文章编号：1003-0077（2017）00-0000-00

面向中朝跨语言文本分类的双语主题词嵌入模型的研究

王琪 田明杰 崔荣一

（延边大学 计算机科学与技术学科 智能信息处理研究室，吉林 延吉 133002）

摘要：日渐丰富的少数民族跨语言文字信息资源，对其进行有效的管理、挖掘与利用有着重要的应用价值。为了解决语言间的差异，解决语言鸿沟问题，针对中朝跨语言文本分类任务，提出了双语主题词嵌入模型。该模型将主题模型与双语词嵌入模型相结合，解决了一词多义引起的歧义性对跨语言文本分类的精度带来的影响。首先，在大规模包含词对齐信息的平行句对中训练双语单词的词嵌入表示；其次，对双语分类语料进行主题模型的建模，并获得双语单词的双语主题词嵌入表示；最后，将双语单词的双语主题词嵌入表示输入至传统文本分类器与深度学习文本分类器，进行模型的训练与分类预测。实验结果表明，在中朝跨语言文本分类任务中 Accuracy 达到了 91.76%，达到了实际应用水平，并且根据双语单词间的相似度进行排序，该文提出的模型可以对一词多义单词的多个词义有很好的表示。

关键词：中朝跨语言文本分类；双语词嵌入模型；主题模型；一词多义

中图分类号：TP391 **文献标识码：**A

Research on Bilingual Topical Word Embedding Model for Chinese-Korean Cross-lingual Text Classification

TIAN Mingjie, Wang Qi, CUI Rongyi

(Intelligent Information Processing Lab., Department of Computer Science and Technology, Yanbian University, Yanji, Jilin 133002 ,China)

Abstract: It is worthwhile to manage and mine the cross-language information resources of ethnic minorities. In order to solve the differences and gap between languages, bilingual topical word embedding model is proposed for the Chinese-Korean cross-lingual text classification task. The model combines the topic model with the bilingual word embedding model to solve the influence of the ambiguity caused by polysemy on the accuracy to cross-lingual text classification. Firstly, the word embedding representation of bilingual words were trained in a large scale parallel sentence pairs with word-aligned. Secondly, modeled the dataset of classification task by topic model, and obtained the bilingual topical word embedding representation of bilingual words. Finally, input the bilingual topical word embedding representation to the traditional text classifier and the deep learning text classifier, and trained the model and predicted the new text. The experimental results show that the accuracy reach 91.76% in the Chinese-Korean cross-lingual text classification task, reaching the practical application level. And according to the similarity between bilingual words, the proposed model can represent the multiple meanings of polysemy words.

Key words: Chinese-Korean cross-lingual text classification; bilingual word embedding model; topic model; polysemy

收稿收稿日期：2017-03-16；**定稿日期**：2017-04-26

基金项目：国家语委“十三五”科研规划项目（YB135-76）；延边大学外国语言文学世界一流学科建设科研项目(18YLPY13)

0 引言

如今的信息资源，不仅在规模上迅猛增长，资源类型及所使用的语言种类也越来越多样化。其中大量少数民族语言文字信息资源接入互联网的大环境，丰富了互联网资源的语言多样性。对少数民族语言文字资源进行有效的管理、挖掘和利用，有着重要的意义和价值。语言种类的多样性丰富了信息资源，但是语言间的差异性，不可避免地给用户利用信息资源带来了阻碍。跨语言文本分类技术能够有效地组织多语言信息资源，解决语言鸿沟的问题，可以消除因语言的差异给人们带来的信息检索和文本分类的困难，帮助人们更好地理解语言信息，有利于进行知识的交流与共享。

跨语言文本分类（cross-lingual text classification）是利用已标注好类别的一种语言的文本训练集训练得到分类器，并对另一种语言未标注类别的文本进行分类的过程。跨语言文本分类研究方法主要有基于双语词典的方法、基于机器翻译的方法、基于主题模型的方法以及基于双语词嵌入的方法。

Bel等[1]使用基于统计的双语词典对预先定义的每个类别源语言文本的前k个词进行翻译，然后使用双语词典将待分类文本翻译成目标语言，最后通过比较文本间相似度确定文本的所属类别。Rigutini[2]使用机器翻译技术对英文和意大利文进行了翻译，并将EM算法应用于跨语言文本分类任务中。Ni[3]提出ML-LDA模型，从维基百科抽取多语言主题，在多语言主题空间上进行跨语言分类。Heyman[4]利用可对比语料库通过主题模型结合迁移学习的方法实现了跨语言文本分类。

基于双语词嵌入的方法指将原向量空间中的单词表示映射为固定维数的另一个空间的向量，双语单词的词嵌入表示共享了只存在于单一语言内部独有的信息。Luong[5]假设两种语言中对齐的单词跨语言地共享上下文，通过在损失函数中添加正则项的方式训练单词的词嵌入。Faruqui[6]利用双语词典与典型相关性分析对双语文本建模，并找出两个语言空间下相关系数最大的向量。Hermann[7,8]在拥有相似句法形式的句子中学习词嵌入，提高跨语言文本分类的准确率。Gouws[9]通过在损失函数中添加句子级别对齐语料的正则项的方式训练词嵌入模型，在英语与法语单词对齐部分使用局部的对齐方式来代替全局的对齐，节省了算法的时间与空间复杂度。Vulić[10]归纳单词的出现规律，并结合双语词嵌入模型对文本级别对齐的语料库建模，最终学习词嵌入表示。

