计算机科学
COMPUTER SCIENCE

基于Kernel-XGBoost的跨语言术语对齐方法

于娟, 张晨

引用本文

于娟, 张晨. 基于Kernel-XGBoost的跨语言术语对齐方法[J]. 计算机科学, 2022, 49(11A): 211000111-6.

YU Juan, ZHANG Chen. Cross-lingual Term Alignment with Kernel-XGBoost[J]. Computer Science, 2022, 49(11A): 211000111-6.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.211000111

基于 Kernel-XGBoost 的跨语言术语对齐方法

于 娟 张 晨
福州大学经济与管理学院 福州 350108
(zhangchenfzu@163.com)

摘 要 跨语言术语对齐是跨语言文本数据分析与知识发现的关键基础。针对跨语言术语对齐研究多为单词术语对齐且严重依赖向量空间对齐的现状,提出一种能够实现跨语言单词及多词术语间一对多对齐的 Kernel-XGBoost 方法。给定跨语言平行语料库,该方法分两步得到同义的跨语言术语对:1)跨语言术语提取与候选术语对生成;2)基于跨语言词嵌入的术语对齐。汉语-西班牙语以及汉语-法语的术语对齐实验表明,该方法在 Top-5 的准确率可达到 80%,能有效支持跨语言信息检索、本体构建等跨语言文本数据挖掘任务。

关键词:跨语言;文本分析;术语对齐;Kernel-XGBoost;汉语;法语;西班牙语
中图法分类号 G202;TP391.1

Cross-lingual Term Alignment with Kernel-XGBoost

YU Juan and ZHANG Chen
School of Economics and Management,Fuzhou University,Fuzhou 350108,China

Abstract Cross-lingual term alignment is a crucial step for cross-lingual text data analysis and knowledge discovery. Current research usually focuses on single-word term alignment and relies heavily on vector space alignment. Therefore,a new Kernel-XGBoost method is proposed for the one-to-many alignment of cross-lingual terms including multi-word terms. Given a cross-lingual parallel corpus,the proposed method obtains synonymous cross-lingual terms in two steps:1) extracting cross-lingual terms and generating candidate term pairs;2) aligning cross-lingual terms based on word embedding. Experiments on Chinese-Spanish and Chinese-French term alignments demonstrate that the proposed method can achieve an accuracy of 80% at Top-5. It can effectively support cross-lingual text mining tasks such as information retrieval,ontology building.

Keywords Cross-lingual,Text analysis,Term alignment,Kernel-XGBoost,Chinese,French,Spanish

1 引言

随着“一带一路”倡议的推进和全球化进程的加快,迅速采集全球信息资源并整合形成全局视图的能力成为跨国组织决策成败的关键。其中,国际新闻、会议、各国际分公司经济发展报告等跨语言文本是跨国组织进行管理决策的主要信息资源。由于跨语言文本数据的体量越来越大,跨语言文本挖掘与分析方法的应用前景日益明朗,研究价值逐渐凸显。采用跨语言文本挖掘方法,可以从不同语言的文本数据中抽取出事先未知的、可理解的、有意义的知识和模式,进而支持跨国组织的信息整合,助力组织管理决策的有效制定。

跨语言术语对齐是跨语言文本分析与知识发现的基础任务之一,是将不同语言中的同义术语进行映射的过程,是跨语言信息检索、本体构建等的关键环节。本文研究跨语言术语对齐的 Kernel-XGBoost 方法。该方法首先从跨语言语料库中提取并生成候选术语对,然后基于 Kernel-XGBoost 方法对齐术语。

本文第 2 节为跨语言术语对齐的研究现状;第 3 节介绍本文方法的框架流程;第 4 节和第 5 节分别详细介绍跨语言候选术语对生成的方法和术语对齐的方法;第 6 节是实验与结果分析;最后总结全文。

2 研究现状

目前,跨语言术语对齐方法多针对单词术语的对齐,少有多词术语对齐,也少有跨语系术语对齐的方法研究。已有的跨语言术语对齐方法研究可分为主要的两类:基于统计的方法和基于深度学习的方法。

在术语对齐研究的早期,多策略相融合的统计对齐方法最受欢迎,该方法需要人工归纳术语翻译对的特征,转化为多个统计学公式计算术语的翻译概率。Sun 等[1]利用规则提取术语,使用术语的长度特征及汉英术语的共现特征计算术语的翻译概率;几年后,Zhang 等[2]对该方法进行了改进,通过 CRF 模型提取汉英术语,在之前的翻译概率中加入了术语的语序位置特征,实验结果显示,与基线方法相比,该方法对于词频排名靠后的术语,对齐的准确率提升较高;文献[3-4]使用双语术语在语料库中的共现特征计算术语的翻译概率;Liu 等[5]利用首尾词性规则提取术语,通过结合独立、停用、语义、首尾及词性相似度、Giza++词对齐工具得到的 g 值对齐术语,在召回率为 1%的前提下准确率可达 96%,远高于 Dice 系数、Giza++等方法得到的术语对。由于同源语言在单词

