展。扩展得到的新特征词不仅要与对应的术语义项相关，也要与文档本身的语境相关。记歧义词 w 的义项标签为 $w\#id$，对应聚类为 $c_{w\#id}\in C_w$。 $c_{w\#id}$ 中的文档记为 $d_{w\#id}$，其 n 个特征词记为 $\{s_1,s_2,...,s_n\}$。扩展使用 3 个相同参数的 Word Embedding 模型，记为 $V1$、$V2$、$V3$。扩展过程如下：

(1)分别计算词 $s_i\in d_{w\#id}$ 在三个向量中语义最接近的 $2r$ 个词，取三者交集，按平均相似度排序后，取前 r 个词得到：

$$VecSim\_r(s_i|V_1,V_2,V_3)=\{s_{i1},s_{i2},...,s_{ir}\} \tag{6}$$

(2)计算所有 s_{ij} 与 $w\#id$ 的相似度均值：$Sim(s_{ij}, w\#id|V_1,V_2,V_3)$，去掉重复词和已有词后，按相似度排序取前 N 项，记为 $\{x_1, x_2, ..., x_N\}$，作为扩展得到的新特征词。过程如图 5 所示。

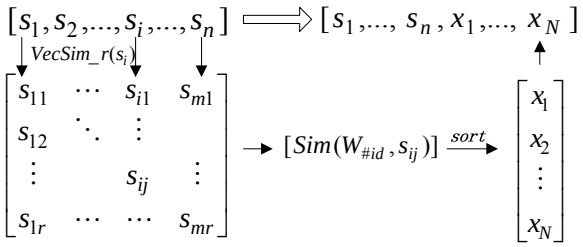

图 5 特征词扩展

在扩展过程中，采用新词的数量 N 非常关键。如果 N 取值太大，将会引入过多的噪声特征，从而降低有效信息；N 取值太小，又无法对原有特征进行有效扩展，合适的 N 值须通过实验得到。扩展得到的新词，能有效弥补当前语境中缺失的语义信息，提高当前文档对歧义词语义的描述能力。

4.5.2 特征词权重的语义线性加权

从直观上，如果特征词与歧义词的语义较为接近，则应该具有更高的权重。而 TFIDF 权重无法考虑这种词与词之间的关联，缺乏对语义信息的描述能力。同样，由 Word Embedding 模型提供的语义向量，能够表示词汇两两间的语义关系，但无法在文档级别计算语义相似程度。本文将两者综合，用特征词与义项的语义相似度对 TFIDF 权重进行线性加权。在计算待消歧文档 d 与义项 $w\#id$ 间相似度时，特征词 $s_i \in d$ 的权重由下式计算：

$$Wt_{w\#id}(s_i) = tfidf(s_i) + Sim(w\#id, s_i|V_1,V_2,V_3)^{\lambda} \tag{7}$$

当 s_i 与义项 $w\#id$ 具有较高语义相似度时，该词特征权重将随之提高。由于语义相似度在[0,1]间，且普遍偏低，故在(7)中添加指数参数 λ，且 0≤λ≤1，提高语义加权幅度。本文中取 λ=0.2。对于第一步聚类而言，可以直接用对应的义项 $C_{w\#id}$ 计算其中各个文档的语义加权。而对于待定的测试文档，则需要在第二步聚类过程中，根据不同的目标义项计算不同的权重，以得到最优聚类结果。

4.5.3 第二步聚类

第二步聚类将待定测试文档归入最有可能的义项类，完成消歧。聚类时，计算待定文档与现有每个义项类的平均相似度，选择相似度最高进行聚类。记包含歧义词 w 的待定文档为 $d_w = \{s_1, s_2, ..., s_n\}$，对应的经过特征词扩展的类簇记为 C'_w。则聚类过程：

（1）求 d_w 特征词对 C'_w 中每个聚类 $c'_{w\#id}$ 的加权权重向量，对任意 $w\#id$，有：

$$V_{(d_w, w\#id)} = \{Wt_{w\#id}(s_1), ..., Wt_{w\#id}(s_n)\} \tag{8}$$

（2）记 $c'_{w\#id}$ 中第 j 个文档特征向量为 $V_{(j,w\#id)}$，则 d_w 与 $c'_{w\#id}$ 的平均相似度用式(9)计算：

$$AvrSim(d_w, c'_{w\#id}) = \frac{1}{|c'_{w\#id}|}\sum_j \text{Cosine}(V_{(j,w\#id)}, V_{(d,w\#id)}) \tag{9}$$

