信息安全领域内实体共指消解技术研究

张晗[1,2]，胡永进[1]，郭渊博[1]，陈吉成[3]
（1. 信息工程大学密码工程学院，河南 郑州 450001；2. 郑州大学软件学院，河南 郑州 450000；
3. 信息工程大学信息技术研究所，河南 郑州 450001）

摘 要：针对信息安全领域内的共指消解问题，提出了一个混合型方法。该方法在原来 BiLSTM-attention-CRF 模型的基础上引入领域词典匹配机制，将其与文档层面的注意力机制相结合，作为一种新的基于字典的注意力机制，来解决从文本中提取候选词时对稀有实体以及长度较长的实体识别能力稍弱的问题，并通过总结领域文本特征，将提取出的待消解候选词根据词性分别采用规则与机器学习的方式进行消解，以提高准确性。通过在安全领域数据集的实验，分别从共指消解以及提取候选词并分类 2 个方面证明了方法的优越性。

关键词：共指消解；混合型方法；领域词典匹配机制；BiLSTM-attention-CRF 模型；信息安全
中图分类号：TP393
文献标识码：A
doi:10.11959/j.issn.1000−436x.2020033

Research on coreference resolution technology of entity in information security

ZHANG Han[1,2], HU Yongjin[1], GUO Yuanbo[1], CHEN Jicheng[3]
1. Department of Cryptogram Engineering, Information Engineering University, Zhengzhou 450001, China
2. Software College, Zhengzhou University, Zhengzhou 450000, China
3. Institute of information technology, Information Engineering University, Zhengzhou 450001, China

Abstract: To solve the problem of coreference resolution in information security, a hybrid method was proposed. Based on the BiLSTM-attention-CRF model, the domain-dictionary matching mechanism was introduced and combined with the attention mechanism at the document level. As a new dictionary-based attention mechanism, the word features were calculated to solve the problem of weak recognition ability of rare entities and entities with long length when extracting candidates from text. And by summarizing the features of the domain texts, the candidates were coreferenced by rules and machine learning according to the part of speech to improve the accuracy. Through the experiments on security data set, the superiority of the method is proved from the aspects of coreference resolution and extraction of candidates from text .

Key words: coreference resolution, hybrid method, domain-dictionary matching mechanism, BiLSTM-attention-CRF, information security

1 引言

实体的共指消解（CR, coreference resolution）旨在解决文档中对实体的重复引用问题，是自然语言处理（NLP, natural language processing）研究的核心内容[1]。它主要用于提高其他 NLP 任务中诸如机器翻译[2-4]、情感分析[5-7]、关系提取[8-10]以及摘要自动生成[11-12]等任务的性能。目前，这项研究多集中在通用领域，主要原因在于：1) 关于通用领域的共指消解技术有丰富的研究经验[13-25]；2) 通用领域

收稿日期：2019−10−14；修回日期：2019−12−27
基金项目：国家自然科学基金资助项目（No.61501515）；河南省重点科技攻关基金资助项目（No.172102210002）；郑州大学青年骨干教师基金资助项目（No.2017ZDGGJS048）
Foundation Items: The National Natural Science Foundation of China (No.61501515), The Project of Henan Provincial Key Scientific and Technology (No.172102210002), The Young Scholar teachers project of Zhengzhou University (No.2017ZDGGJS048)

内的标注语料充足，例如自动内容抽取（ACE, automatic content extraction）语料库[26]、CoNLL-2012[27]、Parcor 语料库[28]等。而关于此项工作在信息安全领域的研究，目前并未找到相关的研究文献。

但是，这并不意味着在信息安全领域内不需要此项工作。例如“As the world’s first cyber ‘super destructive weapon’, Stuxnet has infected more than 45 000 networks around the world. Computer security experts believe the virus is the highest level ‘worm’ ever. The new virus uses a variety of advanced technologies, so it is extremely stealthy and destructive.”，在这句话中，“Stuxnet”“the virus”“the new virus”和“it”代表的都是同一个实体“Stunxnet”。通过共指消解，可以获得“Stunxnet”与“the highest level worm”之间是“is-a”的关系，这将提高从文本中提取实体属性关系的准确性，从而使信息安全领域内知识图谱更加完善，而知识图谱的完善也会使其对威胁的预警更加精确。

常用的共指消解技术有 3 种：第一种是基于规则的方法[13-16]，第二种是基于统计的方法[17-20]，第三种是基于深度学习的方法[21-25]。其中，第一种方法依赖于手工定制的规则，覆盖面较窄，灵活性较差，不能很好地处理丰富的词汇信息；第二种方法虽然可以处理丰富的词汇特征，但是有学者认为从准确性上第二种方法的表现弱于第一种方法[29]；第三种方法更适用于包含大规模标注数据的领域，而信息安全领域缺乏大规模的可用于共指消解的标注数据，因此该方法并不适用于信息安全领域。文献[19]采用将规则与统计相结合的方法来解决共指消解的问题，虽然该文献提出的方法在通用领域达到了比较理想的效果，但是信息安全领域由于其特殊性与通用实体共指消解有所不同，归纳如下。

1) 处理实体类型不同，所以提取候选词的词类不同。在通用领域内，进行提取和消解的实体类型为人名、地名、组织名等；信息安全领域内的实体类型多与“产品”“漏洞”“攻击”相关，因此实体的组合形式通常是以短语的形式出现，例如“Advanced Persistent Threat”。而且，这些实体通常为物体，文本中经常会出现类似于“damage of the virus”形式的短语，其中的“the virus”也是需要提取的待消解词，因此对于信息安全领域中的文本，待提取词除了简单的名词、代词、专有名词之外，还包括名词短语以及一些名词短语中包含的嵌套短语。

2) 待提取的词类不同，所以提取方法不同。例如，文献[19]中提取的候选词包括文本中所有的普通名词、专有名词以及代词，提取时用到的是诸如同位语、谓语主格以及角色同位语之类的句法模式。而对于信息安全领域中的文本来说，根据句法模式提取出的候选词并不能满足需要，因此要用到的提取方法也与文献[19]不同。

3) 对待消解词进行消解时用到的特征不同。例如，在通用领域对实体类型“人名”进行消解时，可以将性别作为一个重要特征进行考量；而信息安全领域的实体类型多为第三人称形式表示，没有性别特征。

4) 相较于通用领域，信息安全领域文本中含有大量的术语和专有名词以及缩写。虽然通用领域中也有一些关于国家或者地名的缩写，但是文献[19]以 OntoNotes[30]作为参考并没有对此类缩写进行专门处理。