基金项目:国家自然科学基金(71771054)
This work was supported by the National Natural Science Foundation of China(71771054).
通信作者:张晨(zhangchenfzu@163.com)

的拼写上较为相似，可以使用编辑距离计算术语的相似程度。Gamallo[6]构建了西葡双语词典，通过计算分布相似性，找到源语言词语的翻译候选词，然后计算候选词与源词的编辑距离，找出最相似的作为翻译词语。以上方法虽然思想简单且准确率较高，但操作过程复杂，需人工总结术语翻译对的特征，制定多个统计学公式，且在低频词对齐上的准确率不高。

词语向量化表示是自然语言处理基础且关键的步骤，随着以神经网络为首的机器学习方法研究热潮的来临，单语向量化及跨语言词嵌入(Cross-Language Word Embeddings，CLWE)工作取得了重要的进展，这也促进了双语词典构建、双语术语对齐等任务的发展。Sanjanasri 等[7]利用 CNN 来提取翻译特征，模型根据源语言单词对之间的相似关系进行训练，给定相应的目标语言单词对之间的余弦相似度作为标签。该网络的核心目标集中在微调以前学到的翻译模式，由于以任意两个词语作为输入，以余弦相似度作为输出，因此计算复杂度十分高，且该模型仅适用于单词术语。Mohiuddin 等[8]认为即使对于密切相关的语言，同构假设也不成立，因此不应该使用线性变换来学习映射。LNMAP 方法首先将嵌入变换到潜在空间中，然后使用非线性变换来学习映射，还加入了反向翻译和原始嵌入重建的约束。由高资源和低资源语言组成的 15 种不同语言对进行的大量实验显示了非线性转换的有效性，尤其是对于低资源和远距离语言。在信息检索领域，也存在很多模型用于查询与文档的匹配度计算，如基于核的神经排序模型[9](Kernel based Neural Ranking Model，K-NRM)，该模型首先应用在单语信息检索领域。2020 年 Yu 等[10]将其用于双语信息检索问题，取得了不错的效果。Conv-KNRM 是 Dai 等对 K-NRM 的改进[11]，在原模型的基础上增加了 CNN 交互层，作用是将检索及文档的上下词信息进行交互，实验表明其在单语信息检索任务上的表现优于 K-NRM 模型。综上，基于词向量的对齐方法需将词语向量化，通过分类器直接对齐术语或学习映射矩阵后在同一空间计算双语术语的相似度。该方法简单，但准确率严重依赖于向量空间对齐的质量[12]。

综上可见，目前跨语系语言间的术语对齐研究成果较少，而已有方法的准确率不令人满意，严重依赖于向量空间对齐的准确性，且缺乏面向单词及多词术语对齐的统一框架。为此，本文研究一种不依赖向量空间对齐的 Kernel-XGBoost 术语对齐方法，实现跨语系的单词和多词术语间的一对多对齐。

3 本文方法框架

基于 Kernel-XGBoost 实现跨语言术语对齐，主要分为两步：候选术语对生成和跨语言术语对齐。前者对输入语料进行文本预处理和术语提取，并生成候选的跨语言术语对；后者将源术语及目标术语进行向量化，然后计算术语的匹配程度，使用 Kernel 对匹配结果进行特征提取，最终通过分类器得到跨语言术语对的匹配程度。本文方法的流程如图 1 所示。

(1)候选术语对生成模块包含 3 个子模块。其中，文本预处理与词语提取子模块采用原子词步长法[13]处理预处理后的文本，提取其中出现的单词及多词词语；术语度计算子模块衡量提取得到的词语是术语的程度；跨语言术语对生成子模块采用基于共句原则的方法，降低术语对生成的时间复杂度，从而降低后续术语对齐概率计算的时间复杂度。后文在第 4 节对该模块进行详细介绍。

(2)跨语言术语对齐模块包含 4 个子模块。其中，跨语言术语向量化子模块，使用由 Facebook 提供的预对齐向量；跨语言术语交互计算子模块，计算双语术语的余弦相似度，构成对齐矩阵；Kernel 相似度分层子模块，利用高斯核将交互后的矩阵做分层处理，从而为分类器提供更准确的特征；XGBoost 术语匹配度计算子模块，将匹配问题转化为分类任务，利用 XGBoost 模型计算双语术语的匹配程度。后文第 5 节详细介绍跨语言术语对齐方法。

