基于深度学习的维吾尔语名词短语指代消解

李 敏[1] 禹 龙[2] 田生伟[1] 吐尔根 · 依布拉音[3] 赵建国[4]

摘 要 针对维吾尔语名词短语指代现象, 提出了一种利用栈式自编码深度学习算法进行基于语义特征的指代消解方法. 通过对维吾尔语名词短语指称性的研究, 提取出利于消解任务的 13 项特征. 为提高特征对文本语义的表达, 在特征集中引入富含词汇语义及上下文位置关系的 Word embedding. 利用深度学习机制无监督的提取隐含的深层语义特征, 训练 Softmax 分类器进而完成指代消解任务. 该方法在维吾尔语指代消解任务中的准确率为 74.5 %, 召回率为 70.6 %, F 值为 72.4 %. 实验结果证明, 深度学习模型较浅层的支持向量机更合适于本文的指代消解任务, 对 Word embedding 特征项的引入, 有效地提高了指代消解模型的性能.

关键词 深度学习, 栈式自编码神经网络, 指代消解, Word embedding, 维吾尔语

引用格式 李敏, 禹龙, 田生伟, 吐尔根 · 依布拉音, 赵建国. 基于深度学习的维吾尔语名词短语指代消解. 自动化学报, 2017, **43**(11): 1984−1992

DOI 10.16383/j.aas.2017.c160330

Coreference Resolution of Uyghur Noun Phrases Based on Deep Learning

LI Min[1] YU Long[2] TIAN Sheng-Wei[1] TurgLm IBRAHIM[3] ZHAO Jian-Guo[4]

Abstract Aimed at the reference phenomena of Uyghur noun phrases, a method using stacked autoencoder model to achieve coreference resolution based on semantic characteristics is presented. Through the study of noun phrases referentiality, we pick up beneficial 13 features for coreference resolution tasks. In order to improve the expression of features for semantic text, Word embedding is added into feature sets, which makes feature sets contain lexical semantic information and context positional relationship. A deep learning algorithm is proposed for unsupervised detection of implicit semantic information, and also introduced is a softmax classifier to decide whether the two markables actually corefer. Experiments show that precision rate, recall rate and F value of coreference resolution reach 74.5 %, 70.6 % and 72.4 %, respectively, which demonstrates that the proposed method on coreference resolution of Uyghur noun phrase and introduction of Word embedding to feature sets are able to improve the performance of coreference resolution system.

Key words Deep learning, stacked autoencoder, coreference resolution, word embedding, Uyghur

Citation Li Min, Yu Long, Tian Sheng-Wei, TurgLm IBRAHIM, Zhao Jian-Guo. Coreference resolution of Uyghur noun phrases based on deep learning. *Acta Automatica Sinica*, 2017, **43**(11): 1984−1992

指代 (Anaphora) 是自然语言的普遍现象, 也是语篇内句与句之间衔接 (Cohesion) 的重要手段之一, 对指代成分准确无歧义的消解有助于机器分析和语篇理解[1]. 随着篇章处理相关应用的日益发展, 指代消解在自动文摘、信息抽取、机器翻译等自然语言处理领域有着广泛的运用.

基于机器学习的方法在指代消解任务中得到广泛的应用. Soon 等[2] 提出一种基于机器学习的指代消解方法, 利用决策树算法消解非限制领域名词短语, 首次给出了完整的实现步骤. Bergsma 等[3] 利用支持向量机 (Support vector machine, SVM) 结合统计信息进行指代消解, 提出一种基于语法路径的代词消解方法. Ng[4] 通过对不同语义类之间指代关系的研究, 提出一种自动推导语义类扩充特征的方法, 提升了指代消解的性能. Bengtson 等[5] 将指代消解视为图式问题, 根据最佳链路决策算法生成指代图, 利用机器学习技术研究不同类型特征对指代消解任务的贡献, 强调距离特征和同位结构特征在消解任务中的重要作用.

与国际上指代消解的研究相比, 国内的指代消解研究起步较晚, 周俊生等[6] 利用了一种无监督的聚类算法来实现名词短语的指代消解, 采用带权图对指代消解问题进行建模, 引入有效模块函数实现

收稿日期 2016-04-12 录用日期 2016-08-02
Manuscript received April 12, 2016; accepted August 2, 2016
国家自然科学基金 (61563051, 61262064, 61662074, 61331011), 自治区科技人才培养项目 (QN2016YX0051) 资助
Supported by National Natural Science Foundation of China (61563051, 61262064, 61662074, 61331011) and Regional Scientific and Technological Personnel Training Project (QN2016YX0051)
本文责任编委 张民
Recommended by Associate Editor ZHANG Min
1. 新疆大学软件学院 乌鲁木齐 830008 2. 新疆大学网络中心 乌鲁木齐 830046 3. 新疆大学信息科学与工程学院 乌鲁木齐 830046 4. 新疆大学人文学院 乌鲁木齐 830046
1. School of Software, Xinjiang University, Urumqi 830008 2. Network Center, Xinjiang University, Urumqi 830046 3. College of Information Science and Technology, Xinjiang University, Urumqi 830046 4. College of Humanities, Xinjiang University, Urumqi 830046

指代等价类的自动划分. 王海东等[7] 探索了语义角色对指代消解性能的影响, 其研究表明, 语义角色信息的引入能够显著提高指代消解的性能. 孔芳等[8] 提出了一种基于卷积树核函数的中英文代词消解方法, 利用卷积树核函数, 探索语义信息对代词消解性能的作用. 奚雪峰等[9] 提出采用深度学习的深层机器学习代替传统浅层机器学习, 通过构建多隐层机器学习模型, 自动挖掘深层语义信息并探索面向指代消解的语义特征泛化表示. 实验结果表明, 多层的学习结构有效地提高了模型的性能, 特征抽象层的引入, 进一步促进模型对深层语义信息的学习.