针对以上问题，本文提出了一种混合的方法来解决信息安全领域内的共指消解问题。本文工作主要分为 2 个部分：1) 从给定文档中提取出所有的候选词语（包括名词性短语、代词、实体以及嵌套短语）并进行分类；2) 对待消解项进行共指消解。本文研究团队在之前的工作中[31]，提出了一种 BiLSTM+attention+CRF 模型来进行文档中的命名实体识别，解决了文档中存在的同一实体标注不一致的问题，例如 Advanced Persistent Threat 和 APT。该模型是在经典 BiLSTM-CRF 模型的基础上加入 Attention 机制来关注当前实体与文档中其他所有单词的相关性，得到该单词在文档层面的特征表示，再进行实体的抽取和分类。但是通过实验发现，该模型对训练集中没有出现过的稀有实体以及长度较长的实体识别能力稍弱，因此本文对该模型进行了改进，提出了一种改进之后的模型 BiLSTM+dic_attention+CRF。该模型引入了领域词典匹配机制，将其与文档层面的注意力机制相结合，作为一种新的基于字典的注意力机制来计算单词特征。此外，由于要提取的候选词除了实体之外还包括一些名词性短语、代词以及嵌套短语。如果只使用 BiLSTM+dic_attention+CRF 模型来提取名词性短语和嵌套短语，需要浪费大量的人力物力来标注数据，而名词短语和嵌套短语具有一定的语法规则可

以进行归纳总结，因此，本文采用规则+BiLSTM+dic_attention+CRF 模型的方式来进行候选词的抽取和分类。本文所做贡献如下。

1) 提出一种将规则与深度学习模型（BiLSTM+dic_attention+CRF）相结合的方法来解决信息安全领域内从文本中提取候选词及分类的问题。

2) 提出一种将规则与机器学习相结合的混合型方法来解决信息安全领域内的共指消解问题。

与现有方法相比，本文所提出的方法在信息安全领域的数据集上达到了更好的性能。

2　相关工作

关于共指消解的研究由来已久，早期主要集中在基于规则的方法，包括基于语法的 Hobbs 理论[13]、基于对话的 Centering 理论[14]和基于语法的 RAP 算法[15]。在 21 世纪早期，一部分学者认为这种基于规则的方法在表现性能上要优于机器学习的方法[29]，但是，基于规则的方法的缺点也非常明显，它过于依赖人们手工制定规则的能力，规则制定的好坏将直接影响方法的性能，并且基于规则的方法灵活性较差，耗费人力过多。共指消解中关于机器学习的研究主要集中在训练分类器[1]，其中决策树和随机森林是最常用的分类器[17-19]。文献[19]提出了一种将规则与统计分类器相结合的方法来进行共指消解，该方法针对每一种要进行共指消解的类型都训练了一个统计分类器，通过实验不仅证明了这种混合型方法优于基于规则的共指消解方法，还证明了随机森林作为共指消解分类器的优越性。但是该方法仅针对通用领域内的共指消解，因此在选择特征时也是根据通用领域内的文本数据进行选择的。随着深度学习模型在自然语言处理领域的应用，它们也逐渐被应用于共指消解任务[21-25]。文献[21]提出了第一个关于共指消解的深度学习模型，它通过对 2 个单独的子任务（回指检测和先行词排序）进行预训练，以学习不同的特征表示。该模型也证明了从实体类群中获取全局特征有利于提高共指消解的性能，但该文献的前提是实体类群是已经事先分类完成的，而本文的工作是首先从文本中提取相关的候选词，因此本文将文献[21]中的实体类群中的全局特征转换成了文档中的全局特征。文献[22]将检测候选词与共指消解相结合，它首先使用卷积神经网络（CNN, convolutional neural network）学习字符的特征，通过 LSTM 学习单词的特征，然后通过 Attention 机制学习候选词的特征表示，并通过一个前馈神经网络对候选词对应的先行词进行排序。该模型所使用的深度神经网络非常庞大，因此很难维护。

除了这些应用于通用领域的研究之外，生物领域内的共指消解研究也有所发展，主要原因在于，生物领域也具有诸如 MEDSTRACT[32]和 MEDCo[33]这样的大型标注语料库。典型的应用包括文献[34]中提出的一种基于机器学习和规则的混合方法，它的 F1 值为 60.9%，是目前生物领域内性能最先进的方法。

针对以上方法中存在的各种问题，本文提出了一种混合型的方法来处理信息安全领域内的共指消解，将从文本中提取待消解项也作为工作的一部分。首先，采用一种规则与深度学习模型相结合的方法来提取文本中的待消解项。其次，采用规则和随机森林的方法进行共指消解。与深度学习的方法相比，这种方法结构简单，需要的训练数据较少。并且本文通过对信息安全领域内文本的研究，挖掘出了适用于信息安全领域共指消解的数据特征并制定出一套规则，可用于信息安全领域的共指消解。

3　基于规则与机器学习的共指消解方法

本文提出了一种将规则与机器学习混合的方法来进行共指消解。该方法的工作分为 2 个部分：1) 从给定文档中提取所有的候选词语（包括名词性短语、代词、实体以及嵌套短语）并进行分类；2) 对待消解项进行共指消解。方法框架如图 1 所示。

从图 1 中可以看出，该方法分为提取候选词和共指消解这 2 个部分。其中，提取候选词部分由规则+BiLSTM+ dic_ttention+ CRF 混合而成，用于提取文本中的待消解词并进行分类，该部分对应 3.1 节的内容。共指消解则主要对分类之后的待消解词进行共指消解，该部分对应 3.2 节的内容。图 1 中，最底层的其他特征代表 3.1.2 节中除了单词特征之外的其他特征；顶层中的其他特征代表 3.2.2 节中的名词短语共指消解除了类型一致性之外的其他特征；$D(w_i)$ 表示单词 w_i 在领域词典中的匹配度；W_d 表示单词匹配度所占的权重；r_i^g 表示单词 w_i 在文档层面的特征表示；g_i 表示单词 w_i 基于领域词典的新的文档层面特征表示。