当前朝鲜语文本分类研究中，周国强[11]使用类TF-IDF估算方法计算权重，利用朴素贝叶斯分类器实现了对朝鲜语文本的分类。Lee[12]通过计算单词与标注类别间的互信息识别多义词，以增加词项数来增加文本向量的维度，降低容错率，提高分类准确率。

基于词典的方法未能很好地解决词的多义性以及对未登录词的处理。基于机器翻译的方法需要依赖机器翻译系统，翻译系统的准确率会较大影响文本翻译质量。单独使用潜在语义文本表示只考虑到单词的全局分布，没有考虑到单词与上下文之间的局部关系。双语词嵌入模型中多义词的所有含义共享同一个向量，存在歧义性问题。对朝鲜语文本分类的研究，当前停留在对单语种的文本分类，还没有以朝鲜语为双语之一的跨语言文本分类的研究，无法满足多语言环境下朝鲜语言文字信息化与智能处理的需求。

本文为了挖掘出中朝双语文本中的隐含语义，提出双语主题词嵌入（Bilingual Topical Word Embedding, Bi-TWE）模型，该模型是在双语词嵌入方法的基础上结合主题模型的思想，弥补词嵌入模型在单词的表示上无法解决歧义性与一词多义对文本分类精度带来的影响。并且利用本文提出的双语主题词嵌入模型结合基于深度学习的文本分类算法提高跨语言文本分类的精确度。

1 相关工作

跨语言文本分类中训练集与测试集的文本属于不同的语种。由于涉及到的语言多于一种，不同于单语种文本分类的文本表示需要将多种语言的文本映射到统一的表示空间，所以统一空间上的文本表示是跨语言文本分类的研究重点。

1.1 单一语言文本表示模型

1.1.1 LDA 主题模型

LDA 主题模型是由 Blei[13]于 2003 年提出的一种文档主题生成模型，LDA 主题模型在文档-主题与主题-词项的分布上引入了 Dirichlet 先验参数[14]，应对了在对大量文本数据建模时可能会出现的过拟合问题。

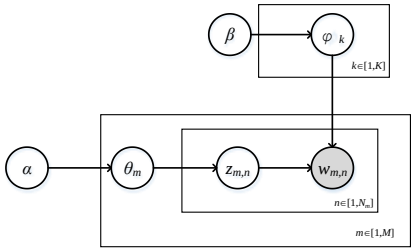

图 1 LDA 主题模型

LDA 主题模型是典型的有向概率图模型[15]，如图 1 所示。其中 $w_{m,n}$ 表示第 m 篇文档第 n 个词，α、β 为根据经验给定的 Dirichlet 分布的超参数，$z_{m,n}$ 是单词 $w_{m,n}$ 所对应的主题，参数 θ_m 为第 m 篇文档在主题上的分布，参数 φ_k 为第 k 个主题在词上的分布。

在构建 LDA 主题模型的过程中 Collapsed Gibbs[16]采样方法对模型的参数进行估计。Collapsed Gibbs 采样可以通过积分避开实际待估计的参数，转而对每个词的主题进行采样。单词序列下主题序列的条件概率计算如下：

$$p(z_i = k \mid \mathbf{z}_{\neg i}, \mathbf{w}) = \frac{n^t_{k,\neg i} + \beta_t}{\sum_{v=1}^{V}(n^v_k + \beta_v) - 1} \cdot \frac{n^k_{m,\neg i} + \alpha_k}{\sum_{l=1}^{L}(n^l_m + \alpha_l) - 1} \quad (1)$$

其中 z_i 表示第 i 个单词对应的主题；$\neg i$ 表示不包括其中的第 i 项；n^t_k 表示 k 主题中出现词 t 的次数；β_t 是词 t 的 Dirichlet 先验；n^l_m 表示文档 m 出现主题 l 的次数；α_l 是主题 l 的 Dirichlet 先验。

1.1.2 词嵌入模型

词嵌入是指把单词的高维向量表示（如 One-hot 表示方法）降低维度到低维的向量空间中，单词向量的每个分量为一个实数。自然语言处理领域指的词嵌入通常是指神经网络语言模型中间产物的单词的向量表示。

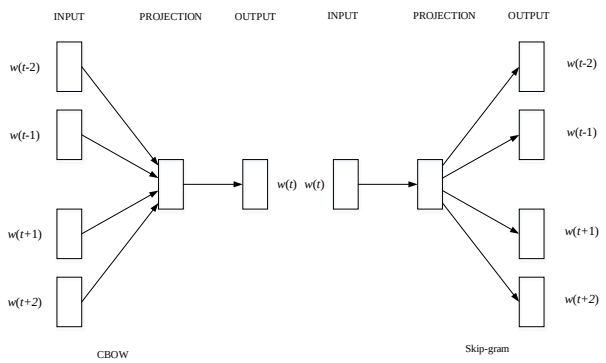

图 2 CBOW 与 skip-gram 模型

Word2Vec 词嵌入模型是谷歌的 Mikolov[17]于 2013 年开源推出的一个用于获取单词的向量表示的工具包，它简单、高效引起很多人的关注。Word2Vec 模型包含两个模型 CBOW 和 skip-gram，其模型结构如图 2 所示。

skip-gram 模型是通过输入中心词 w，通过神经网络预测中心词 w 的前 c 个词和后 c 个词。Mikolov 在文献[18]中提出负采样方法对词嵌入的训练进行了加速。负采样方法与 skip-gram 模型的结合是最为常用的 Word2Vec 词嵌入模型训练方法。