随着指代消解研究的不断深入, 如何有效地获取语义信息在指代消解中起至关重要的作用. 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, 受到广泛的关注[10−11]. 本文提出利用 Word embedding 之间的位置关系表示语义层面上的联系, 作为文本的高层抽象特征, 通过深度学习模型多映射单元提取主要的结构信息, 发现变量之间的真正关系形式, 进而完成指代消解任务. 目前, 许多学者从不同的角度探讨了各种语言的指代现象, 但对于像维吾尔语这样的小语种研究还不够深入. 因此本文针对维吾尔语的名词短语指代现象进行研究, 主要研究代词和名词/名词短语、名词和名词/名词短语、名词短语和名词短语间的指代问题.

1 相关研究

深度学习是机器学习研究中的一个新的领域, 最早于 2006 年 Hinton 等提出, Hinton 指出, 深层神经网络 (Deep neural network, DNN) 具有更好的特征学习能力, 无监督式逐层预训练其抽象后的数学特征对原始数据有更本质的刻画. 随着深度学习方法在图像和语音领域的应用, 越来越多的深度学习方法被用于自然语言处理. Glorot 等[12] 利用深度自编码算法完成文本分类任务, 通过添加纠正激活函数, 有效地提高了分类效果. Glorot 等[13] 在文献 [12] 的基础上, 提取出评论的高层抽象特征, 解决了跨领域文本分类的问题. Lu 等[14] 将深度自编码算法运用到机器翻译任务中, 为基于词汇翻译模型提取到了有效地特征集, 并在中英文翻译过程中取得了很好的效果. 由此可见, 深度自编码算法具有很强的无监督学习特征的能力, 能较好地提取文本中隐含特征.

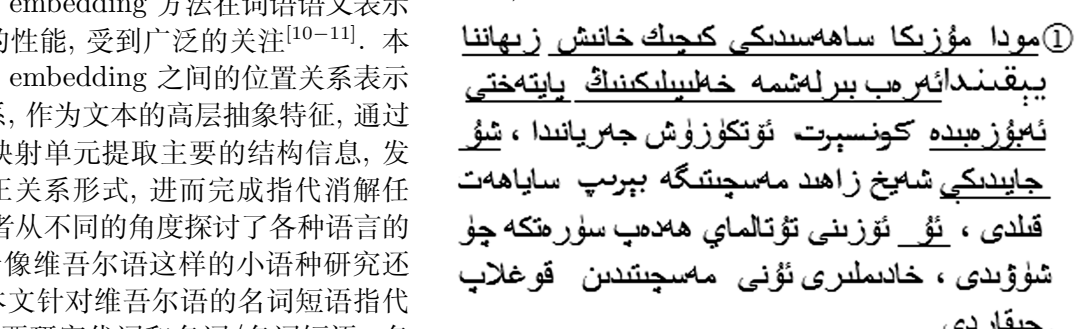
本文利用深度栈式自编码算法无监督学习文本特征, 通过构建多隐层机器学习模型, 逐层变换特征, 将样本在原空间的特征表示变换到一个新特征空间, 获取数据的分布式特征表示, 并自动挖掘深层语义, 从而使分类更加容易.

2 预备知识

为了便于读者理解本文对维吾尔语名词短语指代消解分析研究的方法, 先明确以下定义:

定义 1. 指代消解: 指自然语篇中的一个语言单位 (通常是词或者短语) 确定其指向之前出现的语言的单位的过程[15]. 用于指向的语言单位, 称为照应语 (Anaphora), 被指向的语言单位称为先行语 (Antecedent).

本文采用消息理解会议 (Message understanding conference, MUC) 对指代的定义, 认为指代关系不仅在于代词和名词/名词短语间, 还存在于名词或者名词短语间, 例如 (维吾尔语的书写规范是从右向左):



①مودا مۇزىكا ساھەسىدىكى كىچىك خانىش زېھاننا يېقىنداىئەرەب بىرلەشمە خەلىيلىكىنىڭ پايتەختى ئەبۇزەبىدە كونسېرت ئوتكۈزۈش جەريانىدا ، شۇ جايىدىكى شەيخ زاھىد مەسچىتىگە بېرىپ ساياھەت قىلدى ، ئۇ ئۇزىنى تۇتالماي ھەدەپ سۈرەتكە چۈشۈردى ، خادىملىرى ئۇنى مەسچىتتىدن قوغلاپ چىقاردى.

(流行音乐小皇后蕾哈娜, 最近在 阿联酋首都阿布扎比 举办演唱会过程中去了 当地 的谢赫扎伊德清真寺游览, 她 大肆拍照, 遭到了工作人员的轰赶)

例 ① 摘自实验语料, 反映了多种指代关系, 包括名词和代词的指代关系: “زېھاننا (蕾哈娜)” 和 “ئۇ (她)”; 名词短语和代词的指代关系 “مودا مۇزىكا ساھەسىدىكى كىچىك خانىش (流行乐界小皇后)” 和 “ئۇ (她)”; 名词短语和名词的指代关系: “شۇ جايىدىكى (当地)” 和 “ئەبۇزەبىدە (阿布扎比)”; 名词短语间: “ئەرەب بىرلەشمە خەلىيلىكىنىڭ پايتەختى ئەبۇزەبىدە (阿联酋首都阿布扎比)” 和 “شۇ جايىدىكى (当地)”; 只有准确地找到这些指代关系所指的实体, 才能准确地完成指代消解任务, 从而提高系统的性能.

定义 2. 指称: 语篇中指代成分 (照应语) 与所指对象 (先行语) 之间的相互解释关系. 它旨在帮助说话者向听者指出正在被谈论的对象[16]. 指称性取决于说话者特定使用一个表达式的意图, 是一种潜在的语义特征.

在维吾尔语语篇中, 并不是所有的名词短语都具有指称功能[16], 因此辨别名词短语的指称性是指代消解研究中的一个重要任务.