3.1　提取候选词语并分类

本文要提取的候选词包括名词性短语、代词、实体以及嵌套短语，可将其分为名词性短语和嵌套短语的提取以及实体的提取。名词性短语和嵌套短

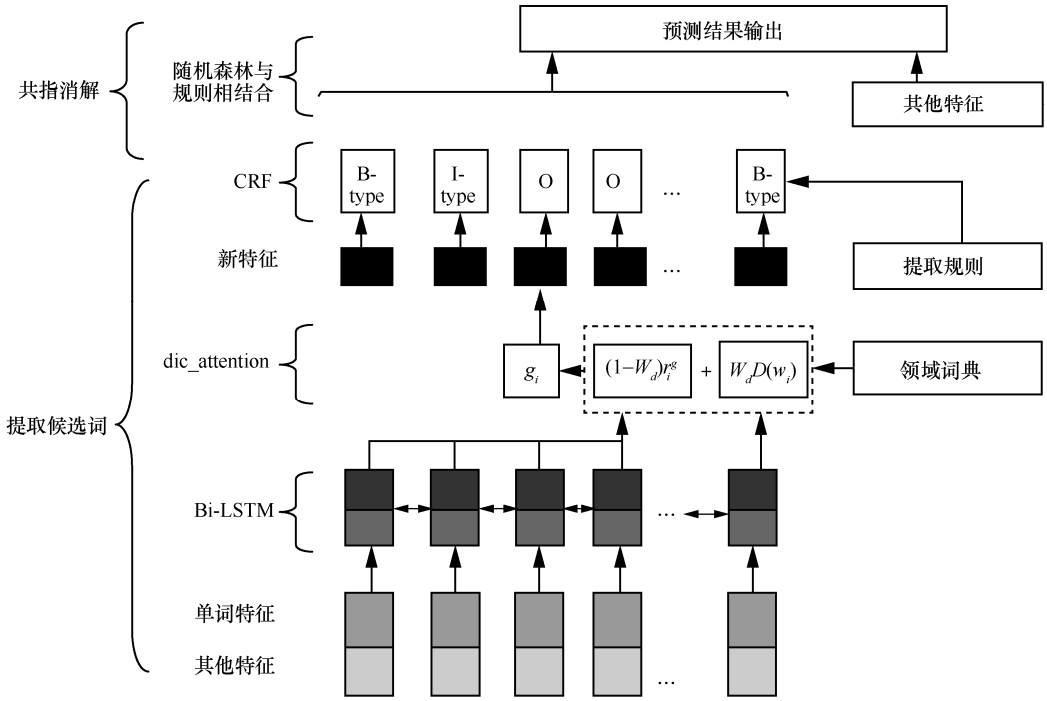


图 1 方法框架

语的提取采用规则的形式进行提取，实体的提取采用 BiLSTM+dic_ttention+CRF 模型进行提取。具体架构如图 1 中的提取候选词部分所示。

3.1.1 *名词短语和嵌套短语的提取规则*

通常情况下，名词短语是由名词以及它的修饰语组成，中心词为名词。名词的修饰语与名词有 2 种位置关系：一是放在被修饰名词的前面，叫作前置定语或定语；二是放在被修饰名词的后面，叫作后置定语。通过对信息安全领域内语料的分析发现，需要进行共指消解的名词短语通常为前置定语名词短语，因此这里只考虑第一种位置关系的情况。

一般来说，作为前置定语的词类有 2 种：其一是限定词，用来限定名词所指范围，例如 these、three、a、the、my 等；其二是形容词，用来表示名词的性质和特征，比如 red、close、new、small 等。因此，可以通过如下规则来获取名词性短语。

假设 U_1 表示冠词集合，U_2 表示形容词性物主代词集合，U_3 表示名词性物主代词集合，U_4 表示指示限定词集合，U_5 表示数量词集合，U_6 表示基数词集合，N 表示名词集合，NP 表示名词短语集合，AD 表示形容词集合，集合 $U=U_1\cup U_2\cup U_3\cup U_4\cup U_5\cup U_6$。

1) 如果单词 a 属于冠词、形容词性物主代词、名词性物主代词、指示限定词、数词、量词、基数词等集合中的任意一个单词，单词 b 属于名词集合，则 ab 构成名词性短语。

2) 如果单词 c 属于形容词，单词 b 属于名词集合，则 cb 构成名词短语。

3) acb 属于名词短语。

可表示为

$$(\forall a)(\forall c)(\forall b)(\mathrm{BEL}(a,U)\wedge\mathrm{BEL}(c,\mathrm{AD})\wedge \mathrm{BEL}(b,N))\rightarrow(\forall ab)(\forall cb)(\forall acb)(\mathrm{BEL}(ab,\mathrm{NP})\vee\mathrm{BEL}(cb,\mathrm{NP})\vee \mathrm{BEL}(acb,\mathrm{NP}))$$

其中，BEL 表示谓语动词“属于”。

除此之外，还要提取嵌套短语。嵌套短语通常存在于所提取出的名词短语中，通过对嵌套短语的语法分析，制定出如下提取规则。

假设 NNP 表示嵌套短语集合，ONP 表示所有格名词短语集合，P 表示介词集合。

1) 嵌套短语来自所有格名词短语。例如，短语“its methods”中的嵌套短语是代词“its”“stuxnet’s damage”中的嵌套短语是专有名词“stuxnet”，可表示为

$$(\forall ab)(\forall a)(\mathrm{BEL}(ab,\mathrm{ONP})\wedge(\mathrm{BEL}(a,U_2)\vee \mathrm{BEL}(a,U_3)))\rightarrow(\forall a)(\mathrm{BEL}(a,\mathrm{NNP}))$$

2) 嵌套短语是名词性短语中的名词或介词。例如，名词性短语“efficiency reduction”中嵌套短语

是“efficiency”，可表示为

$$(\forall ab)(\forall a)(\mathrm{BEL}(ab,\mathrm{NP})\wedge(\mathrm{BEL}(a,N)\vee \mathrm{BEL}(a,P)))\rightarrow(\forall a)(\mathrm{BEL}(a,\mathrm{NNP}))$$

如果提取出的名词性短语中包含实体，那么只提取实体部分。

3.1.2 实体提取和分类

1) 输入特征

在图 1 中，除了使用制定的规则来提取名词短语和嵌套名词短语之外，还要使用模型 BiLSTM-dic_attention-CRF 对文本中的实体进行提取和分类。此时，模型的输入特征包括以下几个方面。

① 单词特征。单词特征又被称为分布式单词表示，可以从一个大型的未标记语料库中获取单词的语义和句法信息。Word2vec 是最常用的单词向量训练工具。为了获得高质量的单词向量表示，本文使用文献[31]中从 CVE 语料库中收集的 94 534 条漏洞记录描述进行单词向量训练。

② 词性标注特征（PoS, part of speech）。词性标注特征也被称为语法标注或词类消疑，是语料库语言学中将语料库内单词的词性按其含义和上下文内容进行标记的文本数据处理技术。人们要提取的候选词，特别是名词性短语和嵌套短语，都要用到单词的词性标注信息，因此，词性标注需要作为重要的输入特征。本文使用 Stanford CoreNLP 作为词性标注工具。

③ 字符特征。字符特征包含实体名称的结构信息，可以表示实体名称的特定组成，特别是在信息安全领域。例如，影响 Windows 的 PE 病毒 Backdoor.Win32.Gpigeon.pd 和 Backdoor.Win32.Gpigeon2010.pc 具有相同的前缀，因此，当模型遇到这些单词时，人们可以根据它的前缀判断出它是 Windows 的 PE 病毒的名称。与传统手工设计的字符特征不同，人们可以通过训练得到单词的字符特征向量。因为英文中的字符个数有限，构造出的字符表远远小于词表，因此本文采用文献[35]中提到的字符训练方法，通过 PoS 标记来对字符进行训练，并进行大小写字符及特殊字符的区分。