在 skip-gram 模型中，已知中心词，需要预测中心词的上下文，因此，对于中心词 w，$Context(w)$中的单词为正样本，负样本为词典中其余的单词。$Context(w)$中的单词 u 选定的负样本集合 $NEG(u)$，对于正样本,模型要最大化目标函数：

$$G = \prod_{w \in C} \prod_{u \in Context(w)} \prod_{z \in \{u\} \cup NEG(u)} p(z \mid w) \quad (2)$$

其中条件概率 $p(z|w)$的定义为：

$$p(z \mid w) = [\sigma(\mathbf{v}(w)^{\mathrm{T}} \theta^z)]^{L^u(z)} \cdot [1-\sigma(\mathbf{v}(w)^{\mathrm{T}} \theta^z)]^{1-L^u(z)} \quad (3)$$

当 z 为上下文单词 u 时，等号右侧乘法的前一项起作用，z 为 u 的负样本时，等号右侧乘法的后一项起作用。

1.2 双语表示模型

1.2.1 双语 LDA 主题模型

双语 LDA 主题模型是 LDA 主题模型的双语扩展。双语 LDA 主题模型使用一组与语言无关的“通用”主题，对文档的 2 种不同语言描述内容进行建模，每个“通用”主题都有 2 种不同的表示形式，每种表示形式与一种语言相对应。双语 LDA 主题模型的概率图模型如图 3 所示：

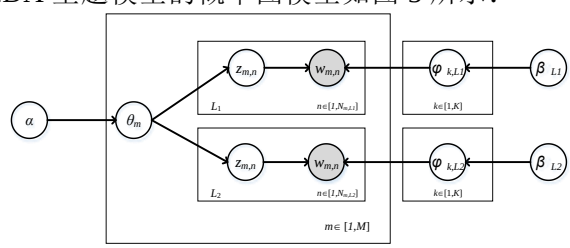

图 3 双语主题模型

其中 $w_{m,n}$ 表示第 m 篇文档第 n 个词，α 为根据经验给定的 Dirichlet 分布的超参数，β_{L_j} 为主题在语言 L_j (j = 1 , 2)上的 Dirichlet 分布的超参数，$z_{m,n}$ 是单词 $w_{m,n}$ 所对应的主题，参数 θ_m 为第 m 篇文档在主题上的分布，ϕ_{k,L_j} 为第 k 个“通用”主题在语言 L_j 的词项上的分布；给定文档集合 D，由

M 篇文档构成，每篇文档包含两种语言的内容表示，第 m 篇文档的语言 L_j 部分包含 $N_{m,Lj}$ 个单词。

在参数估计阶段，需要计算单词序列下的主题序列条件概率 $p(z_{i,L_j}=k|\mathbf{z}_{\neg i,L_j},\mathbf{w}_{L_j})$，其中 $\mathbf{w}_{L_j}$ 表示语言 L_j 的词项的矩阵表示；z 表示这些词项的主题分配；$\mathbf{z}_{\neg i,L_j}$ 表示在不考虑当前 L_j 语言中第 i 个词的主题分配。条件概率的计算公式如下：

$$p(z_{i,L_j}=k|\mathbf{z}_{\neg i,L_j},\mathbf{w}_{L_j}) = \frac{n^t_{k,\neg i,L_j}+\beta^t_{L_j}}{\sum_{v=1}^{V_{L_j}}(n^v_{k,L_j}+\beta^v_{L_j})-1}\cdot\frac{\sum_{j=1}^{2}n^k_{m,\neg i,L_j}+\alpha_k}{\sum_{l=1}^{L}(\sum_{j=1}^{2}n^l_{m,L_j}+\alpha_l)-1} \quad (4)$$

其中 $n^t_{k,\neg i,L_j}$ 表示不考虑当前词项 t 的当前主题分配的情况下，语言 L_j 中词项 t 的主题分配到 k 的次数；$\sum_{v=1}^{V_{L_j}}(n^v_{k,L_j}+\beta^v_{L_j})-1$ 表示不考虑当前词项 t 的当前主题分配的情况下，语言 L_j 中所有词项的主题分配到 k 的次数，$\beta^v_{L_j}$ 为主题 v 在语言 L_j 上分布的 Dirichlet 超参数；V_{L_j} 表示语言 L_j 的词典；$n^k_{m,\neg i,L_j}$ 表示不考虑当前词项 t 的当前主题分配的情况下，文档 m 中语言 L_j 所有词项的主题分配到 k 的次数；$\sum_{l=1}^{L}(\sum_{j=1}^{2}n^l_{m,L_j}+\alpha_l)-1$ 表示忽略当前词项 t 的情况下，文档 m 中两种语言的单词总数，α_l 为决定文档在主题 l 上分布的 Dirichlet 超参数。

1.2.2 双语 skip-gram 模型

双语 skip-gram 模型是单语 skip-gram 模型的双语扩展，该模型利用单语的上下文语境共存信息和双语约束的语义等价信息训练得到双语词嵌入。双语 skip-gram 模型在优化方法上使用负采样方法，采样一定个数的负样例可帮助模型区分正负。