经过深入分析和研究, 结合维吾尔语具体的语言特点, 在参考相关文献后, 实验组维吾尔语语言学专家总结出维吾尔语中具有指称性的名词短语:

1) 专有名词

表示某一特有事物名称的词. 在维吾尔语语篇中, 表示人、地点和机构的专有名词都有一个特有的指称对象, 例如:

②شىنجاڭ ئاپتونۇم رايونى 1955- يىلى قۇرۇلغان، ئۇنىڭ كان بايلىقلىرى ناھايىتى مول.

(新疆维吾尔自治区 成立于 1955 年, 它 的矿产资源丰富)

2) 带领属性人称词尾的名词短语

领属性人称词尾表示一种归属关系, 一般附加在名词之后, 用来明确某一实体的归属性问题. 例如:

③مېنىڭ ئاكام ئاجايىپ خۇش خەت يېزىش ۋە رەسىم سىزىش ماھارىتىگە ئىگە ئىدى . ئۇ خۇراپىي ئادەتلەرنى زاڭلىق قىلىدىغان ھەجۋىي رەسىملەرنى سىزگەن.

(我的哥哥 擅长书法和绘画, 他 画过一幅嘲讽迷信活动的漫画)

在例③中 مېنىڭ ئاكام(我的哥哥) 带有领属性人称词尾 م , 与下文 ئۇ (他) 具有指称关系.

3) 被指示词修饰的名词短语

指示词用来指示或区别事物, 被指示词修饰的名词短语通常作为先行语的解释扩充, 目的是为了帮助读者或者受话人更好地理解所传达的信息. 例如:

④بۇ يىل ئاشلىقتىن مول ھوسۇل ئالالمىدۇق ، مەن بايا سۇلايمان بوغارتىرلار بىلەن بۇ مەسىلىنى پاراڭلاشتىم.

(今年我们没有夺得粮食丰收, 这个问题 我刚才和苏莱曼会计等人已经聊过了)

在例④中 بۇ مەسىلىنى (这个问题) 指代 بۇ يىل ئاشلىقتىن مول ھوسۇل ئالالمىدۇق (今年我们没有夺得粮食丰收). 指示代词 بۇ (这个) 修饰名词 مەسىلىنى (问题), 做照应语.

4) 被形容词或形容词化的成分修饰的名词短语

形容词或形容词化的成分充当限定词对名词短语做了限定, 具有确定性, 一般可作为先行语或者照应语. 例如:

⑤پاكار بويلۇق مەزمۇت گەۋدىلىك ئىسىركىرەپ كەلدى ، ئۇ جاراڭلىق ئاۋاز بىلەن سۆز باشلىدى.

(个头矮小、身体壮实的伊米尔 进来了, 他 开始高声说话)

5) 带宾格标志的名词短语

带宾格的名词短语通常所表达的事物已经被确定或者已被确定将要出现. 例如:

⑥مەن بىر كىتابنى تاپتىم ، ئۇ ئۆيدە.

(我找到了 一本书, 它 在家里)

在例⑥中, كىتابنى(书) 带宾格后缀 "نى", 与下文中 "ئۇ (它)" 具有指代关系.

3 基于栈式自编码的指代消解模型

针对维吾尔语名词短语指代消解问题本文参考 Soon 等[2] 提出的指代消解基本框架. 首先, 确定先行语和照应语对应的候选项, 构建名词短语特征向量; 接着生成训练实例、训练分类器; 最后, 识别测试文档, 构建指代链, 完成消解任务. 本文采用文献 [9] 一致的方式进行实例的初步生成, 引入 Word embedding 作为特征项, 利用栈式自编码非线性逐层贪心算法无监督地学习深层语义特征, 用学习到的特征训练 Softmax 分类器, 进而实现名词短语的指代消解. 整个消解过程如图 1 所示.

3.1 特征提取

提取的特征是否有效对本文提出的栈式自编码指代消解模型有直接的影响, 使用准确的特征对文本进行描述, 实验效果会相应提高. 结合实验组维吾尔语语言学专家总结的具有指称性的名词短语, 选取以下特征进行指代消解:

1) AnProperNoun: 若照应语是专有名词, 该特征值为 1; 否则, 为 0; 本文对维吾尔语专有名词限定为人名、地名、机构名.

2) CaProperNoun: 若先行语是专有名词, 该特征值为 1; 否则, 为 0;

3) AnDefiniteNP: 若照应语是被形容词或形容词化成分修饰的名词短语, 该特征值为 1; 否则, 为 0. 在语料标注过程中, 这类名词短语的词性标注为 AdjNP, 特征提取时判断, 若词性为 AdjNP, 该特征值为 1; 否则, 为 0.

4) CaDefiniteNP: 若先行语是被形容词或形容词化成分修饰的名词短语, 该特征值为 1; 否则, 为 0.