2) BiLSTM-dic_attention-CRF 模型

BiLSTM-dic_attention-CRF 模型是在 BiLSTM-attention-CRF 模型的基础上添加了领域词典，将其与原来的文档层面 Attention 机制相结合，作为一种新的基于字典的注意力机制来计算单词特征，以解决 BiLSTM-attention-CRF 模型对训练集中未出现过的稀有实体以及长度较长实体识别稍弱的问题。本文模型中使用到的领域词典是本文研究团队在之前的工作[36]中，通过 wikipedia 和信息安全领域的 UCO（unified cybersecurity ontology）构造出来的。

假设有文档 $D=\{s_1,s_2,\cdots,s_n\}$，其中，$s_i=\{w_1,w_2,\cdots,w_m\}$ 为组成文档 D 的第 i 个句子，$\boldsymbol{w}_i=\{\boldsymbol{c}_1,\boldsymbol{c}_2,\cdots,\boldsymbol{c}_k\}$ 为句子 s_i 的第 i 个单词表示，$\boldsymbol{c}_i$ 是单词 $\boldsymbol{w}_i$ 的第 i 个字符特征表示，$\boldsymbol{p}_i$ 表示单词 $\boldsymbol{w}_i$ 的词性特征表示，则可以得到关于单词 $\boldsymbol{w}_i$ 的新的特征表示 $\boldsymbol{h}_i$ 为

$$\boldsymbol{h}_i=\mathrm{BiLSTM}(\boldsymbol{w}_i,\boldsymbol{c}_1,\boldsymbol{c}_2,\cdots,\boldsymbol{c}_k,\boldsymbol{p}_i) \tag{1}$$

其中，$\boldsymbol{h}_i$ 作为 Attention 层的输入，Attention 层主要用来计算单词 $\boldsymbol{w}_i$ 与文本中其他单词 $\boldsymbol{w}_j(j=1,2,3,\cdots,i-1,\ i+1,\cdots,mn)$ 的关联度，该权重值 $\boldsymbol{a}_{ij}$ 可表示为式(3)。

$$f(\boldsymbol{h}_i,\boldsymbol{h}_j)=\boldsymbol{h}_i^{\mathrm{T}}\boldsymbol{W}_a\boldsymbol{h}_j \tag{2}$$

$$\boldsymbol{a}_{ij}=\frac{\exp(f(\boldsymbol{h}_i,\boldsymbol{h}_j))}{\sum_{k=1}^{nm}\exp(f(\boldsymbol{h}_i,\boldsymbol{h}_k))} \tag{3}$$

其中，$\boldsymbol{W}_a$ 为需要训练的模型参数。

此时，可以得出文档层面的一个全局特征表示 $\boldsymbol{r}_i^g$ 为

$$\boldsymbol{r}_i^g=\sum_{j=1}^{nm}\boldsymbol{a}_{ij}\boldsymbol{h}_j \tag{4}$$

这里引入了领域词典匹配机制，将其与 Attention 机制相结合，计算出新的基于领域词典的全局特征表示 $\boldsymbol{g}_i$ 如式(8)所示。

$$D(\boldsymbol{w}_i)=\begin{cases}1,\ \boldsymbol{w}_i\text{与词典相匹配}\\ 0,\ \text{其他}\end{cases} \tag{5}$$

$$s(\boldsymbol{h}_i,\boldsymbol{r}_i^g)=\boldsymbol{W}_d D(\boldsymbol{w}_i)\boldsymbol{h}_i+(1-\boldsymbol{W}_d)\boldsymbol{r}_i^g \tag{6}$$

$$\boldsymbol{A}_{ii}=\frac{\exp(s(\boldsymbol{h}_i,\boldsymbol{r}_i^g))}{\sum_{k=1}^{nm}\exp(s(\boldsymbol{h}_i,\boldsymbol{r}_k^g))} \tag{7}$$

$$\boldsymbol{g}_i=\sum_{j=1}^{nm}\boldsymbol{A}_{jj}\boldsymbol{h}_j \tag{8}$$

其中，$s(\boldsymbol{w}_i,\boldsymbol{r}_i^g)$ 表示单词 $\boldsymbol{w}_i$ 在词典中匹配之后其特征 $\boldsymbol{h}_i$ 与全局特征 $\boldsymbol{r}_i^g$ 的关联，用 $\boldsymbol{A}_{ii}$ 表示新的权重值，计算出基于领域词典的全局特征表示 $\boldsymbol{g}_i$。

接下来，使用一个 tanh 层来获取单词 $\boldsymbol{w}_i$ 与文

档中其他单词相关的特征表示 $\boldsymbol{h}_i^{\text{new}}$。

$$\boldsymbol{h}_i^{\text{new}} = \tanh(\boldsymbol{W}_g[\boldsymbol{g}_i, \boldsymbol{h}_i]) \tag{9}$$

将 $\boldsymbol{h}_i^{\text{new}}$ 作为 CRF 的输入，其过程为

$$\boldsymbol{o}_i = \boldsymbol{W}\boldsymbol{h}_i^{\text{new}} \tag{10}$$

$$\text{score}(D, y) = \sum_{i=1}^{N}(\boldsymbol{o}_{i,y_i} + T_{y_{i-1},y_i}) \tag{11}$$

$$y^{\text{result}} = \text{argmax}(\text{score}(D, y)) \tag{12}$$

其中，T_{y_{i-1},y_i} 是标签 y_{i-1} 到 y_i 的转换分数；score() 函数用来计算输入文档 D 的标签序列 $y = y_1 y_2 \cdots y_N$ 的分数，y^{result} 是最终输出的标签序列结果（即 BIO 标签）；$\boldsymbol{W}$ 是模型参数。

3.2 候选词的共指消解

由于在信息安全领域缺乏大规模的可用于共指消解的标注语料，因此，本文提出一种将规则与机器学习相结合的方法来进行候选词的共指消解。需要进行共指消解的候选词包括代词以及名词短语（本文提取的实体属于名词，也划分进名词短语中）。

其中，最难解决的部分是关于代词的共指消解，它与语句的语法结构有着极大的关系[1]，因此这部分工作将采用定制规则的形式完成；关于名词短语的共指消解，则使用机器学习的方法完成。

3.2.1 代词的共指消解

通过对收集来的文本进行分析，需要进行共指消解的代词分为 2 种：第一种是关系代词；第二种是人称代词，由于信息安全领域中不存在人物作为实体，因此本文仅对第三人称代词进行消解。

1) 关系代词消解

关系代词的先行词通常在同一个句子中。对于关系代词，选择所有位于它前面的名词短语作为它的候选先行词。然后根据句子的句法分析树，提取关系代词与候选词之间的句法分析路径并选择最短路径，将距离最近的名词短语作为关系代词的先行词。举例说明如下。

例 1 Sentence: $(\text{Autoruns})_1$ revealed that there are $(\text{two core files})_2$ $(\text{Mrxcls.sys})_3$ and $(\text{Mrxnet.sys})_4$ in $(\text{the Stunex})_5$ $(\text{which})_?$ was (the first malicious code) to damage (the industry control system) in the world。