（3）选择平均相似度最大的作为最后聚类结果，即：$d_w \in \hat{c}'_{w\#id}$，有：

$$\hat{c}'_{w\#id} = \arg\max_{w\#id \in C'_w}(AvrSim(d_w, c'_{w\#id})) \tag{10}$$

至此，完成整个聚类过程。

5 实验结果及分析

本文所述消歧聚类方法属于无监督聚类，仅在参数设计时用到少量答案数据，包括显著相似阈值 *th*1 和特征词扩展数量 N；参数设计采用准确率 P%作为评价指标。整个消歧系统性能的测试，以每个歧义术语义项采用聚类准确率 P%、召回率 R%、F 值为评价指标。在整个测试集上，用所有义项的性能指标均值进行评价。

5.1 显著相似性聚类实验

图 6 给出了在不同阈值条件下，聚类文档占总测试文档的比例与聚类正确率之间的关系。其中横轴为阈值，当 *th*1>2.0 后，聚类结果的正确率达到 96%，此时约有一半数据被聚类。此后，随着 *th*1 提高，聚类正确率没有显著变化，而聚类比例则线性下降。因此，可以根据聚类数据比例来制定阈值。按照第一步聚类 30%左右的数据为准，本文设定 *th*1=3.4。

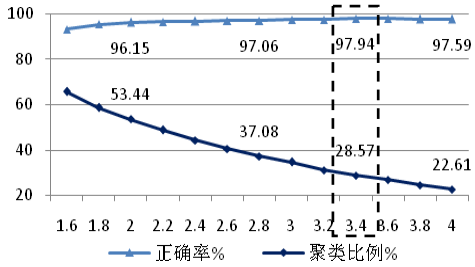

图 6 显著相似阈值对第一步聚类影响

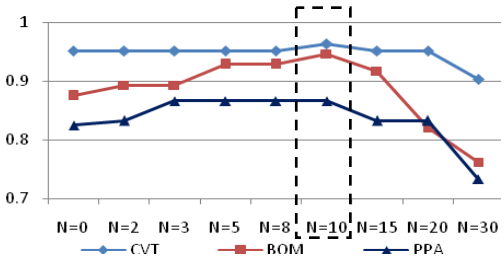

图 7 特征词扩展有效性实验

5.2 特征词扩展实验

对聚类中的文档进行特征词扩展时，扩展词数 N 对最后系统性能有较大影响。以参数 p=q=60 训练 3 个 Word Embedding，特征向量维度均为 100 维。以“CVT”“BOM”“PPA”为例，测试不同的 N 值对第二步聚类准确率的影响。在进行第二次聚类的时候，没有使用语义加权。当 $N \in [2,10]$ 区间时，随着 N 增大，正确率逐渐提高，说明特征词扩展有助于挖掘歧义词语义信息。但当 N 值较大时(N>20)，正确率显著下降，这是由于扩展词过度泛化，引入大量噪声导致。因此，在一定范围内扩展特征词，对提高系统性能有明显效果。本文后续实验中，选取 N=10 进行扩展。实验结果如图 7 所示。

5.3 消歧实验

在前两步实验基础上，对整个测试集进行消歧实验。实验中所用到的参数见表 2。

表 2 实验参数设定

参数名称	取值
显著相似性阈值	3.4
Word Embedding 训练语料重复次数	60
Word Embedding 特征维数	100
扩展特征词数	10

实验设计两个 Baseline 对比消歧系统。Baseline I 选择基本的 TFIDF 权重加权的 VSM 模型，对全部测试数据进行一次聚类，与文献[13]的方法区别在于，其使用歧义词上下文一定窗口内的词作为特征词，而本文中使用文档中除停用词外所有词作为特征词。Baseline II 系统采用与文献[16]类似的两步聚类方法进行。其中，第一步采用显著相似聚类，第二步则利用第一步聚类得到的类簇，不进行特征词和语义加权。Baseline 系统消歧性能见表 3。

实验结果中，利用显著相似聚类得到的结果具有很高的性能。第二步聚类结果的 F 值与待定数据相比有 7%左右的提升，表明第二步聚类能显著改善系统性能。总体性能中，准确率与 Baseline I 相比提升 3.47%，但召回率和 F 值均有超过 10%的提升，该结论与之前相关工作得到的结论较为一致。

表 3 Baseline 消歧实验结果

Type	P%	R%	F	Num

Baseline I		83.30	77.23	75.01	2384
Baseline II	第一步聚类	97.78	95.84	96.16	681
	待定数据	69.89	75.37	67.22	1703
	第二步聚类	80.05	79.36	75.92	1703
	总体性能	86.77	87.01	83.22	2384