双语 skip-gram 模型共同的目标函数中添加了与单语环境下单词与上下文约束关系类似的跨语言约束项，目标函数如下：

$$G=\alpha(Mono_1+Mono_2)+\beta Bi \quad (5)$$

其中每个单语模型 $Mono_1$ 与 $Mono_2$ 旨在获得每种语言的聚类结构，而双语约束项 Bi 将两种语言的表示空间联系在一起。

双语 skip-gram 模型能够对双语单词使用标准的 skip-gram 模型进行交叉预测单词。采用双语 skip-gram 模型预测上下文的示例如图 4 所示。

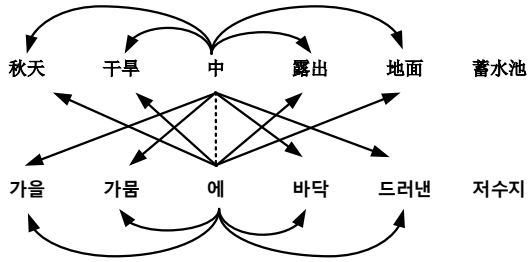

图 4 双语 skip-gram 模型

当给定语言 L_1 中的单词 w_1 和另一种语言 L_2 中的单词 w_2 之间的对齐关系，双语 skip-gram 模型使用单词 w_1 来预测单词 w_2 的上下文，使用单词 w_2 来预测单词 w_1 的上下文。最终对目标函数进行优化，可以获得语言 L_1 与语言 L_2 的单词词嵌入表示。

2 双语主题词嵌入模型

基于双语词嵌入模型的跨语言文本分类是现阶段最为主流的方法，但是在词嵌入模型中，每一个单词用一个向量表示。当针对一词多义的单词进行向量表示时，这些单词虽然有多种词义，但是所有词义都统一表示为一个向量。单词在不同的上下文语境下有着不同的语义，而这些不同的语义都通过同一个向量表示，显然这种表示方式是不合理的，存在歧义性问题。

2.1 平行句对与词对齐信息

用于训练双语词嵌入模型的双语数据类型有词对齐、句子对齐和文档对齐三种对齐类型。关于跨语言词嵌入模型研究表明，双语平行语料对齐级别的选择，即模型所使用的数据类型，比双语词嵌入模型架构更为重要[19]。词对齐级别的数据的生成花销最大，句子对齐次之。在对现有的双语词嵌入方法中，对各个典型数据要求的词嵌入模型进行比对后发现，数据要求越高的双语词嵌入模型在跨语言文本分类上有着越好的效果，文本分类精度与训练词向量的数据的整理程度成正比[20]。使用句子对齐平行语料和平行句子对中的双语单词已进行词对齐的数据的双语 skip-gram 得到了最高的跨语言文本分类精度。

2.2 基于平行句对与词对齐信息的双语 skip-gram 模型

双语 skip-gram 模型训练依靠双语平行句对与词对齐信息。在平行句对中，假设 $\mathbf{S}=(s_1,s_2,\cdots,s_m)$ 为语言 L_1 的长度为 m 的句子，其

中 s_i 为句子 **S** 中的第 i 个单词，$\mathbf{T}=(t_1,t_2,\cdots,t_n)$ 为语言 L_2 的长度为 n 的句子，其中 t_j 为句子 **T** 中的第 j 个单词。在同一种语言中，需要通过 s_i 预测 s_i 的上下文，通过 t_j 预测 t_j 的上下文。不同语言中，如果 s_i 与 t_j 为词对齐关系，需要通过 s_i 预测 t_j 的上下文，通过 t_j 预测 s_i 的上下文。

通过对双语 skip-gram 模型的训练，可以获取到双语单词的词嵌入表示，把双语单词整合为一个词典，并假设 V 为双语单词词典，$|V|$为词典大小，DIM 为词向量的维数，并设 $\mathbf{Embedd},\ \mathbf{Embedd}\in\square^{|V|*DIM}$ 为双语单词词嵌入矩阵，其结构如下

$$\mathbf{Embedd}=[\mathbf{v}(w_1);\mathbf{v}(w_2);\cdots;\mathbf{v}(w_{|V|})] \quad (6)$$

2.3 自适应多原型向量表示

针对一词多义可能会引起的歧义性，Reisinger[21]，Huang[22]分别提出了多原型向量空间模型的概念，对单词进行聚类，并为每个簇生成不同的向量。然而在这些模型下，单词的表示往往随着簇的个数而增长，每个单词在每个簇下都有一个表示，增加了冗余度。

针对上述问题，需要引入“自适应”的多原型模型与词嵌入模型的结合方法，以实现单词的合理表示。LDA 主题模型对语料库整体进行建模，包含全局信息。语料库中的每篇文本由潜在主题的分布所表示，而每个潜在主题由单词分布表示。LDA 主题模型思想是每篇文本中有多个潜在主题，而文本中每个单词在不同的主题背景下出现的概率是不同的。不同于传统多原型模型，在 LDA 主题模型下，把单词视为对象时，单词与主题间的关系为一对一或者一对多的关系，单词在多原型表示上得到了自适应。在 LDA 主题模型下，单词可以属于一个主题，也可以属于多个主题，对应于语言学中的“单义词”与“多义词”。

