doi:10.3969/j.issn.1673-9833.2018.04.010

基于医疗类别的电子病历命名实体识别研究

李 飞[1,2]，朱艳辉[1,2]，王天吉[1,2]，徐 啸[1,2]，冀相冰[1,2]

（1. 湖南工业大学 计算机学院，湖南 株洲 412007；2. 湖南省智能信息感知及处理技术重点实验室，湖南 株洲 412007）

摘 要：基于电子病历命名实体识别对智慧医疗和医疗知识图谱的构建具有重要意义，提出一种基于医疗类别的命名实体识别方法。首先，针对电子病历语料中实体特点进行深度挖掘，将电子病历分为 4 类医疗类别；然后，对各医疗类别分别构建特征集，并使用条件随机场模型对身体部位、症状和体征、检查与检验、疾病与诊断、治疗等 5 类命名实体进行命名实体识别；最后，将基于医疗类别特征集识别效果和通用特征集的识别结果进行对比。实验结果表明，基于医疗类别的电子病历命名实体识别效果显著提升，可以满足应用需求。

关键词：电子病历；命名实体识别；条件随机场；医疗类别

中图分类号：TP391 **文献标志码**：A **文章编号**：1673-9833(2018)04-0061-06

Research on Electronic Medical Record Named Entity Recognition Based on Medical Categories

LI Fei[1,2]，ZHU Yanhui[1,2]，WANG Tianji[1,2]，XU Xiao[1,2]，JI Xiangbing[1,2]

（1. College of Computer，Hunan University of Technology，Zhuzhou Hunan 412007， China；
2. Hunan Key Laboratory of Intelligent Information Perception and Processing Technology，Zhuzhou Hunan 412007， China）

Abstract：Based on the named entity recognition in electronic medical records is of great significance to medical treatment AI and the construction of medical knowledge graph，a proposal has been made of a named entity recognition method based on medical categories. First, the electronic medical record is to be divided into 4 categories according to the entity characteristics of the corpus of electronic medical records. Then, the feature sets are to be constructed respectively for the medical categories, followed by an identification of the named entities of such five named entities as body parts, symptoms and signs, inspection and test, disease and diagnosis, and treatment by using the conditional random field model. Finally, a comparison has been made between the recognition results based on medical class feature sets and the general feature sets. The results show that the effect of named entity recognition based on medical categories has been significantly improved, enabling it to meet the application requirement effectively.

Keywords：electronic medical record；named entity recognition；conditional random field；medical category

收稿日期：2017-12-20

基金项目：国家自然科学基金资助项目（61402165），湖南省教育厅基金资助重点项目（15A049），国家工商行政管理总局科研基金资助项目（2014GSZJWT001KT006），湖南工业大学科研基金资助重点项目（17ZBLWT001KT006），湖南省研究生科研创新基金资助项目（CX2017B688）

作者简介：李 飞（1992-），男，安徽宿州人，湖南工业大学硕士生，主要研究方向为自然语言处理，E-mail：flytoskye@163.com

通信作者：朱艳辉（1968-），女，湖南株洲人，湖南工业大学教授，硕士生导师，主要从事文本处理与知识工程方面的教学与研究，E-mail:swayhzhu@163.com.

1 研究背景

随着医疗信息化的快速发展，电子病历（eclectic medical record，EMR）已经在临床中开始普及，由此产生了大量的病历资料。电子病历[1]是指产生于医务人员在医疗活动过程中，由医务人员撰写，面向患者个体描述医疗活动的记录。电子病历中包含了大量的医疗知识与信息，很多临床辅助决策系统都将电子病历作为重要的知识来源。然而，计算机只能处理结构化的数据。因此，如何让计算机理解以自然语言形式存在的电子病历信息，构建并挖掘大规模医疗知识库以支持医疗人工智能（artificial intelligence，AI）事业的发展，现已成为医疗知识信息化建设亟待解决的问题。

命名实体识别（named entity recognition，NER）作为信息抽取的子任务，是指将非结构化文本中具有特定意义的实体抽取出来，这对于文本的结构化起着至关重要的作用。我国电子病历命名实体识别研究相较发达国家起步较晚。电子病历作为医疗领域的知识载体，其语言风格具有随意性、专业性、句法不完整等特点，且病历中的命名实体繁多，导致识别结果差强人意。

