文章编号：1003-0077(2015)06-0023-07

基于知网义原词向量表示的无监督词义消歧方法

唐共波[1,2]，于 东[1,2]，荀恩东[1,2]

(1. 北京语言大学 大数据与语言教育研究所，北京 100083；
2. 北京语言大学 信息科学学院，北京 100083)

摘 要：词义消歧一直是自然语言处理领域中的重要问题，该文将知网(HowNet)中表示词语语义的义原信息融入到语言模型的训练中。通过义原向量对词语进行向量化表示，实现了词语语义特征的自动学习，提高了特征学习效率。针对多义词的语义消歧，该文将多义词的上下文作为特征，形成特征向量，通过计算多义词词向量与特征向量之间相似度进行词语消歧。作为一种无监督的方法，该方法大大降低了词义消歧的计算和时间成本。在 SENSEVAL-3 的测试数据中准确率达到了 37.7%，略高于相同测试集下其他无监督词义消歧方法的准确率。

关键词：词向量；《知网》；词义消歧；无监督方法

中图分类号：TP391 **文献标识码**：A

An Unsupervised Word Sense Disambiguation Method Based on Sememe Vector in HowNet

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

(1. Institute of Big Data and Language Education, Beijing Language and Culture University, Beijing 100083, China;
2. College of Information Science, Beijing Language and Culture University, Beijing 100083, China)

Abstract: Word sense disambiguation (WSD) is a classical issues in nature language processing. In this paper, we trained a language model with the sememe information in HowNet that can represent word semantic, so as to learn the semantic features of words automatically and improve the efficiency of feature learning. Then, we represent words by vectors of sememes. Meanwhile, the contexts of the polysemes is used as features. And then we disambiguate the polysemant by computing the vectors' cosine similarity between polysemes and feature. We choose SENSEVAL-3 as test set, and achieve 37.7% in precision, which is better than other unsupervised method in the same test data.

Key words: word embedding; *HowNet*; WSD; unsupervised methods

1 引言

自然语言中存在着大量多义词，词义消歧对于具有认知能力的人类来说并不是一件困难的事情，但是对计算机自动识别构成了困难。词义消歧(word sense disambiguation, WSD)就是指计算机根据多义词上下文及其他信息进行词义确定的过程。词义消歧在自然语言处理中是一个较为基础且困难的问题，而且会直接影响到信息检索、机器翻译、文本分类、语音识别等上层任务。

目前主流词义消歧的方法有基于知识库的方法和基于语料库的方法。基于知识库的方法覆盖面较大，可以对知识库中所有词进行消歧，而基于语料库的方法则只能针对部分选择的词进行消歧。基于知识库的方法大多借助相关语言的语义知识库进行消歧，比如中文的《知网》[1]、英文的 WordNet 等。基于语料库的方法又分为有监督的方法和无监督的方法。有监督的方法将词义消歧视作一种分类问题，使用包括决策树、决策表、朴素贝叶斯、神经网络、基

收稿日期：2015-07-10 定稿日期：2015-09-16

基金项目：国家自然科学基金(61300081,61170162)，北京语言大学研究生创新基金项目(中央高校基本科研业务费专项资金)(15YCX100)

于实例、支持向量机、自举、集成等在内的方法。而无监督的方法本质上则是聚类问题，可以根据多义词或者多义词的上下文聚类，也可以基于词语的共现来进行消歧。有监督的方法可以获得比较高的准确率，但是需要费时的人工标记，无监督的方法虽然不需要人工标记语料，但是消歧的准确率却明显不如有监督的方法高。

自从2000年《知网》发布以来，面向中文的词义消歧任务中就出现了大量基于《知网》知识的工作。刘群[2]提出了基于《知网》的词语语义相似度计算方法，余晓峰等[3]在刘群的语义相似度计算的基础上，利用多义词的上下文特征进行消歧。车超等[4]借助《知网》中定义的义原（将会在第二节中进行介绍）之间的关系进行消歧。杨尔弘[5]等提出了基于义原同现频率的词义排歧方法。张明宝等[6]借助义原之间的关系，定义语义联系强度，同时定义四条消歧规则进行词义消歧。以上方法克服了训练语料缺乏、数据稀疏的情况，但是在计算语义相似度时会考虑义原之间的关系等情况，计算起来比较复杂。于东等[7]提出了基于word embedding语义相似度的字母缩略术语消歧方法，将词嵌入的思想应用到消歧工作中。