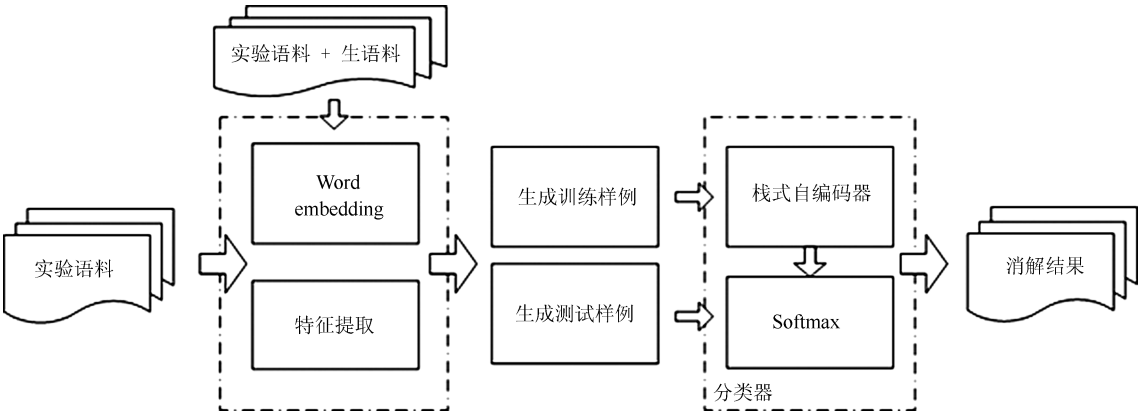


图 1　维吾尔语名词短语指代消解框架

Fig.1　The coreference resolution of Uyghur noun phrases framework

5) AnDemonstrativeNP: 若照应语被指示词修饰的名词短语, 则该特征值为 1; 否则, 为 0. 实验组维吾尔语语言学专家制定了一个指示词库, 提取特征时扫描词库, 当名词短语包含指示词时, 该特征值为 1; 否则, 为 0. 指示词库如表 1 所示 (此处仅列举出部分指代词).

表 1　指示词库

Table 1　The demonstrative thesaurus

指人指物	指性质	指数量	指地点
بۇ/这个	بۇنداق/这样	بۇنچە/这么	مۇشـە/这儿
مانا بۇ/这个	ۇنـداقم/这样	مـۇنچە/这么	مەشە/这儿
ئۇ/那个	شۇنداق/那样	شۇنچە/那么	ئايە /那儿
ئاۋۇ/那个	ئۇنداق/那样	شۇنچۋالا/那么	ئاۋۇيەر/那儿
...	...	...	...

6) CaDemonstrativeNP: 若先行语被指示词修饰的名词短语, 则该特征值为 1; 否则, 为 0.

7) AnPossessionNP: 若照应语是带领属性人称词尾的名词短语, 该特征值为 1; 否则, 为 0; 维吾尔语中属格有两种形式: 加后缀 "نىڭ" 和附加人称词尾. 语料标注过程中对名词的格范畴进行了标注, 特征提取是判断, 若格范畴为领属性人称词尾, 该特征值为 1; 否则, 为 0.

8) CaPossessionNP: 若先行语是带领属性人称词尾的名词短语, 该特征值为 1; 否则, 为 0.

9) AnObjectiveNP: 若照应语是带宾格标志的名词短语, 该特征值为 1; 否则, 为 0; 带宾格标志的名词短语, 通常在名词短语后附加 "نى" 后缀, 判断名词短语是否包含后缀成分, 包含该特征值为 1; 否则, 为 0.

10) CaObjectiveNP: 若先行语是带宾格标志的名词短语, 该特征值为 1; 否则, 为 0.

11) PropertyFit: 若照应语和先行语满足词性一致, 该特征值为 1; 否则, 为 0.

12) SinglePluralFit: 若照应语和先行语满足单复数一致, 该特征值为 1; 否则, 为 0.

13) FullMatch: 若照应语和先行语去掉 "بۇ (这)" "ئۇ (那)" 等修饰性的词后满足全匹配, 该特征值为 1; 否则, 为 0.

3.2　Word Embedding

不同于传统的词袋模型, 基于神经网络训练的 Word embedding 包含丰富的上下文信息, 可以很好地表现目标词在当前文本中的语义规则, 同时也避免了维数灾难[17]. 本文使用 Mikolov 等[10] 提出的 Word2vec 工具进行 Word embedding 的训练, 选择 Skip-gram+HS 模型作为训练框架, 通过计算词之间的余弦相似度表示文本语义上的相关程度. 为了更准确地获取每个词在低维空间中语义的分布情况, 本文在原有实验语料的基础上进行了扩充. 选取天山网、人民网等维吾尔语版网页作为语料来源, 利用网络爬虫下载网页, 进行去重、去噪处理之后获取约 7 000 篇不限题材且未标注任何内容的生语料.

同语料训练的词向量每一维表达的隐含语义相似[17], 本文在词向量的基础上, 通过将短语中的词语向量融合, 构建实验需要的包含隐含语义信息的短语向量 Phrase embedding. 如例 ⑤ 中ئىسمر گەۋدىلىك مەزمۇت بويلۇق پاكار (个头矮小、身体壮实的伊米尔) 的句向量表示: C(S) = (C(پاكار) + C(بويلۇق) + C(مەزمۇت) + C(گەۋدىلىك) + C(ئىسمر))/5.

3.3　训练样例和测试样例的生成

3.3.1　构建训练实例

本文首先对实验语料中的维吾尔语名词短语 (这里的名词短语包括代词、单个名词、名词短语)

进行提取, 并按一定的规则两两组队, 通过名词短语与标注语料中指代链的对应关系, 判断组队的名词短语是否具有指代. 若两个名词短语之间存在指代关系, 即为正例; 不存在则为负例. 具体的算法步骤如下:

步骤 1. 提取语料中的名词短语, 存入集合 $\{NounPhraseSet\}$, 提出语料中指代链信息, 存入集合 $\{ChainNum, ChainContent\}$.

步骤 2. 对于每一个名词短语判断其是否满足 $NP_{i\geq 2} \in \{ChainContent\}$, 若满足继续步骤 3, 不满足存入集合 $\{NagativeSet\}$, $i++$.

步骤 3. NP_i $(i \geq 2)$ 从后向前依次与 NP_{i-1} 至 NP_0 间的每个名词短语组队, 组队结果存入集合 $\{WordsGroup\}$, 其中 $NP_{i-k}NP_i \in \{WordsGroup\}$

步骤 4. 判断 NP_{i-k} NP_i 单复数是否冲突, 不冲突继续步骤 5, 冲突则删除 $NP_{i-k}NP_i$, $k++$, 更新 $\{WordsGroup\}$.

步骤 5. 判断 NP_i 与 NP_{i-k} 是否满足 $NP_{i-k} \in \{ChainContent\}$ 且 $NP_{i-k}.ChainNum = NP_i.ChainNum$. 满足继续步骤 6, 不满足 $k++$.