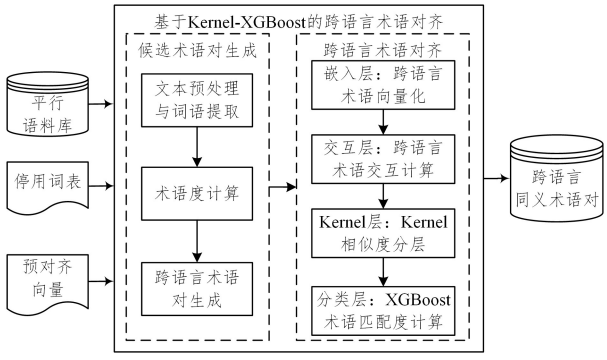

图 1 基于 Kernel-XGBoost 的跨语言术语对齐方法流程图

Fig. 1 Flow chart of cross-language term alignment method based on Kernel-XGBoost

4 候选术语对生成

候选术语对生成是术语对齐的基础，首先进行文本预处理，采用原子词步长法提取各语言文本中出现的单词及多词词语；接着分别计算每一词语的术语度，过滤不满足术语度阈值的词语，形成多语言术语集合；然后基于共句原则生成候选术语对，以大幅降低后续计算术语对齐概率的时间复杂度。

4.1 文本预处理与词语提取

文本预处理包括文本清洗、分词或词形还原、词性标注、停用词及停用词性删除。文本预处理子模块的输入是文本，输出是被停用词、停用词性、标点符号截断的词串序列，为保留句子的断点信息，删除的停用词、停用词性、标点符号均采用换行符“\n”代替。

本文采用原子词步长法提取词语，该方法由频繁词串提取与子串过滤两步组成。词语提取子模块的输入是文本预处理生成的词串序列，输出是候选词语。具体的词语提取方法请见文献[11，14-15]。

4.2 术语度计算

为了判断自动提取得到的词语是否领域专有术语，需要计算词语的术语度(termhood)。本文借鉴并改进了文献[16]提出的领域相关度(Domain Relevance，DR)和领域均匀度(Domain Consensus，DC)算式，通过统计领域术语在领域文档和非领域文档中的分布情况来计算词语的术语度。本文的术语度计算方法需要使用前景语料库和背景语料库。其中，前景语料库是指包含丰富领域术语的文档集合；背景语料库是区别于非领域文档的集合。

DR 衡量词语与领域相关的程度，需结合前景语料库和背景语料库来计算。其计算式如下：

$$DR=\log\left(\frac{freq_{Cf}(t)/N}{(freq_{Cb}(t)+1\times10^{-8})/M}\right) \tag{1}$$

其中，Cf 指前景语料库，Cb 指背景语料库；N 是前景语料库的总文本数，M 是背景语料库的总文本数；$freq_{Cf}(t)$ 是指词

语 t 在前景语料库中出现的总频次，$freq_{Cb}(t)$ 是指词语 t 在背景语料库中出现的总频次，为避免除 0 错误，以 1×10^{-8} 作为平滑因子。

DC 衡量词语在领域文本中分布的均匀程度，仅需结合前景语料库来计算。算式如下：

$$DC=\sum_{Cf_j\in Cf}\left[P(t|Cf_j)\times\log\frac{1}{P(t|Cf_j)}\right] \tag{2}$$

$$P(t|Cf_j)=\frac{freq_{Cf_j}(t)}{freq_{Cf}(t)} \tag{3}$$

其中，Cf_j 是指前景语料库中的文本。

当由式(1)计算得到 $DR>0$ 时，说明词语 t 在前景语料库中出现的频次高于在背景语料库中出现的频次，则 t 是领域术语的概率较大。当由式(2)计算得到 $DC=0$ 时，说明词语 t 仅出现在前景语料库中的一个文本里，即没能在领域语料中均匀分布，几乎不可能是领域术语。因此，过滤掉 $DR\leqslant0$ 及 $DC=0$ 的词语，将剩下的词语作为术语。

4.3　跨语言术语对生成

对于术语度计算子模块输出的双语术语集合，为了降低计算复杂度，本文先基于共句原则生成候选的跨语言术语对，即构建源术语的候选目标术语。若有源术语 m 个，目标术语 n 个，则常用的术语对齐方法的计算时间复杂度一般为 $O(m\times n)$。为了降低时间开销，本文仅将共现于对齐句对中的术语作为候选术语对。即，若有 p 个对齐句对，共现在每个句对中的源术语平均有 a 个，目标术语平均有 b 个，则计算的时间复杂度为 $O(p\times a\times b)$，其中，a 和 b 一般为较小的个位数。

为明晰起见，以两对汉西句子为例，解释基于共句原则的术语对生成方法。图 2 给出了两句汉西平行句子对，仅供举例使用，不具特殊性，其中西语已做词形还原处理。图 3 给出了图 2 句子经原子词步长法及术语度计算后提取出的术语。图 4 给出了图 3 术语经共句原则生成的部分跨语言候选术语对。

句子对一：
价格波动和金融化对进口国和出口国的投资决定都有影响。
la volatilidad de el precio y la financiarización influir en la decisión de inversión tanto en el país importador como en el exportador.
句子对二：
价值链获益与直接外资之间的关系是可以辨明的，后者被视为成功生产流程的一个滞后变量。
ser posible distinguir la relación entre el beneficio de la cadena de valor y la IED，ya que se reconocer que este último ser un variable rezagar con respecto a el proceso de producción eficiente.