本文在 Baseline II 的基础上，通过扩展特征词和特征词语义线性加权两种方法，提升消歧性能，实验结果见表 4。使用“第二步聚类+扩展特征词”方法，各性能指标较 Baseline II 均有 4%左右的提升，总体正确率超过 90%，表明根据 Word Embedding 模型扩展得到的新的特征词能有效补充原有文档中语义缺失，从而对消歧产生显著影响。

表 4 改进后消歧实验结果

Type		P%	R%	F	Num
Baseline I		**83.30**	**77.23**	**75.01**	**2384**
Baseline II	单独	80.05	79.36	75.92	1703
	总体	**86.77**	**87.01**	**83.22**	**2384**
第二步聚类+扩展特征词	单独	85.93	86.65	84.63	1703
	总体	**90.22**	**89.50**	**87.58**	**2384**
第二步聚类+扩展特征词+语义线性加权	单独	89.62	90.88	88.62	1703
	总体	**91.94**	**91.55**	**89.40**	**2384**

在“第二步聚类+扩展特征词+语义线性加权”实验结果中，系统消歧性能进一步提高约 2%。此时，计算特征词在不同义项中的语义相关度，并进行词权重叠加，能使聚类更具有倾向性，但也会导致过拟合。采用线性加权，而非指数加权，可以使权重变化较为平缓，以避免参数过拟合现象。

表 5 歧义术语单独消歧结果

Term	Baseline II			Improved			Effect
	P%	R%	F	P%	R%	F	
ABC	97.80	94.80	95.68	98.80	97.40	97.96	+
ABS	98.33	96.67	97.31	98.16	89.70	93.05	=
AFP	**74.54**	**75.0**	**74.77**	**99.51**	**95.0**	**96.97**	++
AP	95.65	96.96	96.16	95.45	97.06	96.01	=
APC	1.0	1.0	1.0	1.0	1.0	1.0	=
BOM	**83.33**	**95.48**	**86.46**	**93.93**	**98.90**	**96.11**	++
CAD	**64.51**	**85.85**	**66.86**	**60.74**	**73.81**	**62.56**	-
CR	90.0	98.26	92.44	98.0	99.31	98.60	+
CVT	97.02	97.84	97.29	98.26	98.71	98.42	+
DMA	1.0	1.0	1.0	1.0	1.0	1.0	=
EPS	**83.33**	**82.69**	**83.00**	**97.22**	**99.56**	**98.26**	++
FCC	83.33	78.79	73.33	86.67	73.33	70.44	=
MBA	96.92	83.33	88.41	97.69	86.36	90.92	+
MIMO	**67.59**	**63.07**	**62.80**	**88.89**	**97.96**	**92.28**	++
NP	95.0	91.67	92.82	97.5	98.08	97.74	+
OA	98.91	92.30	95.28	94.72	91.75	93.16	-
PCI	91.25	89.06	87.7	96.37	92.19	92.82	+
PCT	1.0	1.0	1.0	1.0	1.0	1.0	=
PPA	**66.67**	**70.0**	**63.33**	**82.52**	**84.34**	**75.72**	++
SAP	**65.34**	**61.25**	**61.08**	**65.38**	**62.1**	**62.20**	=
SAS	82.86	85.07	70.98	86.19	85.42	75.47	+
SIP	99.49	96.67	98.19	1.0	1.0	1.0	+
TT	**82.70**	**83.70**	**82.43**	**90.81**	**98.46**	**93.35**	++
UC	97.62	93.10	95.08	96.15	93.94	94.78	-
UPS	89.07	91.33	89.12	93.53	95.58	94.33	+

表 5 给出所有歧义术语在 Baseline II 和改进方法上的性能比较。表中“+、++、=、-”

分别表明性能有提升、有显著提升、性能可比、性能下降。测试的 25 个术语中，6 个术语的消歧性能有超过 10%的提升，表明 Word Embedding 语义表示方法能够很好地应用于消歧问题；9 条术语的性能有所提升，7 条术语的性能基本持平，说明方法对于大多数术语消歧而言具有一定效果；由于经验参数无法适应所有文档，有 3 个术语的性能没有明显提升。

术语“CAD”和“SAP”在 Baseline II 和改进方法上的消歧性能均较低。经分析，前者有两个义项分别是“计算机辅助设计”和“计算机辅助诊断”，对应文档集合存在许多重叠的特征词，难以区分。后者义项集中有“SAP 软件公司”和“SAP 管理软件”两个定义，分别是公司名和该公司生产的同名软件，因而也具有很高的混淆度。以上义项的区分还需要更深层次的语义关系才能实现。