步骤 6. 将 NP_i 与 NP_{i-k} 组成的词语对构成正例, 存入集合 $\{+instance\}$. 将 NP_i 与 NP_j $(i > j > i-k)$ 且 $NP_j \in \{NagativeSet\}$ 组成的词语队构成负例, 存入集合 $\{-instance\}$, $k++$.

步骤 7. 重复步骤 4～步骤 6, 当 $k = i$ 时, $i++$, 执行步骤 2. 当 $i > NounPhraseSet.length$ 时停止.

3.3.2 构建测试实例

测试实例的构建过程与训练实例过程类似, 由于指代链信息未知, 因此对提取出的每个名词短语都作为照应语依次向前组队, 判断组队中的名词短语单复数是否冲突, 将不冲突的组队供给训练时生成的分类器模型进行分类. 若分类器根据生成的分类模型判断该实例为正列, 则认为这两个名词短语之间具有指代关系, 组队过程结束. 否则, 继续向前与前一个名词短语组队, 直到整篇文档的第一个名词短语为止.

构建训练和测试实例后, 根据第 3.1 节和第 3.2 节给出的文本特征向量和 Phrase embedding 依次获得每一实例对应的特征向量取值, 生成训练样例和测试样例. 以例句⑦为例:

⑦هۆسەيىن نۇۋزامانسىنىڭ كاتتا بىلىم ئىگىلىرىدىن بىرى بولغاچقا ، بۇراخان ئۇنى قەدىرلەيتتى.

(因为 吾斯英 是 当代的大学者之一, 所以 布葛热汗 尊敬 他)

提取出的名词短语包括: هۆسەيىن (吾斯英), نۇۋزامانسىنىڭ كاتتا بىلىم ئىگىلىرىدىن بىرى (当代大学者之一), نۇۋزامانسىنىڭ كاتتا بىلىم ئىگىلىرىدىن (当代大学者), كاتتا بىلىم ئىگىلىرىدىن (大学者), بۇراخان (布葛热汗), ئۇنى (他), 后一项名词短语与前一项名词短语依次组队, 删除单复数冲突的名词短语组队. 通过对比集合 $\{ChainNum, ChainContent\}$, هۆسەيىن(吾斯英) 和ئۇنى(他) 在指代链中, 且在同一条指代链, 经过处理后形成维吾尔语名词短语指代消解的训练和测试样例如表 2 所示.

3.4 栈式自编码

栈式自编码 (Stack autoencoder, SAE) 是深度学习领域中一种重要的无监督学习结构, 它由多个自编码器堆栈组成, 结构如图 2 所示, 前一层自编码器的输出作为后一层自编码器的输入, 通过逐层贪婪训练方法得到最优模型.

自编码器对数据的处理主要分为两个阶段: 编码和解码. 在编码阶段, 栈式自编码接受训练数据 $x \in [0,1]^N$, 首先对其进行线性变化, 在激活函数作用下得到一个编码结果 y $(y \in [0,1]^M)$, 计算如式 (1) 所示. 将编码结果 y 在解码器的作用下, 得到重构训练数据 $\hat{x} \in [0,1]^N$, 计算如式 (2) 所示.

表 2 维吾尔语名词短语指代消解训练和测试样例

Table 2 Training or testing sample format for Uyghur noun phrases

先行语	照应语	样例值 (13 个特征值 + 50 维先行语、照应语 Word embedding)		是否指代
هۆسەيىن	ئۇنى	0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0	0.133, −0.053, 0.114, ···, −0.108 0.177, −0.008, 0.127, ···, −0.055	是
نۇۋز زامانسىنىڭ كاتتا بىلىم ئىگىلىرىدىن بىرى	ئۇنى	0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 0	0.076, 0.099, 0.019, ···, −0.069 0.177, −0.008, 0.127, ···, −0.055	否
بۇراخان	ئۇنى	0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0	0.060, −0.135, 0.277, ···, −0.042 0.177, −0.008, 0.127, ···, −0.055	否

$$y = f_\theta(x) = sigmoid(Wx + b) \tag{1}$$

$$\hat{x} = g_{\theta'}(y) = sigmoid(W^T y + b^{'}) \tag{2}$$

其中, θ 和 $\theta^{'}$ 是编码解码过程中的参数, $W \in \mathbf{R}^{M\times N}$ 代表权重矩阵, b 和 $b^{'} \in \mathbf{R}^M$ 代表隐层的偏置向量, $\text{sigmoid}(\cdot) = (1/1+\exp(-z))$, 函数值域为 [0, 1]. 自编码器的学习过程是无监督的, 其目的是最小化 x 和 $\hat{x}$ 的差距, 降低重构损失, 本文使用的损失函数为 Kullback-Liebler 散度[13], 计算公式如式 (3)、式 (4).

$$\text{argmin}J(x, \hat{x}) = \text{argmin}J(x, g_{\theta'}(f_\theta(x))) \tag{3}$$

$$J(x, \hat{x}) = KL(x||\hat{x}) \tag{4}$$

栈式自编码采用经典的随机梯度下降算法进行训练, 在每次迭代的过程中更新网络参数 $\theta^{'}$, 利用更新得到的最优化参数, 对训练数据进行无监督的学习, 计算公式如式 (5) 所示.

$$a_i^{l+1} = sigm(z_i^{l+1}) = sigm(\sum_{j=1}^{s_l} W_{ij}^{(l)} x_j + {b_i}^l) \tag{5}$$

式 (5) 中, s_l 是第 l 层的神经元总数. a_i^l 表示神经网络的第 l 层中第 i 个神经元的激活值, 即特征值. 令 $h = \sum a_i$, h 为栈式自编码无监督学习算法提取出的一组维吾尔语名词短语对特征, 对于训练数据中 n 组名词短语对构成的特征集合 Feature 可表示为 $\text{Feature} = h^1, h^2, \cdots, h^n$.