本文将《知网》中可以表示词语语义的义原信息融入到语言模型的训练中，通过义原向量对多义词进行向量化表示。同时，将多义词的上下文作为特征，形成特征向量，并通过计算多义词词向量与特征向量之间相似度进行词语消歧。经过实验对比，准确率略高于其他无监督方法。本文的工作流程如图1所示。

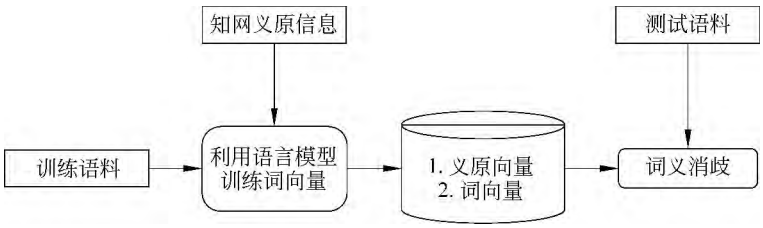

图1 工作流程图

本文将在第二节中介绍基于知网义原词向量表示的词义消歧方法，第三节是实验内容，最后的第四节是对实验的总结与展望。

2 方法

2.1 义原统计信息

《知网》是董振东先生提出的以汉语和英语的词语所代表的概念为描述对象，以揭示概念之间以及概念所具有的属性之间的关系为基本内容的常识知识库。《知网》中包含非常丰富的词语语义信息以及世界知识，为自然语言的各项研究提供了宝贵的资源。本文主要利用了《知网》提供的语义信息，暂未使用世界知识以及概念之间的各种关系。接下来将对知网进行简单的介绍（本文用到的《知网》为2011版）。

《知网》的结构中最重要的两个概念是“概念”和“义原”。“概念”可以理解为词语的一个义项，一个多义词有多个义项，那么也就对应多个“概念”。“义原”是用来描述“概念”的最小意义单元，所有的“概念”都可以用“义原”进行表示。例如，概念“跑酷”的描述语言为：“DEF={fact|事情:CoEvent={exercise|锻炼},domain={sport|体育}}”，其中“事情”、“锻炼”、“体育”均为“跑酷”的描述义原。《知网》并没有《同义词词林》和WordNet那样的树状结构，而是通过义原之间的关系将所有的概念进行关联，组成一个网状的知识库。《知网》中的义原包含实体、事件、属性、属性值、动态角色与属性、次要特征以及专有名词七大类义原，共计2 448个。本文主要利用的就是这些义原信息。

《知网》中以概念为单位进行描述，每一条描述作为一个记录。具体形式如表1所示。其中NO.

表1 《知网》中的记录

NO.=120283
W_C=少
G_C=verb[shao3]
S_C=
E_C=还～两个杯子，～双筷子，～了三块钱
W_E=be short
G_E=verb
S_E=
E_E=
DEF={lack|缺少}

表示在《知网》中记录的序号，W_C、G_C、S_C、E_C 分别表示中文词语、词性、情感极性、例子，W_E、G_E、S_E、E_E 分别表示英文词语、词性、情感极性、例子，DEF 是知网的知识描述语言，是《知网》的核心内容。

由于本文面向中文进行词义消歧，暂时只是使用《知网》中的中文部分。《知网》中，我们将只有一个中文概念，而且概念只用一个义原进行描述的词定义为单义原词，反之则定义为多义原词。经统计，单义原词有 35 347 个，构成单义原词的义原个数为 1 492，占义原总数的 60.95%；多义原词 69 382 个，构成多义原词的义原个数为 2 041，占义原总数的 83.37%。其中多义原词中 71.88%(1 467/2 041)的义原也出现在单义原词中的义原(图 2)，这就说明大部分的词语是可以通过单义原词中的义原来进行语义的表示。平均每个单义原词中的义原可以表示 23.7 个单义原词，而且大部分的单义原词由少量义原构成(图 2)，说明利用义原来表示单义原词可以明显减少特征的训练。绝大多数的多义原词由少于七个义原构成(图 2)，说明利用义原向量来表示多义词也是简单可行的。