6 结语

在中文文档中，缩略术语频繁使用，一词多义现象日益突出。缩略术语消歧方法可以应用在计算机辅助翻译、通用机器翻译和信息检索中，具有较高的应用价值。缩略术语专业性强，消歧依赖专业知识库。近几年，互联网进入大数据时代，建立大规模缩略术语知识库成为可能。因此，如何挖掘术语语义特征，利用语义信息实现义项匹配，成为缩略术语消歧的关键问题。本文利用 Word Embedding 提高缩略术语消歧性能，提出无监督地获取每个义项语义表示的方法，在消歧过程中，利用语义信息对特征词进行扩展和语义线性加权，得到精度较高的消歧结果。实验发现，消歧过程中，语义扩展规模不能过大，否则将导致性能降低。这说明每个义项所涵盖的概念范畴往往十分有限，少数词就能描述义项的核心概念。因此，Word Embedding 的核心作用是挖掘文档中缺失的语义信息。该结论对文本数据挖掘和信息检索领域的许多应用有一定参考价值。

参考文献

[1] Banerjee S, Pedersen T. An adapted Lesk algorithm for word sense disambiguation using WordNet [C]. In Pro-ceedings of the 3rd International Conference on Intelli-gent Text Processing and Computational Linguistics, Mexico City, 2002:17-23.

[2] Collobert R, Weston J. A unified architecture for natural language processing: Deep neural networks with multi-task learning [C]. In Proceedings of the 25th International Conference on Machine Learning, Helsinki, 2008: 160- 167.

[3] Mikolov T, Chen K, Corrado G, et al. Efficient Estima-tion of Word Representations in Vector Space [C]. In Proceedings of Workshop at ICLR, 2013.

[4] Mikolov T, Sutskever I, Chen K, et al. Distributed Re-presentations of Words and Phrases and their Composi-tionality[C]. In Proceedings of NIPS, 2013.

[5] Schütze H. Automatic word sense discrimination [J]. Computational Linguistics, 1998, 24(1):97-123.

[6] Mann G, Yarosky D. Unsupervised Personal Name Disambiguation [C]. In Proceedings of CoNLL-2003, Edmonton, 2003:33-40.

[7] Z Peng, L Sun, X Han. SIR-NERD: A Chinese Named Entity Recognition and Disambiguation System using a Two-Stage Method [C]. In Proceedings of the 2ed CIPS-SIGHAN Joint Conference on Chinese Language Processing, Tianjin, 2012:115-120.

[8] J Liu, R Xu, Q Lu, J Xu. Explore Chinese Encyclopedic Knowledge to Disambiguate Person Names [C]. In Pro-ceedings of the 2ed CIPS-SIGHAN Joint Conference on Chinese Language Processing, Tianjin, 2012.

[9] H Liu, Y Lussier, C Friedman. Disambiguating ambi-guous biomedical terms in biomedical narrative text: An unsupervised method [J]. Journal of Biomedical Infor-matics, 2001, 34:249–261.

[10] Stevenson M, Yikun G, Abdulaziz A A, Robert G. Dis-ambiguation of Biomedical Abbreviations [C]. In Pro-ceedings of the Workshop on BioNLP, Boulder, 2009: 71–79.

[11] 王瑞琴,孔繁胜. 无监督词义消歧研究[J]. 软件学报, 2009,20(8):2138-2152.

[12] 张刚,刘挺,等. 隐马尔可夫模型和 HowNet 在汉语词义标注中的应用[J]. 计算机应用研究, 2004,10(增刊): 67-69.

[13] 鲁松,白硕,黄雄. 基于向量空间模型中义项词语的无导词义消歧[J]. 软件学报, 2002,13(6):1082-1089.

[14] 何径舟, 王厚峰. 基于特征选择和最大熵模型的汉语词义消歧[J]. 软件学报, 2010,21(6):1287-1295.

[15] 丁海波, 肖桐, 朱靖波. 基于多阶段的中文人名消歧聚类技术的研究[C]. 第六届全国信息检索学术会议, 牡丹江, 2010:316-324.

[16] 李广一, 王厚峰. 基于多步聚类的汉语命名实体识别和歧义消解[J]. 中文信息学报, 2013, 27(5):29-34.

[17] 杨欣欣, 李培峰, 朱巧明. 基于查询扩展的人名消歧[J]. 计算机应用, 2012, 32(9):2488-2490.

[18] 张榕, 宋柔. 基于互联网的汉语术语定义提取研究[C]. 全国第八届计算语言学联合学术会议, 南京, 2005.

作者简介：于东（1982——），男，博士，讲师，主要研究领域为自然语言处理。Email:yudong_blcu@126.com；
荀恩东（1967——），男，博士，教授，主要研究领域为计算语言学，语言教育技术。 Email: edxun@126.com。

于东

荀恩东