词嵌入模型与主题模型结合的对单词的表示考虑到局部和全局特征，可以对单词表示“嵌入”局部与全局信息，可以解决一词多义引起的歧义性问题。

2.4 主题的嵌入表示

通过双语 LDA 主题模型，双语平行语料每个语言中每个文本的每个单词都会被分配到各自的主题。主题作为文本与单词的中间一级，每个双语主题可以被每一个语言的单词所解释。

词嵌入模型与主题模型相结合的单词表示考虑到局部和全局特征。为了对单词表示添加主题信息，需要得到主题在词嵌入空间的表示。双语 LDA 主题模型对平行分类语料的训练集进行建模后，可以获取到主题-词项矩阵 **NW**，其表示如下：

$$\mathbf{NW}=\begin{bmatrix} n_{11} & n_{12} & \cdots & n_{1|V|} \\ n_{21} & n_{22} & \cdots & n_{2|V|} \\ \cdots & \cdots & \cdots & \cdots \\ n_{K1} & n_{K2} & \cdots & n_{K|V|} \end{bmatrix} \quad (7)$$

其中$|V|$为语料的词典大小，K 为双语主题模型的主题数，$n_{ij}(i=1,2,\cdots,k;\ j=1,2,\cdots,|V|)$为匹配到第 i 个主题的第 j 个词的次数。

令 nw_i 为 **NW** 矩阵中第 i 行，即第 i 个主题在词项空间下的表示。通过如下公式获得第 i 个主题在词嵌入空间上的表示 $\mathbf{TopicEmbedd}_i$

$$\mathbf{TopicEmbedd}_i=\frac{\mathbf{nw}_i*\mathbf{Embedd}}{\|\mathbf{nw}_i\|_1} \quad (8)$$

其中 **Embedd** 为从式(6)中获取的单词的词嵌入矩阵，$\|\mathbf{nw}_i\|_1$为第 i 个主题中包含的单词数。LDA 主题模型中的主题作为由单词构成的语义集合，可以使用单词的表示对主题进行解释。式(8)通过加权平均主题中包含的单词的词嵌入表示，得到了双语主题的词嵌入空间的表示，代表了主题中包含的单词的通用语义。

2.5 单词的双语主题词嵌入表示

在主题模型中，语料中的每个单词会分配到不同的主题，而词典中的词项在语料中会分配到一个或多个主题。将自适应多原型特征结合词嵌入模型后，既考虑了单词之间的上下文，又考虑到了主题模型兼顾全局的多原型特征。如果 N 为预先定义的簇个数，$|V|$为词典大小，双语主题词嵌入模型的单词表示由双语词嵌入表示和双语主题嵌入表示的组合构成，通过组合双语词嵌入表示和双语主题嵌入表示来对单词进行描述，总共利用 $N+|V|$个嵌入表示描述语料中的单词。

使用双语主题词嵌入模型对分类语料的训练集和测试集进行建模以后，对于一个文本中的任意单词 w，若该单词在词典中的下标为 i，该单词所分配到的主题为第 j 个主题，则我们可以通过如下方式得到单词 w 的双语主题词嵌入 $\mathbf{v}(w)$。

$$\mathbf{v}(w)=(Embedd_i,TopicEmbedd_j)^{\mathrm{T}} \quad (9)$$

在构造语料中任意一个单词的双语的词嵌入时，可通过该单词在词典中的下标与该单词分配到的主题，在双语单词词嵌入矩阵与双语主题嵌入矩阵进行查找并进行拼接操作。

3 实验结果及分析

3.1 数据集

训练双语主题词嵌入模型语料包含训练双语词嵌入模型以及训练双语主题&分类器两类不同的平行语料数据，前者所需要的平行语料在规模上要远大于训练、测试分类器的语料，在涉及到的领域上训练双语词嵌入的语料要包含分类语料。

3.1.1 双语词嵌入

本文训练双语词嵌入模型使用 TED 平行语料，其内容来源于 TED 讲座视频的平行字幕数据，由中文、朝鲜语以及日语组成，本文使用其中中文与朝鲜语字幕平行句对。本文使用结巴分词系统对 TED 平行语料的中文部分进行分词与词性标注。针对朝鲜语部分，本文使用 Open Korean Text 朝鲜语分词系统进行分词与词性标注。本文使用筛选平行语料中的单词词性的方法来抽取语义等价单元，在分词系统的词性列表中剔除没有语义的词性,以此来决定语义等价单元的集合。在词对齐信息的获取上，使用 Fast_Align[23]词对齐工具在双语平行句对中抽取了单词间的对齐关系。

通过上述语料的获取、分词和筛选等步骤，最终通过 277747 对平行句抽取了 26694 个中文词项与 28025 个朝鲜语词项。

3.1.2 双语主题训练及跨语言文本分类

训练双语主题模型是面向分类语料进行的，对分类语料中的一词多义单词的词嵌入表示添加潜在主题层面的语义。本文利用爬虫技术对韩国“东亚日报”的中朝双语新闻进行爬取。通过人工筛选的方式选取了政治、经济、文化以及体育四类 4188 篇新闻，其中包含 1250 篇政治、995 篇经济、875 篇文化以及 1068 篇体育新闻。训练集与测试集按 8:2 比例进行分割。包含其中中文文本平均包含 126 个单词，朝鲜语文本平均包含 165 个单词。