消解例 1 分析结果如表 1 所示。通过对例句进行句法分析，提取从关系代词到各候选词之间的分析路径，其中最短的一条为“NP-NP-SBAR-WHNP”，对应的候选词为“the Stunex”，则该候选词即是关系代词“which”的先行词。

表 1 消解例 1 分析结果

名称	举例
关系代词	which
候选词	$(\text{Autoruns})_1$
	$(\text{two core files})_2$
	$(\text{Mrxcls.sys})_3$
	$(\text{Mrxnet.sys})_4$
	$(\text{the Stunex})_5$
句法分析路径	NP-S-VP-SBAR-S-VP-VP-NP-SBAR-WHNP
	NP-S-VP-VP-PP-NP-SBAR-WHNP
	NP-PP-NP-SBAR-WHNP
	NP-PP-NP-SBAR-WHNP
	NP-NP-SBAR-WHNP
最短路径	NP-NP-SBAR-WHNP(the Stunex)

2) 第三人称代词消解

人称代词的先行词最有可能位于同一句或前一句。首先在同一个句子中搜索候选先行词，如果候选集为空，则从前一个句子中重新提取候选词并找到可能的先行词。由于人称代词必须指代实体，因此只保留安全领域实体候选词。如果候选集不为空，则将语法解析树从人称代词节点开始从下往上移动，如果有并列结构（包括并列名词短语、并列动词短语和并列从句），则选取第一段子结构中距离最远的候选词（按词距计算）作为人称代词的先行词；否则，从语法解析树中找到最近的子句或句子，选择其中距离最远的候选词作为先行词。举例说明如下。

例 2 Sentence：$(\text{Stuxnet})_1$ searches for $(\text{specific programs})_2$, accesses $(\text{industrial control systems})_3$,$((\text{its})_?$ attack object) is the target program development tool。

消解例 2 分析结果如表 2 所示。“its”的实体候选词为“Stuxnet”“specific programs”和“industrial control systems”，其中，有 2 个并列结构作为候选，分别是“$(\text{Stuxnet})_1$ searches for $(\text{specific programs})_2$”和“accesses $(\text{industrial control systems})_3$”，这里，“$(\text{Stuxnet})_1$ searches for $(\text{specific programs})_2$”为句子中的第一段并列结构，里面包含了 2 个实体候选词“$(\text{Stuxnet})_1$”和“$(\text{specific programs})_2$”，“$(\text{Stuxnet})_1$”距离“its”最远，即为先行词。

表 2　　消解例 2 分析结果

名称	举例
第三人称代词	its
实体候选词	$(\text{Stuxnet})_1$ $(\text{specific programs})_2$ $(\text{industrial control systems})_3$
并列结构候选	$(\text{Stuxnet})_1$ searches for $(\text{specific programs})_2$ accesses $(\text{industrial control systems})_3$
最远候选词	$(\text{Stuxnet})_1$

3.2.2　名词短语的共指消解

首先，介绍进行机器学习时需要使用到的特征向量，每个特征是通过比较 2 个待消解项之间相应的属性得来的，如下所示。

1) 所属类别是否一致。在 3.1 节抽取候选词的同时对其进行了分类，直接比较 2 个待消解项的类型是否一致，为二值属性，一致为真，不一致为假。

2) 别名和简称是否一致。如果 2 个待消解项其中一个是另外一个的别名或者简称，则值为真，反之则为假。

3) 单复数是否一致。分析 2 个待消解项后所跟动词或者系动词的形式，判断单复数是否一致，一致为真，不一致为假。

4) 2 个待消解项在文本中的距离。用 2 个待消解项在文本中所间隔的句子条数表示。

5) 名称相似性。例如短语“the virus”通常与某个病毒的名称具有相同的指代，与含有“product”“company”等词语的短语则不会。

6) 同位语。通过句法分析器可以判断出 2 个待消解项中的一个是否为另外一个的同位语，同时获取这 2 个待消解项的同位语成分。

7) 中心词的相似性。通常情况下，认为名词短语中的中心词是名词，本文通过余弦相似度来比较 2 个名词短语的中心词的相似性。

8) 结尾词的相似性。使用余弦相似度比较 2 个名词短语的最后一个单词的相似性。

接下来，进行构造训练集。假设文档中含有一条指代链 A_1-A_2-A_3-A_4，在这条链中直接相邻的指代项对（如 A_1-A_2,A_2-A_3,A_3-A_4）生成正训练样本。在这样的指代项对中，第一个名词短语通常被认为是先行词，而第二个名词短语则是后置词。负训练样本的提取如下。例如，有 B_1 和 B_2 是出现在 A_1 和 A_2 之间的其他对象，那么可以得出负训练样本为 A_1-B_1,A_1-B_2,B_1-A_2,B_2-A_2。举例说明如下。

例 3　Sentence：As the world’s first cyber “super destructive weapon”, $(\text{Stuxnet})_{A_1}$ has infected more than 45,000 networks around the world. $(\text{Computer security experts})_{B_1}$ believe $(\text{the virus})_{A_2}$ is $(\text{the highest level})_{B_2}$ $(\text{“worm”})_{B_3}$ ever. $(\text{The new virus})_{A_3}$ uses a variety of advanced technologies, so it is extremely stealthy and destructive。

例 3 中，有指代链$(\text{Stuxnet})_{A_1}$- $(\text{the virus})_{A_2}$-$(\text{The new virus})_{A_3}$，则可以生成正训练样本$(\text{Stuxnet})_{A_1}$ - $(\text{the virus})_{A_2}$和$(\text{the virus})_{A_2}$ -$(\text{The new virus})_{A_3}$。虽然指代对象间具有传递性，但是为了减少误差，这里只考虑短距离的指代关系。同样地，可以生成负训练样本$(\text{Stuxnet})_{A_1}$ - $(\text{Computer security experts})_{B_1}$、$(\text{Computer security experts})_{B_1}$- $(\text{the virus})_{A_2}$等。

共指消解的问题事实上就是一个对候选词进行分类的问题，因此，本文采用随机森林算法来解决此问题。随机森林算法是一个包含多棵决策树的分类器，易于实现，计算开销也很少。

图 2 举例说明了使用随机森林算法进行共指消解的过程。假设此时要对候选词“the virus”进行消解，算法首先根据该候选词找出某个范围内的所有可能的先行词（一般情况下选择前后相邻的 2 个句子中的所有名词短语）。选择具有最高置信度的指代链中的先行词作为该候选词的先行词。通过设置一个最小的置信阈值 t_i 来控制指代链过度生成，如果不存在置信度值大于 t_i 的指代链，则该候选词没有共指先行词（此状态可能会在后续消解过程中更改）。其中，t_i 的值可通过训练得出。