图 2　汉西平行句对示例

Fig. 2　Example of Chinese-Western parallel sentence pairs

句子对一：
价格；价格波动；金融；金融化；进口；进口国；出口；出口国；投资；影响。
volatilidad（波动）；precio（价格）；volatilidad de el precio（价格波动）；financiarización influir（金融化影响）；influir en la decisión（影响决定）；decisión de inversión（投资决策）；influir（影响）；financiarización（金融化）；inversión(投资)；país importador（进口国）；país（国家）；exportador（出口国）。
句子对二：
价值；价值链；获益；外资；关系；辨明；视为；成功；生产流程；滞后变量。
distinguir（辨明）；beneficio de la cadena de valor（价值链获益）；cadena de valor（价值链）；beneficio（获益）；valor（价值）；IED（直接外资）；último（后者）；producción eficiente（成功生产）；eficiente（成功）。

图 3　经原子词步长法及术语度计算后提取出的术语

Fig. 3　Terminology extracted after atomic word step method and terminology calculation

句子对一：	句子对二：
价格- volatilidad；	价值-distinguir；
价格-precio；	价值-beneficio de la cadena de valor；
价格-volatilidad de el precio；	价值-cadena de valor；
价格-financiarización influir；	价值-beneficio；
价格-influir en la decisión；	价值-valor；
价格-decisión de inversión；	价值-IED；
价格-influir；	价值-último；
价格-financiarización；	价值-producción eficiente；
价格-inversión；	价值-eficiente；
价格-país importador；	价值链-distinguir；
价格-país；	价值链-beneficio de la cadena de valor；
价格-exportador；	价值链-cadena de valor；
价格波动-volatilidad；	价值链-beneficio；
价格波动-precio；	价值链-valor；
价格波动-volatilidad de el precio；	价值链-IED；
价格波动-financiarización influir；	价值链-último；
价格波动-influir en la decisión；	价值链-producción eficiente；
价格波动-decisión de inversión；	价值链-eficiente；
价格波动-influir…	获益-distinguir…

图 4　部分跨语言候选术语对示例

Fig. 4　Some examples of cross-language candidate term pairs

5　跨语言术语对齐

目前，跨语言单词嵌入的研究取得了重要的进展[17-18]。CLWE 是将两个单独训练的词向量在同一空间中对齐，其中源语言中的词语接近其在源语言中的同义词和在目标语言中的翻译，反之亦然，从而可以在同一空间中计算双语词语的相似性，而 CLWE 的发展能够为双语术语对齐任务提供强有力的支持。

本文在 CLWEs 的基础上，使用 Kernel 层对源术语及其对应的候选目标术语的相似度进行分层，用 XGBoost 分类器计算术语的匹配度。由于候选术语对中的术语包含单词和多词术语，因此需要首先统一术语的长度，且能够在向量对齐质量不高的同时进一步提高术语对齐的准确度。本文的跨语言术语对齐模型如图 5 所示。

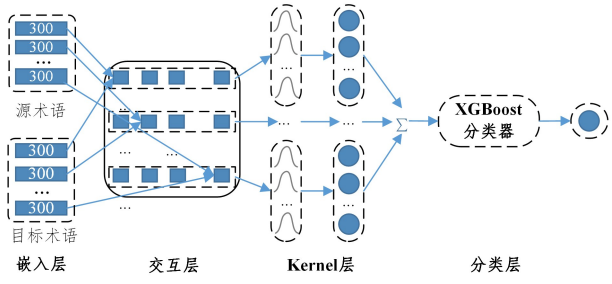

图 5　跨语言术语对齐流程图

Fig. 5　Cross-language term alignment flowchart

5.1　嵌入层

嵌入层的输入是已分词或直接根据空格切分的源术语及目标术语，输出是由词向量组成的嵌入矩阵。本文的术语向量使用由 Facebook 提供的跨语言预对齐向量 fastText[19]。

本文在进行跨语言术语对齐时，同时处理单词术语和多词术语，即术语不等长，所以需要先统一术语的长度。为此，将不满足固定长度的术语自动补 0 向量。若术语长度统一为 8 个单词，而源术语 s 由 n 个单词组成，表示为 $\{w_1^s,w_2^s,\cdots,w_n^s\}$，目标术语 t 由 m 个单词组成，表示为 $\{w_1^t,w_2^t,\cdots,w_m^t\}$，则向量化后的源、目标术语分别转换为 $\{\overrightarrow{w_1^s}^T,\overrightarrow{w_2^s}^T,\cdots,\overrightarrow{w_8^s}\}$，$\{\overrightarrow{w_1^t},\overrightarrow{w_2^t},\cdots,\overrightarrow{w_8^t}\}$。其中，$\overrightarrow{w_1^s}$ 为源术语中单词 w_1^s 对应的 300 维 fastText 预对齐词向量。