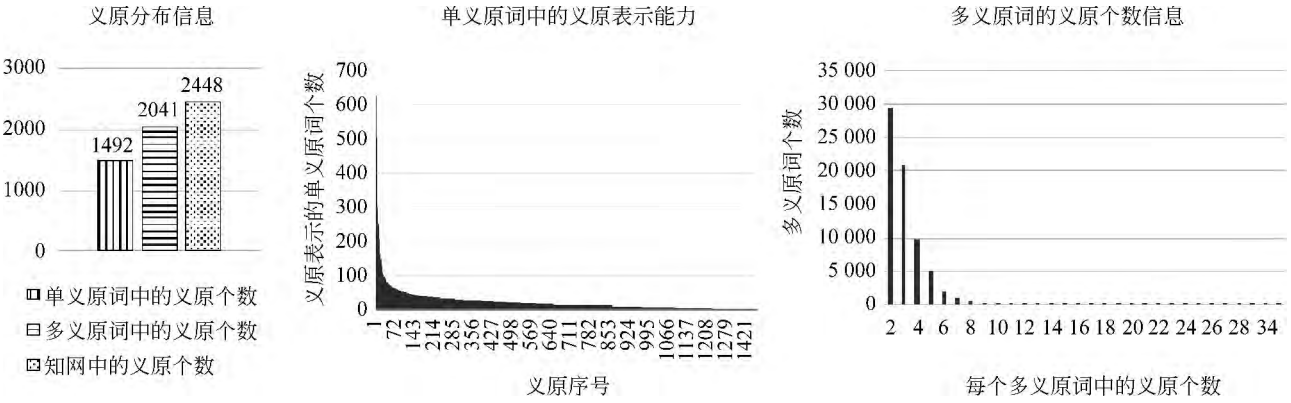

图 2　义原统计信息

2.2　义原词向量的训练

要将自然语言交给机器学习模型处理，通常首先需要将语言形式化，我们选择词向量来实现这一过程。一种最简单的词向量方式是用维度为词表大小的向量来表示一个词，而且只有一个维度的值为 1，这个维度表示当前的词语，其他维度全是 0。这种词的表示有两个缺点：一是容易受维数灾难的困扰；二是不能很好地刻画词与词之间的相似性，这种情况下，词向量之间的距离都是相同的。另外一种是分布式表征方法，通过训练将某种语言中的每一个词映射成一个固定维度的向量，将所有这些向量放在一起形成一个词向量空间，而每一向量则为该空间中的一个点，在这个空间上引入“距离”，就可以计算词语之间在语义语法之间的相似度。

Word2vec[8] 是 Google 公司开源的一个用于将词语进行向量化表示的工具，是神经网络语言模型的简化，有 CBOW 和 Skip-gram 两种模型[9-10]。只需要输入要训练的语料，便可以输出语料中的词语对应的词向量。得到的向量可以在自然语言处理和机器学习的任务中用来表示特征。Mikolov 等[11] 发现通过 word2vec 训练出来的词向量具有一定的表示词汇语法和语义关系的能力，因此可以通过计算词语对应的向量之间的相似度来得到词汇之间的语义相似度。

在普通的词向量训练模型中，一个词语只有一个词向量，对于多义词来说，一个词向量显然是不够的，Huang 等[12] 为多义词训练多个词向量，Chen 等[13] 为多义词的每个义项训练相应的词向量。因此我们用不同的词向量来表示多义词的每一个概念。为了增强词向量的语义表达能力，我们将《知网》中的具有语义表征能力的义原信息融入到词向量的训练过程中，既可以得到普通词语的词向量，又可以得到义原的表示向量，为下一步多义词词向量的表示以及词义消歧提供支持。词向量的训练步骤为：

a) 将原始语料进行分词，其中语料选取自 1.6G 的现代汉语语料；

b) 在保留原始语料的基础上，将语料中出现的单义原词替换为对应的义原，同时加上标签，例如，“参观”被替换为“【看】”。通过训练我们就可以得到表示词语“参观”和义原“看”的向量。由于单义原词