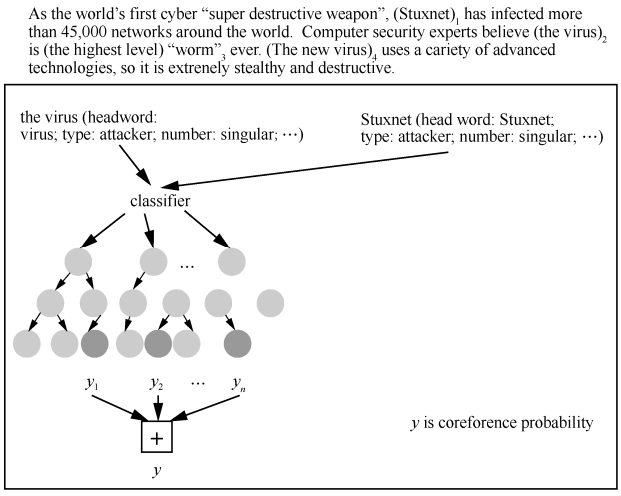


图 2　共指消解示例

4 实验及结果分析

本文验证部分进行了 4 个实验。实验 1 获取了所提方法在领域词典匹配度权重 W_d 的最佳取值；实验 2 通过与其他基准模型的比较，验证了所提方法在信息安全领域语料上进行共指消解的优越性能；实验 3 证明了单个特征对名词短语共指消解的影响程度；实验 4 验证了所提方法与其他基准模型相比在候选词提取时的优越表现，并验证了领域词典匹配机制的加入对之前工作[31]在实体提取和分类时的改进。

4.1 数据来源

本文使用的实验数据来自之前工作[31]所收集的信息安全领域文本，包括 we live security、threatpost 等处的博客文章，CVE（common vulnerabilities and exposures）描述，微软安全公告以及信息安全类文章摘要，从中摘取了 20 篇摘要、45 篇博客文章、59 段 CVE 描述以及 50 篇微软安全公告，共包含 9 123 条句子。在之前的工作中，已经对这些文本中的实体标注了类型，因此这些标注语料可作为 BiLSTM+dic_attention+CRF 模型的训练数据，从中抽取了 20 篇安全报告和 20 篇博客标注了指代链，共获取指代链 45 932 条，其中正训练样本为 7.5%。这些指代链将作为机器学习的训练数据。

由于负训练样本的数据远多于正训练样本，为了减少训练时间，本文采取了文献[19]的负样本抽取方法，具体步骤如下。

1) 使用训练数据集中所有的正训练样本，随机抽取 10%的负训练样本，进行分类器的训练。

2) 检查所有负训练样本的分类器置信值（即估计概率），只保留前 10%最模糊的负训练样本，即与正训练样本相比置信值最高的负训练样本。本文使用这些信息量更大的负训练样本和所有正训练样本来训练最终的分类器。

4.2 实验设置

设置特征向量的维度为 300，BiLSTM 中的神经单元数为 1 000，最小批次个数为 64，最大迭代次数为 100。使用文献[37]中提到的方法更新模型参数，并设置学习效率为 10^{-3}，l_2 为 10^{-5}。为了避免过拟合，本文采用 dropout 技术。BiLSTM 和 Attention 层的 dropout 值分别为 0.3 和 0.5。随机森林中的参数设置主要是对其中单个决策树的参数设置，最小置信阈值为 30%，叶子节点最小样本数设置为 5，最大深度为默认值，决策树的个数为 100。这些参数都是在训练集合中通过 10 倍交叉验证得出的。

实验是在 2 个 NVIDIA GTX 1080Ti GPU 和内存为 64 GB 的机器上完成的，模型训练时间约为 1 h。

4.3 实验结果与分析

实验 1 验证 BiLSTM-dic_attention-CRF 模型中领域词典匹配度权重 W_d 的取值。在实验 1 中，提取的信息安全领域实体包括 4 种类型，分别是“product”“vulnerability”“attacker”和“company”，以 0.05 的步长将权重值从低到高进行设置，其他参数保持不变，其表现性能如图 3 所示。其中，P 代表准确率，R 代表召回率，F 代表 F-Measure。

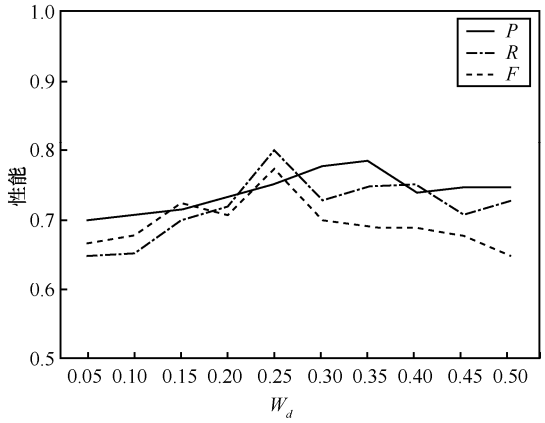


图 3 W_d 取不同值时模型的性能

从图 3 可以看出，W_d =0.25 时模型性能最优，此时 F=0.78。

实验 2 验证本文所提方法在信息安全领域共指消解的优越性。使用的 4 个基准模型为文献[19]所提的 scaffolding approach、Soon 等[17]方法、Zhang 等[23]方法、Wiseman 等[24]方法。将基准模型与本文所提模型一起应用在信息安全领域数据中，实验结果如表 3 所示。

表 3 各模型性能

方法	P	R	F
scaffolding approach[19]	63.8%	69.9%	66.7%
Soon 等[17]	60.3%	57.2%	58.7%
Wiseman 等[24]	61.2%	69.5%	65.1%
Zhang 等[23]	65.4%	68.3%	66.8%
本文所提方法	69.7%	74.2%	71.9%

从表 3 中可以看出，本文所提模型在信息安全

领域数据的表现性能要优于其他 4 个模型。通过对错误样本的分析可以看出，Wiseman 等[24]方法的主要核心在于使用 RNN 学习实体类群中每一个实体的潜在全局表示，再通过 RNN 对这些实体进行共指消解。但是本文模型并没有给出具体的聚类方式，而是默认实体已经聚类完成。于是，先利用本文所提模型将文本中需要进行共指消解的候选词提取出来，然后采用最简单的 *k*-means 聚类法对这些候选词进行聚类，但是实验后的结果并不理想。通过分析得出，在对这些实体聚类时，通常会把不具备领域特征的代词聚在一起，对这些类群学习全局特征表示时，很难学习到领域特征，这无疑会影响后续共指消解的性能。scaffolding approach[19]和 Soo 等[17]方法所处理的文本都是通用领域的文本，所制定的特征大都针对“organization”“person”等通用领域内的类型实体，因此在对信息安全领域内数据共指消解时的表现较差。Zhang 等[23]方法通过 Biaffine Attention 机制以及优化候选词提取损失函数进行共指消解，需要大量的标注训练数据来训练参数，因此虽然在 CoNLL-2012 数据集中取得了优良的表现，但是在标注数据有限的信息安全领域数据集中表现稍弱。