5.2 交互层

交互层的输入是双语词向量，输出是双语对齐矩阵，矩阵中的每个元素为源术语中词语与目标术语中词语的余弦相似度。算式如下：

$$M_{ij}=\frac{\overrightarrow{w_i^s}\cdot\overrightarrow{w_j^t}}{\|\overrightarrow{w_i^s}\|\cdot\|\overrightarrow{w_j^t}\|} \tag{4}$$

其中，$\overrightarrow{w_i^s}$为源术语中第 i 个单词对应的向量，$\overrightarrow{w_j^t}$为目标术语中第 j 个单词对应的向量，$\|A\|$ 表示向量的模运算。

5.3 Kernel 层

Kernel 层的输入是对齐矩阵，输出是一个 k 维向量。Kernel 层对源术语及其对应的候选目标术语的相似度进行分层，通过高斯核将相似度分为 k 个级别，从中捕捉模糊对齐的特征，从而在对齐向量质量不高的情况下为分类器提供更可靠的特征。算式如下：

$$K_k(M_i)=\sum_j\exp\left(-\frac{(M_{ij}-\mu_k)^2}{2\sigma_k^2}\right) \tag{5}$$

$$\overrightarrow{K}(M_i)=\{K_1(M_i),K_2(M_i),\cdots,K_k(M_i)\} \tag{6}$$

$$\varphi(M)=\sum_{i=1}^{n}\log\overrightarrow{K}(M_i) \tag{7}$$

式(5)中，μ_k 为第 k 个高斯核的均值，用来控制相似度分层的界限；σ_k 为第 k 个高斯核的方差，控制高斯核的局部作用范围。

5.4 XGBoost 分类层

XGBoost 分类层的输入是 k 维向量，输出是跨语言术语对齐的概率。为了强化分类效果，本文将信息检索模型 K-NRM[9]的分类层改为基于决策树的有监督分类器 XGBoost[20]，采用前向分布算法进行贪婪学习。训练完成得到多棵树结构，要预测新样本的分数时，根据该样本的特征，走遍每棵树结构，将样本落在相应的叶结点上，最终累加每棵树对应的叶结点分值即可得到该样本的预测结果。

6 实验分析

为验证本文方法在跨语系术语对齐任务上的有效性，特选取语言距离较大的汉西及汉法两种语言进行实验。由于目前还没有检验汉西术语对齐及汉法术语对齐的专用语料库，也没有统一、标准的评价指标，本文选取联合国平行语料库作为本文的数据集[21]，并用 $P@N$ 作为衡量术语对齐质量的评价指标。

6.1 数据集

对于汉西术语对齐实验，本文的前景语料库使用联合国语料库 2012－2014 年贸易和发展会议的 200 篇汉西平行文本，包含平行句子 42984 对，共计 11.28 MB；背景语料库选取联合国语料库 2012－2014 年外层空间、环境会议的 477 篇汉西平行文本，包含平行句子 92663 对，共计 24.5 MB。

对于汉法术语对齐实验，本文的前景语料库使用联合国语料库 2012－2014 年贸易和发展会议的 200 篇汉法平行文本，包含平行句子 41999 对，共计 11.22 MB；背景语料库选取联合国语料库 2012－2014 年外层空间、环境会议的 477 篇汉西平行文本，包含平行句子 90693 对，共计 24.5 MB。

6.2 评价指标

$P@N$ 是信息检索领域常用的评价指标[8,22]。由于本文将术语对齐任务转化为查找源术语的最佳目标术语问题，因此可以借鉴信息检索的评价指标 $P@N$。计算如下式所示：

$$P@N(t)=\frac{\sum_{i=1}^{Z}\|align(t)\cap d(t)\|}{Z} \tag{8}$$

$$\|align(t)\cap d(t)\|=\begin{cases}1, & \text{预测值与实际结果存在交集}\\ 0, & \text{其他}\end{cases} \tag{9}$$

其中，$align(t)$ 表示本文方法预测的源术语 t 在目标术语 TOP-N 上的对齐；$d(t)$ 表示源术语 t 在文本中对应的真实目标术语；$\|align(t)\cap d(t)\|$ 表示预测到的目标术语与实际目标术语是否存在交集，如果有交集则值为 1，否则值为 0；Z 表示源术语的总数。