3.2 双语单词的双语主题词嵌入表示

朝鲜语是表音文字，使用少量的字母记录语言中的语音，从而形成记录语言的文字，比如“yi”这个音，不同于汉语的“一”、“依”等多种写法，朝鲜语只有“이”这一种表示。汉字作为表意文字，大多数常规的中文单词是单义的，而朝鲜语普遍存在一个单词对应好几种词义。

表 1 与朝鲜语单词不同词义最相似的中文单词

朝鲜语单词	词典释义	所属主题	主题内最相似的中文单词
이상	理想	Topic1	理想，时间，能力，缺少，强大
	以上	Topic4	超过，以上，较，增加，数量
총	总	Topic4	超过，总，收入，计算，少
	枪	Topic6	射击，卫星，共计，军队，威胁
강도	强盗	Topic2	犯罪，监禁，温室，力量，腐败
	强度	Topic6	精确，代价，高强度，距离，造成
초	秒	Topic2	秒钟，后，刚刚，当时，之前
	初	Topic4	今年，最初，每周，月份，最近
공사	公社	Topic0	设施，海运，公社，研究，现代
	工事	Topic4	工程，费用，规划，领域，原料
	工事	Topic6	工程，海军，到，火箭，卫星
보고	报告	Topic4	演示，电视，公开，拥有，搜索
	看	Topic9	观看，看，想，说，听到
정상	正常	Topic3	普通，最终，稳定，经济，结果
	顶上	Topic5	活跃，成功，美国，个人，顶峰
삼성	三星	Topic1	三星队，攻破，大胜，连败，女篮
	三星	Topic4	三星，东芝，物产，占有率，电子

以主题数为 10 的情况为例，通过双语主题词嵌入模型，对双语语料中的单词进行词嵌入表示以后，根据双语单词间的相似度进行排序，并返回与朝鲜语单词最相似的 5 个中文单词，其结果如表 1 所示。从表可以看出，朝鲜语单词的每个词义，都有潜在主题所表示，并且在主题中的单词语义与词典释义接近。另外，“삼성”对应的翻译“三星”也分配到了两个主题，其中 Topic1 的内容与体育相关，而 Topic4 的内容与经济相关。由此可以发现，双语主题词嵌入模型不仅可

以解释传统的一词多义，而且也对不同领域的同一个单词也可以有多个词嵌入表示。

3.3 跨语言文本分类实验方案

3.3.1 双语文本表示

本文实验中所用到的双语表示方法如下：

1）双语主题模型（Bi-LDA）

在本实验中，双语主题模型的主题数 K 分别取 10、20、40、60、80 和 100，先验参数取值为：α=50/K, β=0.01，对训练集进行 1000 次迭代，对测试集进行 100 次的迭代。

2）双语词嵌入模型（Bi-skipgram）

本实验训练 128、256 和 512 三种维数的词向量，预测中心词前后各 2 个上下文单词；每个样例采样 5 个负样例，每个批次计算 50 个句子的损失后更新参数，对语料整体进行 20 次遍历（epoch），初始化学习率为 0.02,每训练 10000 句后对学习率进行调整。

3）双语主题词嵌入模型（Bi-TWE）

首先通过双语词嵌入模型获取 128、256 以及 512 维的词嵌入表示；之后对分类语料进行双语主题模型的建模，主题数 K 取 10、20、40、60、80 和 100，获得双语主题在两种语言单词上的表示后，用双语词嵌入表示构建主题表示。最后根据分类语料中每个单词以及分配到的主题构建该单词的双语主题词嵌入表示。

3.3.2 文本分类算法

传统文本分类算法中，本实验使用感知机算法的优化版本-多分类平均感知机算法，以 0.05 的学习率进行 20 次的迭代训练多分类平均感知机分类器。同时将双语主题模型结合朴素贝叶斯分类器进行跨语言文本分类，对于零概率问题，进行了拉普拉斯平滑。并且支持向量机的核函数选定为多项式内核。

在深度学习文本分类算法中，在本文实验使用了 TextCNN[24]与 HAN[25]算法。在 TextCNN 算法中，根据经验使用高度为 3、4 和 5 的三种卷积核窗口，每种高度卷积窗数量选取 100 个，以 multi-channel 多通道的方式训练。Pooling 方式选择 1-Max Pooling ，Dropout 选择 0.5，采用 0.02 的学习率，使用 Adadelta 优化方法，每批训练 50 个样例共对语料整体进行 200 次遍历（epoch）。HAN 层级注意力网络中学习率采用 0.02，使用随机梯度下降法，momentum 动量值取 0.9，以每批 50 个训练样例，共训练 100 个 epoch。

3.4 实验结果及分析

本文跨语言文本分类实验将分类平行语料以 8:2 的比例进行训练集与测试集的分割，进行了 3 组分类实验:1）用中文训练集训练分类器，对朝鲜语测试集进行测试，表示为 cn2kr；2）用朝鲜语训练集训练分类器，对中文测试集进行测试，表示为 kr2cn；3）用中文与朝鲜语训练集一同训练分类器，对中文与朝鲜语测试集进行测试，表示为 all2all

本文提出的双语主题词嵌入模型结合深度学习分类算法的分类精度与对比实验结果如表 2 所示，其中第 1 列为文本表示模型，第 2 列为分类算法，第 3 列为词嵌入表示的维数，第 4 列为主题个数，第 5 列到第 7 列为跨语言文本分类的实验方案。