中义原数量远远小于单义原词本身(4%),因而这一过程可以大大减少数据稀疏对训练造成的困扰;

c) 将处理后的语料作为 word2vec 模型的输入,词向量的维度设为 100,上下文窗口为 5,选择 CBOW 方法进行训练。

2.3 词义消歧方法

本文主要通过计算多义词与上下文的特征向量之间的语义相似度来进行词义消歧,相似度最高的义项作为多义词在该语境下的语义。

2.3.1 多义词与特征向量的相似度

多义词与特征向量之间的相似度计算主要通过计算多义词的概念与多义词的特征向量之间的相似度来实现。选取相似度最大的概念作为该多义词在当前上下文的解释。假设多义词有 N 个概念,那么概念向量集合为$\{c_1,c_2,\cdots,c_n\}$。特征向量设为 F。则有式(1)。

$$C' = \max_{i=1,\ldots n} \mathrm{sim}(c_i,F) \tag{1}$$

C′即为多义词在当前上下文的最佳的候选概念(义项)。而 $sim(c_i,F)$表示概念与特征向量的相似度。

2.3.2 概念与特征向量的相似度

概念一般是由多个义原进行表示,概念的向量表示有以下三种方法:

a) sumVec:将表示概念的义原的向量进行累加,累加结果用来表示概念。

b) averVec:求表示概念的义原的向量平均值,平均值用来表示概念。

c) allVec:表示概念的所有义原的向量均用来表示概念。

概念与特征向量的相似度计算方法:

a) 当使用 sumVec 和 averVec 方法表示向量时,相似度为两个向量的余弦距离。

b) 当使用 allVec 表示向量时,采用 Mihalcea[14]的向量对齐的方法计算相似度。

向量对齐方法:假设有 J 个向量表示特征,特征向量集合为 F:$\{f_1,f_2,\cdots,f_j\}$,有 K 个向量表示概念,概念向量集合为 C:$\{c_1,c_2,\cdots,c_k\}$,F 和 C 中的向量两两之间计算余弦相似度,取相似度最大的一对作为已对齐的向量,插入到集合 P 中。

P:$\{\{(f_1,c_m)\cdots\}(f_1\in F,c_m\in C)$。然后从 F 和 C 中将已选择的向量删除,循环执行此步骤,直至 F 或者 C 变成空。最后计算所有对齐向量的平均余弦相似度,如式(2)所示。

$$\mathrm{sim}(F,C) = \frac{1}{n}\sum_{i=1,p_i\in P}^{n}\cos p_i \tag{2}$$

其中

$$n = \min(j,k) \tag{3}$$

3 实验

3.1 特征的选取与表示

本文选取多义词的上下文信息作为特征。Zheng-Yu Niu[15]认为词语的上下文窗口为 10 以内的时候作特征最好。Ke Cai[16]的实验结果表明上下文特征窗口为 5(−5,+5)的时候结果最好。Huang Heyan[17]的实验结果则表明上下文窗口为 1 的时候效果最好。本文分别选取上下文窗口为 1 的实词和整句话中除多义词以外的所有实词作为特征进行了两组实验。

由于选取的上下文特征词一般都不止有一个,所以如何表示特征向量又是一个问题。针对这个问题,我们同样采用了三种方法:

a) sumVec:将上下文窗口内实词的向量进行累加,累加结果作为特征向量;

b) averVec:求上下文窗口内实词的向量平均值,平均值作为特征向量;

c) allVec:上下文窗口内所有实词的向量均作为特征向量。

3.2 实验与分析

本实验训练语料来自北京语言大学现代汉语语料库(BCC)[18]的文学综合语料,包括文学、报刊类,共计 13 亿字左右。总共训练了 182 398 个义原实例,测试数据来自 SENSEVAL-3 的中文词义消歧评测任务,有 20 个汉语词的 379 个实例。如“材料”这个多义词的消歧示例如图 3 所示。

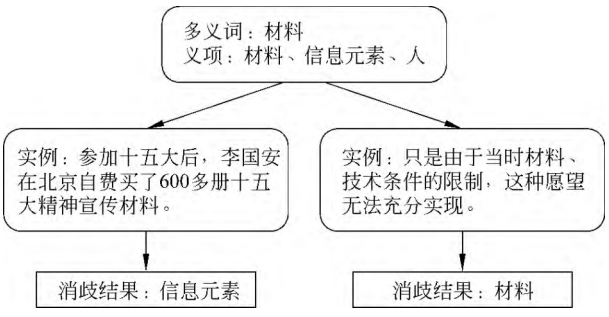

图 3 “材料”的消歧示例

由于未参加当年评测,故无法获取官方的测试

集标注结果，本实验的测试集标注方法为在《知网》2011 版的基础上进行的手工标注。

正确率的计算公式如式(4)所示。

$$P = \frac{n_c}{n_a} \tag{4}$$

其中 n_c 表示标注正确的实例数，n_a 表示所有的实例数。

本文做了六个对比实验，如表 2 所示。

表 2　对比实验

向量表示方法 / 特征选取	sumVec	averVec	allVec
最近邻实词	实验一 (29.6%)	实验三 (31.1%)	实验五 (26.1%)
上下文所有实词	实验二 (32.7%)	实验四 (37.7%)	实验六 (25.6%)