6.3 实验与分析

汉西术语对齐实验中，首先采用原子词步长法提取汉、西词语，得到汉语词语 26544 条，西语词语 40237 条；接着计算词语的术语度，过滤掉低于阈值的词语，剩余汉语术语 23719 条，西语术语 31352 条，汉、西术语具体长度分布情况如图 6、图 7 所示；然后基于共句原则生成汉西候选术语对，共计 3728463 对。

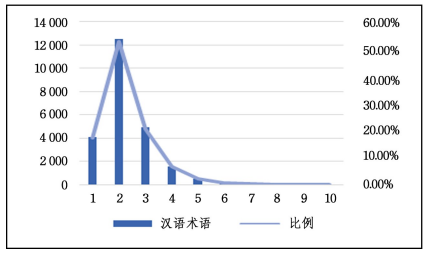

图 6 汉语术语长度分布

Fig. 6 Length distribution of Chinese terms

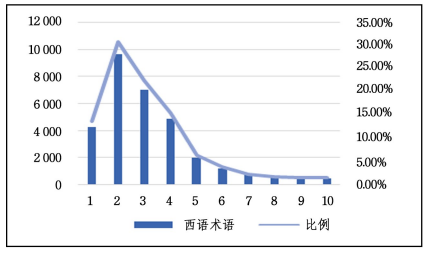

图 7 西语术语长度分布

Fig. 7 Length distribution of Spanish terms

汉法术语对齐实验中，采用原子词步长法提取汉、法词语，得到汉语词语 26544 条，法语词语 43544 条；经术语度计算，剩余汉语术语 23719 条，法语术语 34269 条，然后基于共句原则生成汉法候选术语对，共计 3888590 对。由于法术语长度分布情况与汉语及西语相似，此处不再详细列举法语的分布情况。

在汉西术语对齐模块，本文采用的训练集为团队之前在路透社语料库中提取出的 10053 条汉西术语对，为每个正样本随机分配 5 个负样本；为使验证集与测试集的分布保持一致，在提取出的 3728463 对候选术语对中，忽略掉因分词、词形还原、语料库的意译、增译、删译、词语提取等方面带来的错误，这些错误具体体现在：在西语术语中无法找到与汉语术语对应的翻译，随机选取 1320 个汉语术语及其对应的候选西语术语作为验证集；由于候选术语对数量多，人工高质量标记术语对工作量巨大，因此在提取出的 3728463 对候选术语对中，再随机选取 1320 个汉语术语（与验证集无交集）及其对应

的候选西语术语作为测试集。

在汉法术语对齐模块，本文的训练集为研究团队在欧洲议会议事录平行文本语料库提取的 5000 条汉法术语对，同样为每个正样本随机分配 5 个负样本，在提取出的 3888590 对候选术语对中选取 625 个汉语术语及其对应的候选法语术语作为验证集，再随机选取 625 个汉语术语及其对应的法语术语作为测试集。基于谷歌翻译、汉西-西汉词典及汉法-法汉词典对验证集及测试集进行人工标记。

为验证本文提出的改进方法与 K-NRM 方法的差别，实现 K-NRM 模型，并实现术语对齐任务常用的余弦相似度方法，以验证两种方法对向量空间对齐质量的依赖程度，在汉西术语对齐任务中，对于 Kernel-XGBoost 模型，设置学习率为 0.1，决策树的最大深度为 7，决策树为 170 棵，学习目标设置为“rank：pairwise”；对于 K-NRM 模型，设置优化器为 Adadelta，迭代次数为 100 次，批大小设置为 32，学习率为 0.001，损失函数设置为三元组损失。在汉法术语对齐任务中，对于 Kernel-XGBoost 模型，设置学习率为 0.01，决策树的最大深度为 4，决策树为 100 棵，学习目标设置为“rank：pairwise”；对于 K-NRM 模型，设置优化器为 Adadelt，迭代次数为 100 次，批大小设置为 32，学习率为 0.001，损失函数为三元组损失。所有的模型，均设置 4 个 Kenel，术语长度固定为 8。以上所有参数均在验证集调整得到。3 种方法在汉西术语对齐任务上的比较结果如图 8 所示，在汉法术语对齐任务上的比较结果如图 9 所示。