表 2 中朝跨语言文本分类准确率(%)（1/2）

表示模型	分类算法	*DIM*	*K*	cn2kr	kr2cn	all2all
Bi-LDA	NB	None	10	83.17	84.48	83.94
			20	84.72	85.44	85.14
			40	84.00	86.87	85.38
			60	84.84	87.23	85.91
			80	85.44	87.58	86.21
			100	84.00	87.58	86.03
Bi-skip gram	Per-cep-tron	128	None	47.73	33.41	60.08
		256		54.05	41.64	61.27
		512		58.47	48.68	62.47
Bi-skip gram	SVM	128	None	70.88	86.51	87.41
		256		85.79	87.30	90.03
		512		86.99	88.26	90.27
Bi-skip gram	HAN	128	None	89.26	89.85	90.81
		256		89.97	89.73	90.57
		512		89.37	89.85	90.39
Bi-skip gram	Text-CNN	128	None	88.06	85.44	88.72
		256		90.33	88.18	90.75
		512		89.61	88.30	90.81
Our Model	HAN	128	10	89.73	89.85	90.93
			20	90.21	89.97	90.45
			40	89.49	89.61	90.21
			60	87.82	89.02	90.33
			80	88.90	90.09	90.39
			100	88.54	89.49	89.91
		256	10	89.61	89.14	90.39
			20	89.49	89.73	90.69
			40	90.33	88.90	90.45
			60	89.97	90.21	90.39
			80	89.97	90.21	90.75
			100	89.49	89.85	90.69

表 2 中朝跨语言文本分类准确率(%)（2/2）

表示模型	分类算法	*DIM*	*K*	cn2kr	kr2cn	all2all
Our Model	HAN	512	10	88.90	89.26	90.63
			20	89.61	89.49	90.75
			40	89.73	89.85	90.99
			60	88.06	89.85	90.51
			80	88.66	**90.57**	91.10
			100	89.02	90.21	91.28
Our Model	Text-CNN	128	10	88.90	89.26	90.39
			20	89.61	88.41	90.75
			40	80.21	88.54	91.10
			60	90.21	88.78	91.10
			80	90.57	89.26	90.93
			100	89.26	89.26	90.75
		256	10	90.09	88.90	91.52
			20	90.93	88.66	91.10
			40	90.21	89.21	91.34
			60	90.93	89.73	91.64
			80	90.93	89.21	91.22
			100	90.33	89.37	90.69
		512	10	**91.16**	88.54	91.34
			20	89.73	88.18	90.81
			40	89.49	88.42	90.87
			60	90.33	88.90	91.34
			80	90.33	88.90	**91.76**
			100	90.21	88.90	91.10

在利用中文训练分类器，并对朝鲜文进行的测试中，双语主题词嵌入模型结合 TextCNN 取得了最好的结果，准确度达到了 90.57%。在利用朝鲜文训练分类器，并对中文进行的测试中，双语主题词嵌入模型结合 HAN 取得了最好的结果，准确度达到了 91.16%。在利用中文与朝鲜文一同训练分类器，并对中文与朝鲜语进行的测试中，双语主题词嵌入模型结合 TextCNN 取得了最好的结果，准确度达到了 91.76%。

本实验的最高准确率也在以 512 为词嵌入表示维数的双语主题词嵌入模型中出现。这说明较高维数的词嵌入表示能表达更完善的语义信息。

在双语主题词嵌入结合两种深度学习算法的跨语言分类实验中，当主题数 *K* 为 80 时，最高准确率出现次数最多，60 次之。在 10~100 的主题中，主题数过少会使主题之间的边界模糊，主题数过大会细分潜在的主题概念，都无法描述某一类的潜在的概念。

比较同一维数的双语词嵌入表示下的双语 skip-gram 模型与双语主题词嵌入模型，如 128 维的双语 skip-gram 模型与 128 维词嵌入表示和各主题数下的双语主题词嵌入模型。在同一种深度学习算法下，可以发现最高准确率出现在双语主题词嵌入模型为文本表示的跨语言文本分类实验中。说明附加的潜在主题信息，在训练分类器的过程中，很好地得到了利用。

在训练数据的规模上 all2all 分类方案在数据量上是其他两种方案的两倍，中文训练集与朝鲜文训练集不同的数据分布可能会降低分类精度，但是分类器通过更大的数据量，捕捉到更多的语义，提高了分类准确率。该分类方案实则多语种文本分类，其分类效果验证了在具备多语种平行语料的前提下，本文中提出的文本表示方法与分类方案能够有效解决多语种分类问题。

4 结论及下一步工作

本文将双语词嵌入模型与 LDA 主题模型进行结合，提出了双语主题词嵌入模型。将 LDA 主题模型应用于解决一词多义所引起的歧义性问题。双语主题词嵌入模型作为自适应的多原型向量表示方法，可以解决单词的一词多义引起的单词嵌入表示上的歧义性。在中朝跨语言文本分类任务中，精确度最高达到了 91.76%，结果表明跨语言文本分类方案达到了实际应用水平。并且通过选取各种主题数 *K* 与词嵌入维数，得到了不同的跨语言文本分类实验结果，从中分析了各种参数对分类准确率的影响，总结出了最佳参数设定方法。