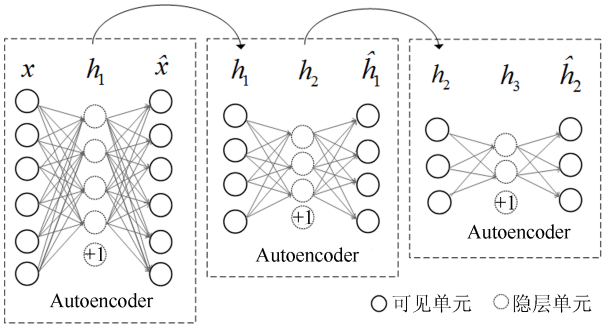


图 2　栈式自编码网络结构图

Fig. 2　Structure of SAE

3.5　Softmax 回归

Softmax 回归模型是 Logistic 回归模型在多分类问题上的推广, 本文利用 Softmax 解决维吾尔语名词短语指代消解的分类问题. 其基本思想是通过计算给定样本属于某一类别的概率, 确定样本的所属类别. 假设 $\{(x^1, y^1), \cdots, (x^n, y^n)\}$ 为给定样本, 且 $x^i \in Feature$ 其中 Feature 为第 3.4 节中栈式自编码器提取的特征集合, $y^i \in \{1, 2\}$ 为 x^i 对应的标签, 则估计样本 x^i 在 softmax 函数中所对应的类别标签概率 p 为

$$h_\theta(x^i) = \begin{bmatrix} p((y^i)=1|x^i;\theta) \\ p((y^i)=2|x^i;\theta) \end{bmatrix} = \frac{1}{e^{\theta_1^T x} + e^{\theta_2^T x^{(i)}}} \begin{bmatrix} e^{\theta_1^T x} \\ e^{\theta_2^T x} \end{bmatrix} \tag{6}$$

其中, θ 是 $h_\theta(x^i)$ 的参数, $1/(e^{\theta_1^T x} + e^{\theta_2^T x^{(i)}})$ 是对概率分布进行归一化, 使得所有概率和为 1.

4　实验与分析

4.1　语料准备

目前比较知名的指代消解标注语料有 MUC (仅有 English 语料) 和 ACE (有 Arabic、English、Chinese 三种语料), 而关于维吾尔语指代消解评测语料还未见报道. 因此, 本文针对维吾尔语指代消解任务对语料进行了筛选和标注.

叙述文是话语分析中典型的语料, 本文选取天山网、人民网、论坛和博客等维吾尔语版网页作为语料来源, 筛选出以记人和叙事为题材且至少包含两条指代链信息的叙述文作为实验语料. 在实验组维吾尔语专家的指导下对语料进行标注, 所标注的内容包括指代链信息、名词短语、名词的类别、名词的数范畴、名词的格范畴. 对标注后的语料选用 XML 文件进行存储.

本次实验采用 200 篇标注完成的语料其中包括 489 条指代链信息, 7 924 个名词短语. 经过第 3.3 节生成训练和测试数据, 其中具有指代关系的名词短语对有 7 721 组, 不具有指代关系的名词短语对有 20 432 组.

4.2　实验结果与分析

为了便于比较, 本文采用 MUC 评测对指代消解结果进行技术评估. 采用准确率 P (Precision)、召回率 R (Recall) 和 F 值三个重要指标衡量消解结果. 其中, 准确率指正确消解的对象占实际消解的对象的百分比. 召回率是指正确消解的对象占消解系统应消解对象的百分比. F 值是正确率、召回率的综合评价指标, 即: $F = R \times P \times 2/(R + P)$.

本文所有实验都采用 5 折交叉验证, 为保证结果的稳定性, 取平均值作为最终结果. 通过反复试验尝试了网络模型不同的参数, 本实验数据量下的最优参数如下表 3 所示. (ρ: 稀疏性参数, β: 稀疏惩罚因子权重, λ: 权重衰减参数, $maxIter$: 迭代次数)

表 3　SAE 模型最优参数

Table 3　Optimal parameters of SAE

参数	ρ	β	λ	maxIter
值	0.1	3	3E−3	800

4.2.1 基于 SAE 模型的有效性验证

在不引入 Word embedding 的前提下, 使用第 3.1 节提取的特征向量作为输入, 利用 SAE^i (i 表示 SAE 包含的自编码器层数) 和 SVM 模型进行指代消解实验. SVM 是处理非线性数据较好的浅层机器学习模型, 与 SAE 模型有较好的对比性, 通过实验验证, 利用多项式核函数的 SVM 模型性能优于线性核函数和 RBF 核函数. 因此本文选用性能最优的 SVM 模型进行对比实验. 结果如表 4 所示.

从表 4 可以看出: SAE^1 与 SVM 相比召回率高了 3.2 %, 但准确率和 F 系数都低于 SVM; 增加 2 层自编码器之后的三个指标已经超过 SVM; 增加 3 层自编码器时, F 系数较 SVM 有了一定的增长; 当增加 4 层自编码器时, 正确率较 SVM 提高了 1.9 %, 召回率提高了 1.6 %, F 系数提高了 1.8 %, 实验结果证明了针对维吾尔语名词短语指代消解任务, 基于 SAE 的模型其效果要优于基于 SVM 的模型. 从表 4 中还可以看出, 随着自编码器层数的增多, 实验结果的准确率均有所提高, 这是因为深度学习通过多层映射单元提取出主要的结构信息, 其计算能力优于单层结构.