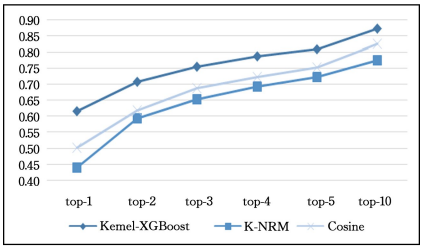

图 8　汉西术语对齐指标 $P@N$ 的比较

Fig. 8　Comparison of Chinese-Spanish term alignment index $P@N$

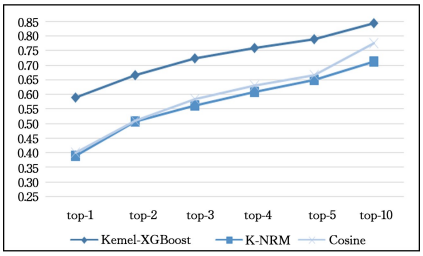

图 9　汉法术语对齐指标 $P@N$ 的比较

Fig. 9　Comparison of Chinese-French term alignment index $P@N$

由实验可知：

(1)使用一个小型汉西单词词典对汉西预对齐 fastText 向量的对齐准确率进行测试，结果只有 35%，而汉西预对齐 fastText 向量在术语对齐中的准确率高达 50%；汉法预对齐 fastText 向量的对齐准确率为 34%，而汉法预对齐 fastText 向量在术语对齐中的准确率为 40%，这是由于本文提取出的术语大多为多词术语，在计算汉西、汉法术语相似度时多个词语的余弦相似度可以互相支持。

(2)本文提出的 Kernel-XGBoost 模型在跨语言术语对齐实验中表现最优，其次是依赖于术语对齐向量的余弦相似度方法。由 K-NRM 与 Kernel-XGBoost 模型的比较实验可知，全连接层构成的分类器性能不如 XGBoost 分类器的表现好。

(3)本文方法在汉西和汉法术语对齐实验中都取得了不错的效果。汉西术语对齐任务使用了大规模训练集，而汉法对齐任务仅使用了一个由 5000 词对组成的小型训练集，但在对齐准确度上，与余弦相似度相比，汉法对齐结果的提升程度要高于汉西对齐结果，因此本文方法不会过分受限于训练集的规模。

此外，对 Kernel-XGBoost 模型的结果进行定量分析可发现，本文方法可以为源术语找到多个合适的目标术语，即实现了一对多对齐。表 1 列出了部分汉西术语对齐的结果。

表 1　汉西术语对齐结果示例

Table 1　Examples of Chinese-Spanish term alignment results

汉语术语	西语术语	对齐概率	标签
双边贸易协定	acuerdo comercial bilateral	0.990419791	1
双边贸易协定	acuerdo bilateral	0.934938778	0
双边贸易协定	acuerdo comercial	0.910885671	0
双边贸易协定	bilateral establecer	0.910885671	0
双边贸易协定	acuerdo	0.78372043	0
经济治理	gobernanza económico	0.998737501	1
经济治理	gobernanza económico mundial	0.956281859	0
经济治理	crecimiento económico	0.952277016	0
经济治理	crisis económico	0.945731399	0
经济治理	politica macroeconómicas	0.927332616	0
物流管理	gestión logística	0.998737501	1
物流管理	materia de gestión	0.998089288	0
物流管理	logística portuario	0.973585021	0
物流管理	gestión portuario	0.956281859	0
物流管理	transporteintermodal	0.927332616	0
宏观经济变量	variable macroeconómicas	0.993127213	1
宏观经济变量	macroeconómicas	0.78372043	0
宏观经济变量	crisis financiero	0.648079064	0
宏观经济变量	fluctuación	0.561499306	0
宏观经济变量	variable	0.555135961	0
创新机制	mecanismo innovador	0.998737501	1
创新机制	mecanismo innovadores	0.998737501	1
创新机制	crear mecanismo	0.952277016	0
创新机制	mecanismo	0.945103941	0
创新机制	innovador	0.822523327	0

综上所述，本文提出的 Kernel-XGBoost 模型受益于特征提取器 Kernel 及分类器 XGBoost，Kernel 层通过对相似度的分层处理来捕获模糊对齐信息，XGBoost 层提供了强大的分类功能，在跨语言向量空间对齐质量不高的情况下，进一步提高了术语对齐的质量；可同时对齐单词及多词术语，并实现术语间的一对多对齐。当然，由于 XGBoost 模型的存在，本文方法所需调节的参数比 K-NRM 模型多，但在较高的准确率及调参工具的支持下，该缺点是可以接受的。

结束语　目前，作为跨语言文本分析与知识发现基础任务之一的术语对齐，在跨语系语言间的研究成果较少，现有方法的准确率尚不令人满意，且缺乏面向单词及多词术语对齐的统一框架。为支持跨国组织的信息整合，助力国际化战略管理决策的有效制定，本文提出了一种基于 Kernel-XGBoost 的跨语言术语对齐方法。该方法首先从语料库中提取跨语言术语并生成候选术语对，然后基于跨语言词嵌入进行术语对齐。汉西、汉法术语对齐的实验表明，本文方法的准确率显著高于经典的余弦相似度、K-NRM 等方法。

本文提出的 Kernel-XGBoost 模型是一个跨语言术语对

齐的框架，从本质上讲是一种相似度计算方法。该模型使用Kernel提取术语的模糊对齐特征，采用分类器XGBoost学习对齐术语对的特征，因此也适用于单词、短语、句子、篇章级别的相似度计算，如跨语言信息检索中检索词与待检索文档之间的匹配度计算。

本文方法能够实现跨语言单词和多词术语的一对多对齐，在提高准确率的同时，无需人工归纳总结术语翻译对的特征，不严重依赖于向量空间对齐的质量，不过分受限于训练集规模，能够为跨语言信息检索、本体构建、大粒度对齐等跨语言文本分析任务提供支持。另一方面，由于本文方法在平行语料库中分别提取术语，因此会受到语料库对齐中的增译、省译、删译等翻译现象的影响。

参 考 文 献

[1] SUN L, JIN Y B, DU L, et al. Automatic extraction of bilingual term dictionaries from parallel corpus[J]. Journal of Chinese Information Processing, 2000, 14(6): 33-39.