近年来，随着医疗 AI 的发展，许多机构组织了大量针对临床文本的 NER 任务，其中 I2B2[2]（the Center of Informatics for Integrating Biology and the Bedside）会议组织方提供了大量的标注语料，由此催生了一些识别效果较好的机器学习算法，如隐形马尔科夫模型（hidden Markov model，HMM）、条件随机场（conditional random field，CRF）等。国内的全国知识图谱与语义计算大会[3]（China Conference on Knowledge Graph and Semantic Computing，CCKS）组委会也迎头赶上，组织了中文电子病历命名实体识别竞赛。由于中文相较于英文没有明显的字符边界，不具备良好的分词标识，加之电子病历的领域特殊性，存在句法结构不完整、构词模式复杂、实体嵌套严重、专业性较强等特点，所以电子病历的命名实体识别研究仍具有一定的挑战性，吸引了较多科研工作者的关注。王云吉[4]提出了一种基于层叠条件随机场的电子病历命名实体识别方法，建立层叠的 CRFs 框架，将复杂的电子病历分为相对简单、相互关联的子层后，分别进行实体识别，取得了较好的识别效果。王世昆等[5]利用 CRFs 和支持向量机（support vector machine，SVM）等方法对中医医案中的症状、病机的自动识别标注问题进行了研究。曲春燕[6]对哈尔滨医科大学附属第二医院的电子病历进行了标注，并且使用最大熵模型（maximum entropy model，ME）、结构化支持向量机（structured support vector machine，SSVM）、条件随机场（conditional random fields，CRF）以及组合算法，完成了对电子病历命名实体的识别。

本文通过深度挖掘电子病历语料中的实体信息，发现不同的医疗记录中实体的特征及类型存在差异。如电子病历中的门诊记录多包含身体部位、症状等实体，出院记录多包含药物、症状等实体。因此，本文通过对电子病历中实体及其特征进行进一步挖掘，提出一种基于医疗类别的命名实体识别方法。并且采用 CRF 模型对身体部位、症状和体征、检查与检验、疾病与诊断、治疗 5 类实体进行命名实体识别。实验结果表明，基于医疗类别的命名实体识别取得了良好的效果。

2 CRF 理论

条件随机场 CRF 是给定一组输入随机变量条件下另一组输出随机变量的条件概率分布模型，该模型不仅克服了马尔科夫模型的缺点，而且解决了最大熵模型存在的标记偏移问题，现已经被广泛地运用于自然语言处理序列标注领域中[7]。其中，适用于序列标注并且应用较为广泛的是线性链条件随机场，其结构如图 1 所示。

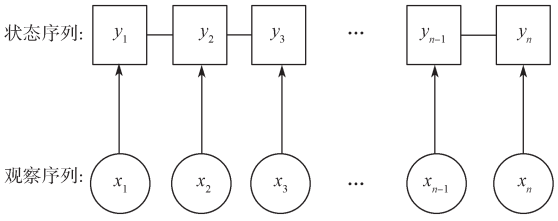

图 1 线性链 CRF 结构

Fig. 1 Liner chain CRF structure

条件随机场的参数化形式可描述如下：对于观察序列 $x=(x_1, x_2, \cdots, x_n)$ 和状态序列 $y=(y_1, y_2, \cdots, y_n)$，设 $P(y|x)$ 为线性链条件随机场，则在随机变量 X 取值为 x 的条件下，随机变量 Y 取值为 y 的条件概率形式如下：

$$P(y \mid x)=\frac{1}{Z(x)}\exp\left\{\sum_{i,k}\lambda_k f_k\left(y_{i-1}, y_i, x, i\right)+\sum_{i,l}\mu_l h_l\left(y_i, x, i\right)\right\}, \tag{1}$$

其中，

$$Z(x)=\sum_{y}\exp\left\{\sum_{i,k}\lambda_k f_k\left(y_{i-1}, y_i, x, i\right)+\sum_{i,l}\mu_l h_l\left(y_i, x, i\right)\right\}。 \tag{2}$$