表 4 基于 SAE 模型的有效性验证

Table 4 The validation of SAE effectiveness

模型	P (%)	R (%)	F (%)
SAE^1	61.775	73.319	67.054
SAE^2	66.064	71.256	68.562
SAE^3	66.134	71.995	68.940
SAE^4	68.695	71.743	70.186
SVM	66.727	70.115	68.379

4.2.2 特征选取对模型性能的影响

在指代消解过程中, 特征的选择对模型的性能影响较大, 为了探讨单个特征对指代消解的贡献度, 本文基于第 4.2.1 节中实验效果最佳的 SAE^4 对各个特征所起的作用分别进行了详细实验, 采用的特征集逐步在上一级特征集扩展, 即下级采用的特征集包含上级所有特征集, 实验结果如表 5 所示. 从表 5 中可以看出, 随着特征集的不断扩展, 指代消解识别结果的各项指标均有所提高. 当照应语判定为专有名词时指代消解的准确率为 46.079 %, 但召回率较低, 这是因为特征集很少的情况下自编码算法无法学习到充分语义特征. 当加入带领属性名词短语的特征时, F 系数较前一项有了一定的提升, 这是因为本文选取的语料以记人或叙事为题材, 这类文章多出现 مېنىڭ دوستۇم (我的朋友) ئۇنىڭ دادىسى (他的爸爸) 等带领属性人称词尾的名词短语. 随着特征集的不断加入, 当特征集包含了本文提取的 13 个特征项时正确率达到 68.695 %, 召回率达到 71.743 %, F 系数达到了 70.186 % 最高值. 实验证明本文提取的特征集合是有效的.

表 5 特征集对结果的影响

Table 5 The influence of introducing features sets

特征项	P (%)	R (%)	F (%)
AnProperNoun	46.079	0.856	1.681
CaProperNoun	66.159	0.713	1.411
AnDefiniteNP	75.897	1.102	2.172
CaDefiniteNP	59.932	4.579	8.508
AnDemonstrativeNP	65.432	8.409	14.903
CaDemonstrativeNP	57.092	9.112	15.716
AnPossessionNP	60.411	26.912	37.222
CaPossessionNP	44.439	38.403	41.201
AnPossessionNP	48.231	51.082	49.616
CaPossessionNP	45.831	70.334	55.498
PropertyFit	64.470	51.108	57.017
SinglePluralFit	58.631	80.205	67.742
FullMatch	68.695	71.743	70.186

4.2.3 Word Embedding 对实验的影响

Word embedding 富含词汇语义及上下文位置关系, 为探讨 Word embedding 对模型的分层结构学习性能的影响, 实验选用 10 维的 Word embedding 加入特征集, 依次训练 SAE^1, SAE^2, SAE^3, SAE^4 模型, 实验结果如表 6 所示, SAE^i + WE 表示, 在原有特征集中加入 Word embedding, 训练 SAE^i 模型.

表 6 Word embedding 的引入对实验的影响

Table 6 The influence of introducing word embedding

模型	P (%)	R (%)	F (%)
SAE^1	60.915	74.569	67.054
SAE^1 + WE	64.382	70.103	67.121
SAE^2	66.064	71.256	68.562
SAE^2+ WE	66.571	71.419	68.910
SAE^3	66.134	71.995	68.940
SAE^3 + WE	68.215	72.375	70.233
SAE^4	68.695	71.743	70.186
SAE^4+ WE	72.352	69.743	71.024

从表 6 中可以看出, 引入 Word embedding 作为特征项对包含不同层自编码器的 SAE 模型都是

有效的. 用加入 Word embedding 的特征集训练 SAE^2 时, 正确率、召回率、F 系数都有所提高, 各项指标接近包含 3 层自编码器的 SAE 模型. 当训练时 SAE^3, 召回率和 F 系数已经超过不引入 Word embedding 的 SAE^4. 这是因为 Word embedding 每一维都包含丰富的上下文信息, 能够很好地表示语义特征, 并且使语义类似的词语, 其向量表示也比较接近, 进一步促进模型对语料深层语义的学习, 进而提高了模型指代消解的性能.

Word embedding 维度选择对 SAE 模型的性能有一定的影响, 为探讨维度为多大时能更好地表达文本语义信息, 将 Word embedding 的维度设定为 10 维、50 维、100 维、150 维、200 维, 分别作为特征项, 引入特征集, 训练 SAE^4 和 SVM 模型. 实验结果如表 7 所示.

表 7 Word embedding 维度对实验的影响

Table 7 The influence of adjusting word embedding dimension

模型 / 维度	SAE^4 + WE			SVM + WE		
	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)
10	72.4	69.7	71.0	67.0	70.3	68.6
50	73.9	69.8	71.8	70.5	69.8	70.1
100	74.5	70.6	72.4	69.9	69.9	69.9
150	75.8	68.4	71.9	69.0	70.4	69.7
200	77.0	67.0	71.9	68.2	70.9	69.4

从表 7 可以看出, Word embedding 的维度选择对 SAE 和 SVM 模型的性能都有影响. 当特征集中 Word embedding 维度为 50 维时 SVM 模型的性能最佳, F 系数为 70.1%. 当维度为 100 维时, SAE 模型的性能最佳, F 系数达到了 72.4% 最高值, 较 SVM 提高了 2.3%. 随着维度的不断地增加, 两个模型的正确率开始回落, 性能下降. 其原因是当维度过高时, 产生过拟合现象, 模型对数据的泛化能力降低.

5 结论

指代消解的研究有助于自然语言处理技术的发展, 具有很大的研究价值和实用价值. 现有的研究主要针对英语、汉语等大语种, 而对于维吾尔语指代消解的分析研究还很少. 针对以上不足, 本文提出利用栈式自编码算法同时基于语义特征的维吾尔语名词短语指代消解方法. 与以往的研究方法相比, 本文利用深度学习机制无监督地提取文本中主要的结构信息, 挖掘深层语义. 通过引入 Word embedding 作为高层抽象特征, 进一步地提高特征对语义的表达. 实验结果证明, 深度学习模型较浅层机器学习模型更适用于本文维吾尔语名词短语指代消解任务, 同时, Word embedding 的引入, 有效地提高了模型对语义的学习和理解能力.

6 致谢

感谢张晶、李冬白、史新宇、高双印、周兴发、陶豆豆、秦越、黎红和张灯柯等同学在本文实验和论文撰写方面提供的热情帮助, 在此谨向他们致以诚挚的谢意和崇高的敬意.