实验发现，实验四的准确率最高，达到了 37.7%，最差的是实验六，准确率只有 25.6%。三种向量的表示方法效果相比：

averVec ＞ sumVec ＞ allVec

同时，选取上下文所有的实词作为特征的实验结果普遍好于使用最近邻实词作为特征的方法。

表 3 是实验四中多义词“钱”的部分实验结果，其中“钱”共有“单位”、“货币”、“钱财”、“姓”、“资金”五个义项。

具体的实验结果如表 4 所示。

SENSEVAL-3 中对应的中文评测最优的系统准确率可以达到 66.5%，但是该系统采用的是有监督的学习方法，我们的实验是采用完全无监督的。同时经过仔细观察 SENSEVAL-3 的训练数据和测试数据发现，测试语料与训练语料的相似度比较高，

表 3　“钱”的部分消歧结果与分析

消歧句	正确义项	消歧结果	分　析
丁佩玲(女)、…… 顾煜麟、＜head＞钱＜/head＞月宝(女)、钱红菱(女)、钱易(女)、钱海鑫、徐家基、徐德郁(女)……	姓	无	消歧句的上下文完全由人名构成，由于数据稀疏的关系导致没有对应的词向量，无法进行消歧。
她又用这笔＜head＞钱＜/head＞买了一头耕牛，发展生产，效果显著。	资金	资金	消歧正确
临走，李清林留下 500 元＜head＞钱＜/head＞，说这是个人的一点心意，暂时解决一下困难。	货币	资金	由于表示货币和资金的义原比较相似，未能够区别。 货币：货币、商业、金融 资金：资金、金融

表 4　实验结果

词	义项数	实例数	实验一正确数	实验二正确数	实验三正确数	实验四正确数	实验五正确数	实验六正确数
路	9	28	6	6	6	8	8	6
冲击	4	13	4	7	4	8	6	6
运动	5	27	6	9	6	9	5	6
日子	3	21	13	11	13	14	10	9
老	12	26	4	5	4	5	3	4
没有	4	15	4	5	4	7	7	8
穿	5	14	3	2	3	2	4	2
地方	7	17	6	2	6	5	0	0
活动	7	16	7	8	7	8	4	6
少	4	19	8	9	8	10	7	11
坐	6	12	1	2	1	5	2	6
分子	3	16	5	2	5	6	6	1

续表

词	义项数	实例数	实验一正确数	实验二正确数	实验三正确数	实验四正确数	实验五正确数	实验六正确数
把握	4	15	8	7	8	7	0	0
突出	6	15	6	6	6	6	0	0
走	9	24	8	8	8	8	7	7
包	11	36	2	6	5	6	5	2
起来	6	20	2	3	3	3	3	0
钱	5	20	5	7	6	7	5	7
研究	2	15	7	11	7	11	11	11
材料	3	10	7	8	7	8	6	5
总数	115	379	112	124	118	143	99	97
正确率/%			29.6	32.7	31.1	37.7	26.1	25.6

在某种程度上降低了有监督学习的难度，而且20个多义词只有79个义项需要消歧，而我们使用的是《知网》2011版，共有115个义项，义项数量是评测任务的146%，消歧难度较之前工作大大增加。另外，Wanyin Li[19]也使用了SENSEVAL-3的数据进行实验，将词语的搭配信息和主题信息作为特征，利用贝叶斯分类器实现词义的消歧，准确率为37.6%。虽然实验设计不同，但是同样作为无监督的方法，本文提出的方法能够得到更高的准确率，说明该方法是有效的。由于没有2002版《知网》数据，所以没有再进行进一步的实验，但是理论上正确率应该会更高。