[2] ZHANG L, LIU Y X. Research on Automatic Extraction of Chinese-English Term Pairs Based on Word Order Position Features[J]. Journal of Nanjing University(Natural Science), 2015(4): 707-713.

[3] LI X Y. Research on term alignment based on bilingual parallel corpus of historical classics[D]. Dalian: Dalian University of Technology, 2010.

[4] ZENG W, WANG H L, XU H J. Research on Automatic Construction Technology of Chinese-English Bilingual Dictionary [J]. Journal of Information, 2011, 30(4): 402-409.

[5] LIU S Q, ZHU D H. Term alignment method based on multi-strategy fusion Giza++[J]. Journal of Software, 2015, 26(7): 1650-1661.

[6] GAMALLO P. Strategies for Building High Quality Bilingual Lexicons from Comparable Corpora[J]. Parallel Corpora for Contrastive and Translation Studies: New resources and applications, 2019, 90: 251.

[7] SANJANASRI J P, MENON V K, SOMAN K P. BUCC2020: Bilingual Dictionary Induction using Cross-lingual Embedding [C]// Proceedings of the 13th Workshop on Building and Using Comparable Corpora. 2020: 65-68.

[8] MOHIUDDIN T, BARI M S, JOTY S. Lnmap: Departures from isomorphic assumption in bilingual lexicon induction through non-linear mapping in latent space[J]. arXiv: 2004. 13889, 2020.

[9] XIONG C, DAI Z, CALLAN J, et al. End-to-end neural ad-hoc ranking with kernel pooling[C]// Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2017: 55-64.

[10] YU P, ALLAN J. A Study of Neural Matching Models for Cross-lingual IR[C] // Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2020: 1637-1640.

[11] DAI Z, XIONG C, CALLAN J, et al. Convolutional neural networks for soft-matching n-grams in ad-hoc search[C]// Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining. 2018: 126-134.

[12] ARTETXE M, LABAKA G, AGIRRE E. Bilingual lexicon induction through unsupervised machine translation[J]. arXiv: 1907. 10761, 2019.

[13] YU J, DANG Y Z. Word extraction method combining part-of-speech analysis and string frequency statistics[J]. Systems Engineering Theory and Practice, 2010, 30(1): 105-111.

[14] YU J, WU X P, LIAO X, et al. Extracting Terms form French Corpora with FP Sequence Tree[J]. Journal of University of Electronic Science and Technology of China, 2021, 50(1): 84-90.

[15] YU J, YAN Y L, JIAN Z W, et al. Extracting Terms from Spanish Corpora Based on DC-Value[J]. Computer Systems & Applications, 2021, 30(6): 271-277.

[16] YU J, DANG Y Z. Research on Extraction Method of Domain Feature Words[J]. Journal of Information, 2009(3): 368-373.

[17] GOIKOETXEA J, SOROA A, AGIRRE E. Bilingual Embeddings with Random Walks over Multilingual Wordnets[J]. Knowledge-Based Systems, 2018, 150(JUN. 15): 218-230.

[18] GLAVA G, VULI I. Non-Linear Instance-Based Cross-Lingual Mapping for Non-Isomorphic Embedding Spaces[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020.

[19] JOULIN A, BOJANOWSKI P, MIKOLOV T, et al. Loss in Translation: Learning Bilingual Word Mapping with a Retrieval Criterion[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018.

[20] CHEN T, GUESTRIN C. Xgboost: A scalable tree boosting system[C]// Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining. 2016: 785-794.

[21] ZIEMSKI M, JUNCZYS-DOWMUNT M, POULIQUEN B. The united nations parallel corpus v1. 0[C]// Proceedings of the Tenth International Conference on Language Resources and Evaluation(LREC'16). 2016: 3530-3534.

[22] CRASWELL N, LIU L, ZSU M T. Precision at n[M]// Encyclopedia of Database Systems. Boston: Springer, 2009: 2127-2128.

YU Juan, born in 1981, professor, Ph.D supervisor. Her main research interests include data science and knowledge engineering, intelligent information system.

ZHANG Chen, born in 1997, postgraduate. Her main research interests include cross-language text analysis and knowledge discovery.