References

1 Zelenko D, Aone C, Tibbetts J. Coreference resolution for information extraction. In: Proceedings of the 2004 ACL Workshop on Reference Resolution and its Applications. Barcelona, Spain: ACL, 2004. 9−16

2 Soon W M, Ng H T, Lim D C Y. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 2001, **27**(4): 521−544

3 Bergsma S, Lin D K. Bootstrapping path-based pronoun resolution. In: Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics. Sydney: Association for Computational Linguistics, 2006. 33−40

4 Ng V. Semantic class induction and coreference resolution. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics. Prague, Czech Republic: ACL, 2007. 536−543

5 Bengtson E, Roth D. Understanding the value of features for coreference resolution. In: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing. Honolulu: Association for Computational Linguistics, 2008. 294−303

6 Zhou Jun-Sheng, Huang Shu-Jian, Chen Jia-Jun, Qu Wei-Guang. A new graph clustering algorithm for Chinese noun phrase coreference resolution. *Journal of Chinese Information Processing*, 2007, **21**(2): 77−82
(周俊生, 黄书剑, 陈家骏, 曲维光. 一种基于图划分的无监督汉语指代消解算法. 中文信息学报, 2007, **21**(2): 77−82)

7 Wang Hai-Dong, Hu Nai-Quan, Kong Fang, Zhou Guo-Dong. Research on semantic role information in anaphora resolution. *Journal of Chinese Information Processing*, 2009, **23**(1): 23−29
(王海东, 胡乃全, 孔芳, 周国栋. 指代消解中语义角色特征的研究. 中文信息学报, 2009, **23**(1): 23−29)

8 Kong Fang, Zhou Guo-Dong. Pronoun resolution in English and Chinese languages based on tree kernel. *Journal of Software*, 2012, **23**(5): 1085−1099
(孔芳, 周国栋. 基于树核函数的中英文代词消解. 软件学报, 2012, **23**(5): 1085−1099)

9 Xi Xue-Feng, Zhou Guo-Dong. Pronoun resolution based on deep learning. *Acta Scientiarum Naturalium Universitatis*

Pekinensis, 2014, **50**(1): 100−110
(奚雪峰, 周国栋. 基于 Deep Learning 的代词指代消解. 北京大学学报 (自然科学版), 2014, **50**(1): 100−110)

10 Mikolov T, Sutskever I, Chen K, Corrado G S, Dean J. Distributed representations of words and phrases and their compositionality. In: Proceedings of the 2013 Advances in Neural Information Processing Systems 26. Lake Tahoe, Nevada, USA: Curran Associates, Inc., 2013. 3111−3119

11 Kim Y. Convolutional neural networks for sentence classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar: Association for Computational Linguistics, 2014. 746−1751

12 Glorot X, Bordes A, Bengio Y. Deep sparse rectifier neural networks. In: Proceedings of the 14th International Conference on Artificial Intelligence and Statistics. Fort Lauderdale, USA, 2011. 315−323

13 Glorot X, Bordes A, Bengio Y. Domain adaptation for large-scale sentiment classification: a deep learning approach. In: Proceedings of the 28th International Conference on Machine Learning. Bellevue, Washington, USA: Omnipress, 2011. 513−520

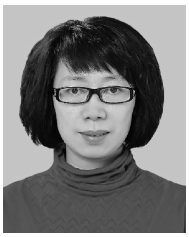
14 Lu S X, Chen Z B, Xu B. Learning new semi-supervised deep auto-encoder features for statistical machine translation. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore, Maryland, USA: ACL, 2014. 122−132

15 Wang Hou-Feng, Mei Zheng. Robust pronominal resolution within Chinese text. *Journal of Software*, 2005, **16**(5): 700−707
(王厚峰, 梅铮. 鲁棒性的汉语人称代词消解. 软件学报, 2005, **16**(5): 700−707)

16 Patgul · Mamat. Uyghur Discourse Anaphora based on Centering Theory [Ph. D. dissertation], Minzu University of China, China, 2010
(帕提古力 · 麦麦提. 基于向心理论的维吾尔语语篇回指研究 [博士学位论文], 中央民族大学, 中国, 2010)

17 He Yu, Pan Da, Fu Guo-Hong. Chinese explanatory opinionated sentence recognition based on auto-Encoding features. *Acta Scientiarum Naturalium Universitatis Pekinensis*, 2015, **51**(2): 235−240
(贺宇, 潘达, 付国宏. 基于自动编码特征的汉语解释性意见句识别. 北京大学学报 (自然科学版), 2015, **51**(2): 235−240)

李 敏 新疆大学硕士研究生. 主要研究方向为自然语言处理.
E-mail: limin_xju@163.com
(**LI Min** Master student at Xinjiang University. Her main research interest is natural language processing.)



禹 龙 新疆大学教授. 主要研究方向为计算机智能技术与计算机网络. 本文通信作者.
E-mail: yul_xju@163.com
(**YU Long** Professor at Xinjiang University. Her research interest covers computer intelligence technology and computer networks. Corresponding author of this paper.)

田生伟 新疆大学教授. 主要研究方向为自然语言处理与计算机智能技术.
E-mail: tianshengwei@163.com
(**TIAN Sheng-Wei** Professor at Xinjiang University. His research interest covers natural language processing and computer intelligence technology.)

吐尔根 · 依布拉音 新疆大学教授. 主要研究方向为计算机智能技术与自然语言处理. E-mail: mytlgxj@126.com
(**TurgLm IBRAHIM** Professor at Xinjiang University. His research interest covers computer intelligence technology and natural language processing.)

赵建国 新疆大学副教授. 主要研究方向为维汉双语对比.
E-mail: 13899951918@126.com
(**ZHAO Jian-Guo** Associate professor at Xinjiang University. His main research interest is Uyghur-Chinese bilinguals comparison.)