式（1）~（2）中：f_k、h_l 为特征函数；

λ_k、μ_l 为对应的权值；

$Z(x)$ 为归一化因子。

命名实体识别过程就是序列标注过程，即先将句子看作一个观察序列，把句中每个字符或者词看作一个符号；然后，给各符号赋予一个状态；接下来通过训练集进行最大化参数 λ_k、μ_l，即可以得到满足条件的条件概率。

对于一个输入序列，最可能的输出标记序列（也即最佳状态序列）为

$$\hat{y} = \arg\max P(y \mid x) \text{ 。}$$

3　特征设计

本文采用 2017 年全国知识图谱与语义计算大会任务二的评测语料，通过对其进行深度挖掘，观察到不同的医疗记录所产生的实体类型及特征具有一定的差异性。电子病历中实体统计信息见表 1。

表 1　实体统计信息

Table 1　Entity statistics information　　例

医疗类别	实　体				
	身体部位	症状和体征	检查和检验	疾病和诊断	治疗
门诊记录	181	375	1	72	0
病史记录	4 720	3 423	3 298	300	80
诊疗记录	590	437	239	164	538
出院记录	3 290	2 118	2 849	4	8

由表 1 可以得知，门诊记录病历中的大部分实体类型为身体部位、症状和体征；病史记录病历中的大部分实体类型为身体部位、症状和体征、检查和检验；诊疗记录病历中的大部分实体类型为身体部位、症状和体征、治疗；出院记录中的大部分实体类型为身体部位、症状和体征、检查和检验。进一步对电子病历语料中的实体数量和平均长度进行统计，所得结果如表 2 所示。

表 2　实体总数及平均长度统计

Table 2　Statistics of the total number of entities and the average lengths

实体名称	总计 / 例	实体平均长度 / 字节
身体部位	8 781	2.80
症状和体征	6 353	2.20
检查和检验	6 387	3.22
疾病和诊断	540	5.76
治疗	626	5.97

结合上述结论，本文进一步统计了各医疗记录中包含的实体数量及类型。根据实体类型所占比例，结合各实体平均长度计算各医疗类别中的实体长度，计算方式见式（4）。

$$L_j = \sum_i len(j,i) * \frac{\lambda_{j,i}}{\sum_{j,i} \lambda_{j,i}} \text{ 。} \quad (4)$$

式中：L_j 为各医疗类别中实体平均长度；

$len(j, i)$ 为各实体类型平均长度；

$\lambda_{j,i}$ 为医疗类别对应的各实体类型数量，$i \in (1, 2, 3, 4, 5)$，$j \in (1, 2, 3, 4)$。

通过计算，得到表 3 所示医疗类别中实体类型及长度统计结果。

表 3　医疗类别中实体类型及长度统计结果

Table 3　Statistics of entity types and lengths in medical categories

医疗类别	主要实体类型	实体平均长度 / 字节
门诊记录	身体部位、症状和体征	2.76
病史特点	身体部位、症状与体征、检查与检验	2.80
诊疗经过	身体部位、症状和体征、治疗	3.80
出院情况	身体部位、症状与体征、检查与检验	2.79

基于以上论述，本文针对不同医疗类别中所包含实体类型和实体平均长度的差异，提出一种基于医疗类别的特征抽取策略，设计了语言特征、上下文特征、外部实体词典特征和实体边界特征，并结合医疗类别中实体差异性构成针对各医疗类别的特征集。

1）语言特征

语言特征可以反映出字符的基本信息，是一种基本特征。中文命名实体识别任务中常用基于字粒度和词粒度两种方法，由于电子病历语言的随意性和自由性，对电子病历分词会出现分词错误，最终导致实体无法被识别，而字粒度相较于词粒度包含更多的实体内部结构等信息，可以进一步提高识别效果，故本文采用字粒度作为语言特征。字粒度的语言特征如表 4 所示。

表 4　字粒度语言特征表示

Table 4　Representation of word size language features

编号	特　征	描 述
1	Character（−2）	前两个字符
2	Character（−1）	前一个字符
3	Character（0）	当前字符
4	Character（1）	后一个字符
5	Character（2）	后两个字符