受限于语料库的语种，现阶段双语主题词嵌入模型只能应用于中朝跨语言文本分类。下一步工作中将通过构建包含更多语种的语料库进行中英、朝英等更广泛的跨语言文本分类研究。而且双语主题词嵌入模型无法直接确定潜在主题数 *K*，要在一定范围采样典型的主题数进行实验验证。针对此问题，在下一步的工作中，需要研究确定主题数 *K* 的方法，以提高双语主题词嵌入模型的训练效率。

参考文献

[1] Bel N ,Koster C H A,Villegas M. Cross-Lingual Text Categorization[J]. Lecture Notes in Computer Science. 2003, 18(2769): 126~139

[2] Rigutini L,Maggini M,Liu B. An EM Based Training Algorithm for Cross-Language Text Categorization[C]//The Proceedings IEEE/WIC/ACM International Conference on Web Intelligence. Compiègne University

of Technology, France, 2005: 529~535
[3] Ni X,Sun J,Hu J,Zheng C. Mining Multilingual Topics from Wikipedia[C]//WWW. Madrid, Spain, 2009: 1155~1156
[4] Heyman G,Vulić I,Moens M F. C-BiLDA Extracting Cross-Lingual Topics from Non-ParallelTexts by Distinguishing Shared from Unshared Content[J]. Data Mining & Knowledge Discovery. 2016, 30(5): 1299~1323
[5] Luong T,Pham H,Manning C D. Bilingual Word Representations with Monolingual Quality in Mind[C]//Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Denver, USA, 2015: 151~159
[6] Faruqui M,Dyer C. Improving Vector Space Word Representations Using Multilingual Correlation[C]//Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics. Gothenburg, Sweden, 2014: 462~471
[7] Hermann K M,Blunsom P. Multilingual Distributed Representations without Word Alignment[J/OL]. ArXiv Preprint ArXiv:1312.6173, 2013.
[8] Hermann K M,Blunsom P. Multilingual Models for Compositional Distributed Semantics[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore, USA, 2014: 58~68
[9] Gouws S,Bengio Y,Corrado G. Bilbowa: Fast Bilingual Distributed Representations without Word Alignments[C]//Proceedings of the 32nd International Conference on Machine Learning. Lille, France, 2015: 748~756
[10] Vulić I,Moens M F. Bilingual Word Embeddings from Non-Parallel Document-Aligned Data Applied to Bilingual Lexicon Induction[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Beijing, China, 2015: 719~725
[11] 周国强，崔荣一. 基于朴素贝叶斯分类器的朝鲜语文本分类的研究[J]. 中文信息学报. 2011, 25(04): 16~19
[12] Lee S. Korean Document Classification Using Extended Vector Space Model[J]. Kips Transactions Partb. 2011, 18B(2): 93~108
[13] Blei D M,Ng A Y,Jordan M I. Latent Dirichlet Allocation[J]. Journal of Machine Learning Research. 2003, 3(Jan): 993~1022
[14] 徐谦，周俊生，陈家骏. Dirichlet 过程及其在自然语言处理中的应用[J]. 中文信息学报. 2009, 23(5): 25~32,46
[15] 徐戈，王厚峰. 自然语言处理中主题模型的发展[J]. 计算机学报. 2011, 34(8): 1423~1436
[16] Yerebakan H Z,Dundar M. Partially Collapsed Parallel Gibbs Sampler for Dirichlet Process Mixture Models[J]. Pattern Recognition Letters, 2017, 90(): 22~27
[17] Mikolov T,Chen K,Corrado G,etal. Efficient Estimation of Word Representations in Vector Space. Computer Science[J/OL]. ArXiv Preprint ArXiv: 1301.3781, 2013
[18] Mikolov T,Chen K,Corrado G,etal. Distributed Representations of Words and Phrases and their Compositionality[C]//Proceedings of the 26th International Conference on Neural Information Processing Systems. Lake Tahoe, USA, 2013: 3111~3119
[19] Ruder S,Vulic I,Sogaard A. A Survey of Cross-Lingual Word Embedding Models[J/OL]. ArXiv Preprint ArXiv: 1706.04902, 2017
[20] Upadhyay S,Faruqui M,Dyer C,etal. Cross-lingual Models of Word Embeddings: An Empirical Comparison[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany, 2016: 1661~1670
[21] Reisinger J,Mooney R J. Multi-Prototype Vector-Space Models of Word Meaning[C]//Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL. Los Angeles, USA, 2010: 109~117
[22] Huang E H,Socher R,Manning C D,Ng A Y. Improving Word Representations via Global Context and Multiple Word Prototypes[C]//Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. Jeju, Republic of Korea, 2012: 873~882
[23] Dyer C,Chahuneau V,Smith N A. A Simple, Fast, and Effective Reparameterization of IBM Model 2[C]//Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Atlanta, USA, 2013: 644~648
[24] Kim Y. Convolutional Neural Networks for Sentence Classification[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar, 2014: 1746~1751
[25] Yang Z,Yang D,Dyer C,etal. Hierarchical Attention Networks for Document Classification[C]//Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, USA, 2016: 1480~1489

王琪（1996—），硕士，主要研究领域为自然语言处理。
E-mail：1094376724@qq.com

田明杰（1990—），硕士，主要研究领域为自然语言处理。
E-mail：834876787@qq.com

崔荣一（1962—），博士，教授，主要研究领域为智能计算，模式识别，机器学习，自然语言处理。
E-mail：cuirongyi@ybu.edu.cn