实验 3　针对单个特征对模型的影响进行实验。8 种特征对应的数字编号如表 4 所示。

表 4　　特征对应的数字编号

特征	编号
所属类别	1
别名和简称	2
单复数形式	3
文本距离	4
名称相似性	5
同位语	6
中心词相似性	7
结尾词相似性	8

单个特征值对模型性能的影响如图 4 所示。

从图 4 中可以看出，在所有特征中，同位语特征对共指消解性能的影响最小，主要原因是对同位语的认定比较复杂，对于一些语法结构相对比较复杂的语句，仅依靠句法分析工具判定句子中的同位语准确率并不高。此外，中心词相似性及别名和简称特征对共指消解性能的影响最高，主要原因分析发现如下：1) 在信息安全领域的文本中含有很多领域内的专业术语及简称，例如，Advanced Persistent Threat 简称 APT；2) 对名词短语来说，中心词往往决定了该短语的主要含义。

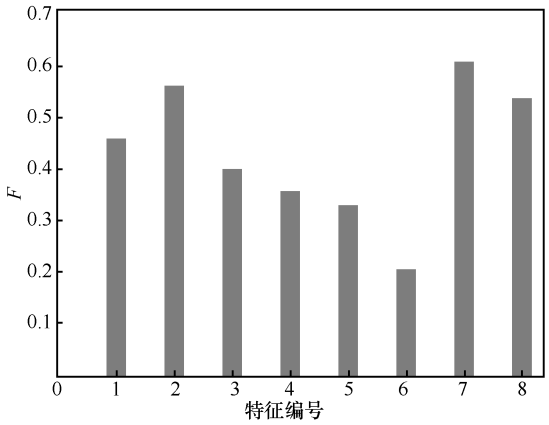


图 4　单个特征值对模型性能影响

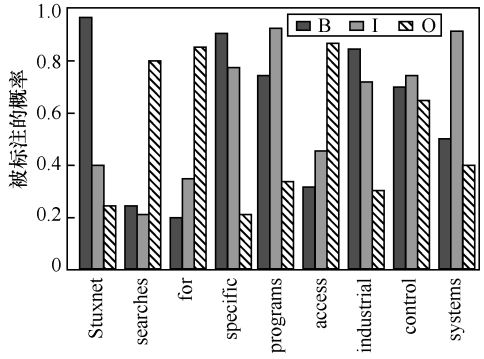
实验 4　对文本中候选词提取的效果也会影响共指消解的效果，因此验证本文所提的将规则与神经网络模型（BiLSTM-dic_attention-CRF）相结合进行提取候选词的方法（简称 rule-based BiLSTM-dic_attention-CRF）的性能。这里使用的基准模型包括实验 2 中的 3 个基准模型，以及本文之前工作[31]中的方法，此处用 previous work 表示。实验 4 提取的信息安全领域实体包括 4 种类型，分别是“product”“vulnerability”“attacker”和“company”。实验结果如表 5 所示。

表 5　　各模型提取候选词的性能表现

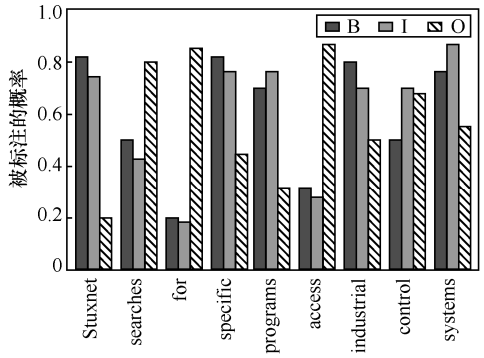
方法	*P*	*R*	*F*
Zhang 等[23]	69.3%	65.9%	67.5%
scaffolding approach[19]	65.1%	63.8%	64.4%
Soon 等[17]	61.7%	55.4%	58.3%
previous work[31]	70.1%	65.2%	67.6%
本文所提方法	75.4%	80.1%	77.7%

从表 5 中可以看出，本文所提方法（rule-based BiLSTM-dic_attention-CRF）在信息安全领域数据上的性能优于目前已知的提取候选词方法，主要原因在于，本文所提方法除了依赖深度学习之外，还通过对信息安全领域中的数据文本进行分析，总结出了一套与之相契合的提取规则，两者结合在一起才能达到较好的效果。图 5 展示了 BiLSTM-dic_ttention- CRF 模型和本文之前的模型（previous work）对例句“Stunex searches for specific programs access industrial control

systems”的实体提取分类效果。其中，B、I、O 表示的是 BIO 标注，如果单词被标注为 B，表示该单词为某个片段的开头，同理，I 表示该单词在片段的中间位置，O 表示不属于任何类型。



(a) BiLSTM_dic_attention-CRF



(b) previous work

图 5 2 种模型对例句的实体提取分类效果

基于文档层面特征向量在进行实体提取工作时的优越性已经在本文之前的工作[31]中进行了验证，此处不再赘述。

通过对 rule-based BiLSTM-dic_attention-CRF 提取出的错误结果进行分析，本文方法依然存在着候选词提取边界过长、先行词中单词缺失以及候选词为无用候选词（即无共指关系）等问题，有待于进一步解决。

5 结束语

本文提出了一种混合方法来解决信息安全领域内的共指消解任务中的 2 个问题：1) 从给定文档中提取出所有的候选词语并进行分类；2) 筛选出符合条件的待消解项进行共指消解。本文针对信息安全领域内文本的特点制定出一套规则并与深度学习模型（BiLSTM-dic_attention-CRF）相结合来解决对文本中候选词语的提取和分类问题，将共指消解分解成代词的共指消解和名词短语的共指消解，代词消解通过规则完成，而名词短语的共指消解通过机器学习完成。实验证明，本文所提方法相较于其他基于通用领域构造的模型，在信息安全领域上的应用性能更加优越。

参考文献：

[1] SUKTHANKER R, PORIA S, CAMBRIA E, et al. Anaphora and coreference resolution: a review[J]. arXiv Preprint, arXiv:1805.11824, 2018.

[2] VASWANI A, BENGIO S, BREVDO E, et al. Tensor2tensor for neural machine translation[J]. arXiv Preprint, arXiv:1803.07416, 2018.

[3] LAMPLE G, OTT M, CONNEAU A, et al. Phrase-based & neural unsupervised machine translation[J]. arXiv Preprint, arXiv:1804.07755, 2018.

[4] CHEN M X, FIRAT O, BAPNA A, et al. The best of both worlds: Combining recent advances in neural machine translation[J]. arXiv Preprint, arXiv:1804.09849, 2018.

[5] CAMBRIA E, PORIA S, HAZARIKA D, et al. SenticNet 5: discovering conceptual primitives for sentiment analysis by means of context embeddings[C]//Thirty-Second AAAI Conference on Artificial Intelligence. 2018: 1795-1802.

[6] ETTER M, COLLEONI E, ILLIA L, et al. Measuring organizational legitimacy in social media: assessing citizens' judgments with sentiment analysis[J]. Business & Society, 2018, 57(1): 60-97.