4 结论与未来工作

词义消歧是自然语言处理中的基础与难点，我们将《知网》中可以表示词语语义的义原信息融入到词向量的训练中，利用义原向量对多义词进行向量化表示，并将其应用到词语消歧。实验结果表明：使用averVec这种方法来表示向量，以及选取整句的实词作为特征是可行的。

本实验只是初步的对此方法进行了探索，后续需要完善和优化的地方还有很多。由于词向量的训练本身准确率并不能达到100%，而且《知网》中表示单义原词的义原也只能表示81.4%的词语，这就决定了实验最后的准确率肯定达不到100%，鉴于词义消歧的困难性，这些环节有待进一步的优化。今后将会在注入本体知识的词向量的训练、特征的选取与表示、《知网》的常识信息的利用等方面进行更深入的研究。

参考文献

[1] 董振东，董强.《知网》.[DB]http://www.keenage.com

[2] 刘群，李素建. 基于《知网》的词汇语义相似度的计算[C]. 第三届汉语词汇语义学研讨会. 台北，2002:59-76.

[3] 余晓峰，刘鹏远，赵铁军. 一种基于《知网》的汉语词语词义消歧方法[C]. 第二届全国学生计算语言学研讨会，北京：中国中文信息学会，2004.

[4] 车超，金博，滕弘飞，等. 基于义原关系的多策略汉语词义消歧方法[J]. 大连理工大学学报，2010，50(4)：603-608.

[5] 杨尔弘，张国清，张永奎. 基于义原同现频率的汉语词义排歧方法[J]. 计算机研究与发展，2001，38(7)：833-838.

[6] 张明宝，马静. 一种基于知网的中文词义消歧算法[J]. 计算机技术与发展，2009，19(2)：9-11，15.

[7] 于东，荀恩东. 基于Word Embedding语义相似度的字母缩略术语消歧[J]. 中文信息学报，2014，28(5)：51-59.

[8] Mikolov T. Word2vec Project[DB/OL]. http://code.google.com/p/word2vec/.

[9] Mikolov T，Kai Chen，Greg Corrado，et al. Efficient estimation of word representations in vector space[C]//Proceedings of the ICLR Workshop，2013.

[10] Mikolov T，Yih W，Zweig G. Linguistic regularities in continuous space word representations[C]//Proceedings of the HLT-NAACL. 2013.

[11] Mikolov T，Sutskever I，Chen K，et al. Distributed representations of words and phrases and their com-

positionality[C]//Proceedings of the Advances in Neural Information Processing Systems. 2013：3111-3119.

[12] Huang E H, Socher R, Manning C D, et al. Improving word representations via global context and multiple word prototypes[C]//Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics：Volume 1. Association for Computational Linguistics, 2012：873-882.

[13] Chen X, Liu Z, Sun M. A unified model for word sense representation and disambiguation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. 2014：1025-1035.

[14] Mihalcea R, Corley C, Strapparava C. Corpus-based and knowledge-based measures of text semantic similarity[C]//Proceedings of the American Association for Artificial Intelligence MA, 2006.

[15] Niu Z Y, Ji D H, Tan C L. Optimizing feature set for Chinese word sense disambiguation[C]//Proceedings of Senseval-3, Third International Workshop on Evaluating Word Sense Disambiguation Systems. 2004.

[16] Ke Cai, Xiaodong Shi, Yidong Chen, et al. Chinese Word Sense Induction based on Hierarchical Clustering Algorithm[C]//Proceedings of the CLP, 2010.

[17] Huang Heyan, Yang Zhizhuo, Jian Ping. Unsupervised Word Sense Disambiguation Using Neighborhood Knowledge[C]//Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation, 2011：333-342

[18] **北京语言大学汉语语料库**[DB]. http://www.bcc.blcu.edu.cn/

[19] Li W, Lu Q, Li W. Integrating Collocation Features in Chinese Word Sense Disambiguation[C]//Proceedings of the Fourth Sighan Workshop on Chinese Language Processing. 2005：87-94.

唐共波(1988—)，**硕士，主要研究领域为自然语言处理。**

E-mail：tanggongbo@126.com

荀恩东(1967—)，**通信作者，博士，教授，主要研究领域为计算语言学、语言教育技术。**

E-mail：edxun@126.com

于东(1982—)，**博士，讲师，主要研究领域为自然语言处理。**

E-mail：yudong_blcu@126.com