[7] MA Y, PENG H, CAMBRIA E. Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive LSTM[C]//Thirty-Second AAAI Conference on Artificial Intelligence. 2018: 5876-5883.

[8] ZENG D, DAI Y, LI F, et al. Adversarial learning for distant supervised relation extraction[J]. Computers, Materials & Continua, 2018, 55(1): 121-136.

[9] GÁBOR K, BUSCALDI D, SCHUMANN A K, et al. Semeval-2018 task 7: Semantic relation extraction and classification in scientific papers[C]//The 12th International Workshop on Semantic Evaluation. 2018: 679-688.

[10] QIN P, XU W, WANG W Y. DSGAN: generative adversarial training for distant supervision relation extraction[J]. arXiv Preprint, arXiv:1805.09929, 2018.

[11] LIU F, FLANIGAN J, THOMSON S, et al. Toward abstractive summarization using semantic representations[J]. arXiv Preprint, arXiv:1805.10399, 2018.

[12] CHEN Y C, BANSAL M. Fast abstractive summarization with reinforce-selected sentence rewriting[J]. arXiv Preprint, arXiv:1805.11080, 2018.

[13] HOBBS J R. Resolving pronoun references[J]. Lingua, 1978, 44(4): 311-338.

[14] BRENNAN S E, FRIEDMAN M W, POLLARD C J. A centering approach to pronouns[C]//The 25th Annual Meeting on Association for Computational Linguistics. 1987: 155-162.

[15] LAPPIN S, LEASS H J. An algorithm for pronominal anaphora resolution[J]. Computational Linguistics, 1994, 20(4): 535-561.

[16] LEE H, CHANG A, PEIRSMAN Y, et al. Deterministic coreference resolution based on entity-centric, precision-ranked rules[J]. Computa-

tional Linguistics, 2013, 39(4): 885-916.

[17] SOON W M, NG H T, LIM D C Y. A machine learning approach to coreference resolution of noun phrases[J]. Computational Linguistics, 2001, 27(4): 521-544.

[18] AONE C, BENNETT S W. Evaluating automated and manual acquisition of anaphora resolution strategies[C]//The 33rd Annual Meeting on Association for Computational Linguistics.1995: 122-129.

[19] LEE H, SURDEANU M, JURAFSKY D. A scaffolding approach to coreference resolution integrating statistical and rule-based models[J]. Natural Language Engineering, 2017, 23(5): 733-762.

[20] 钱伟, 郭以昆, 周雅倩, 等. 基于最大熵模型的英文名词短语指代消解[J]. 计算机研究与发展, 2003, 40(9): 1337-1343.
QIAN W, GUO Y K, ZHOU Y Q, et al. English noun phrase coreference resolution via a maximum entropy model[J]. Journal of Computer Research and Development,2003, 40(9): 1337-1343.

[21] WISEMAN S, RUSH A M, SHIEBER S, et al. Learning anaphoricity and antecedent ranking features for coreference resolution[C]//The 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. 2015: 1416-1426.

[22] LEE K, HE L, LEWIS M, et al. End-to-end neural coreference resolution[J]. arXiv Preprint, arXiv:170707045, 2017.

[23] ZHANG R, SANTOS C N, YASUNAGA M, et al. Neural coreference resolution with deep biaffine attention by joint mention detection and mention clustering[J]. arXiv Preprint, arXiv:1805.04893, 2018.

[24] WISEMAN S, RUSH A M, SHIEBER S M. Learning global features for coreference resolution[J]. arXiv Preprint, arXiv:160403035, 2016.

[25] CLARK K, MANNING C D. Deep reinforcement learning for mention-ranking coreference models[C]//The 2016 Conference on Empirical Methods in Natural Language Processing. 2016: 2256-2262.

[26] DODDINGTON G R, MITCHELL A, PRZYBOCKI M A, et al. The automatic content extraction (ACE) program-tasks, data, and evaluation[C]//LREC. 2004: 1.

[27] PRADHAN S, MOSCHITTI A, XUE N, et al. Conll-2012 shared task: modeling multilingual unrestricted coreference in ontonotes[C]//Joint Conference on EMNLP and CoNLL-Shared Task, Association for Computational Linguistics. 2012: 1-40.

[28] GUILLOU L, HARDMEIER C, SMITH A, et al. Parcor 1.0: a parallel pronoun-coreference corpus to support statistical mt[C]//9th International Conference on Language Resources and Evaluation (LREC). 2014: 3191-3198.

[29] HAGHIGHI A, KLEIN D. Simple coreference resolution with rich syntactic and semantic features[C]//The 2009 Conference on Empirical Methods in Natural Language Processing. 2009: 1152-1161.

[30] BBN Technologies. 2006. Coreference Guidelines for English OntoNotes – Version 6.0.

[31] 张晗, 郭渊博, 李涛. 结合 GAN 与 BiLSTM-Attention-CRF 的领域命名实体识别[J]. 计算机研究与发展, 2019, 56(9): 1851-1858.
ZHANG H, GUO Y B, LI T. Domain named entity recognition combining GAN and BiLSTM-Attention-CRF[J]. Journal of Computer Research and Development, 2019, 56(9): 1851-1858.

[32] PUSTEJOVSKY J, CASTANO J, SAURI R, et al. (2002) Medstract: creating large-scale information servers for biomedical libraries[C]//The ACL-02 Workshop on Natural Language Processing in the Biomedical Domain. 2002: 85-92.

[33] SU J, YANG X, HONG H, et al. Coreference resolution in biomedical texts: a machine learning approach[C]//Dagstuhl Seminar Proceedings. 2008.

[34] D'SOUZA J, NG V. Anaphora resolution in biomedical literature: a hybrid approach[C]//ACM Conference on Bioinformatics. ACM, 2012: 113-122.

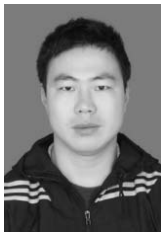
[35] 韩旭. 基于神经网络的文本特征表示关键技术研究[D]. 北京: 北京邮电大学, 2019.
HAN X. Research on key technologies of text feature representation based on neural network[D]. Beijing: Beijing University of Posts and Telecommunications, 2019.

[36] ZHANG H, GUO Y B, LI T. Multifeature named entity recognition in information security based on adversarial learning[J]. Security and Communication Networks, 2019, 2019(2): 1-9.

[37] KINGMA D P, BA J. Adam: a method for stochastic optimization[J]. arXiv Preprint, arXiv: 1412.6980, 2014.

[作者简介]

张晗（1985− ），女，河南项城人，信息工程大学博士生，主要研究方向为自然语言处理、信息安全。



胡永进（1981− ），男，山东潍坊人，信息工程大学讲师，主要研究方向为主动防御、态势感知。

郭渊博（1975− ），男，陕西周至人，博士，信息工程大学教授、博士生导师，主要研究方向为大数据安全、态势感知。

陈吉成（1984− ），男，江苏涟水人，信息工程大学博士生，主要研究方向为复杂网络、信息内容安全。