2）上下文特征

上下文特征是指实体词汇窗口长度内观测值之间的相互依赖关系，该特征可以很好地刻画出实体内部的依赖关系以及实体与非实体的相互关系。结合表

3 对各医疗类别中实体特点进行分析，本研究针对不同的医疗类别选择不同的实体窗口构成上下文特征，各医疗类别实体窗口大小具体如表 5 所示。

表 5 各医疗类别实体窗口大小

Table 5 Sizes of medical category entity windows

医疗类别	实体平均长度 / 字节	窗口大小
门诊记录	2.76	3
病史记录	2.80	3
诊疗记录	3.80	4
出院记录	2.79	3

3）外部实体词典特征

由于电子病历语言的专业性，有必要引入实体词典。本文通过对训练语料中命名实体进行提取，并加入搜狗诊断学词库 [8] 和医院电子病历词库构成外部实体词典特征，具体如表 6 所示。

表 6 实体词典来源及大小

Table 6 Source and size of the entity dictionary

词典名称	词典来源	词典大小 / 个
医院电子病历词库	搜狗词库	2 753
诊断学词库	搜狗词库	339
语料病历词典	训练语料实体	22 687

4）实体边界特征

实体边界特征是确定字符边界特征位置信息的重要依据，确定命名实体的边界对命名实体识别起着至关重要的作用。本文采用 BIO 编码模式描述观测序列的词边界特征，其中 B 表示实体的开头，I 表示实体的剩余部分，O 表示非实体序列，并且对实体类型进行编码。表 7 给出了本文命名实体的边界特征及其类型编码。

表 7 边界特征及实体类型编码集

Table 7 Boundary feature and entity type coding sets

编码	含 义	编码	含 义
B-body	身体部位的开头	B-dis	疾病和诊断的开头
I-body	身体部位的剩余部分	I-dis	疾病和诊断的剩余部分
B-sas	症状和体征的开头	B-tre	治疗的开头
I-sas	症状和体征的剩余部分	I-tre	治疗的剩余部分
B-che	检查和检验的开头	O	非实体序列
I-che	检查和检验的剩余部分		

4 实验设计与结果分析

4.1 实验流程与数据选择

本实验采用 2017 年 CCKS 评测任务二的语料作为实验数据，CCKS 的电子病历语料不仅已做了去隐私处理，而且详细标注了实体、实体类型等信息。基于条件随机场的开源工具有 CRF++[9]、FlexCRF 等，本实验采用 CRF++ 作为模型训练和预测的工具。电子病历实体抽取的实验流程如图 2 所示。

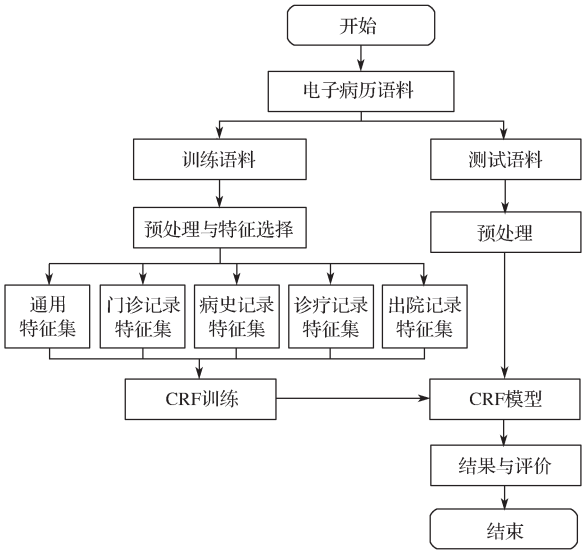

图 2 电子病历实体抽取实验流程图

Fig. 2 Electronic medical record entity extraction experiment flow chart

4.2 特征模板设计

本研究中，采用 CRF++ 工具进行特征模板的设计。CRF++ 工具需要利用用户制定的模板文件（Template File）对训练语料进行训练，模板文件的每 1 行代表 1 个特征模板（Feature Template），从而可确定训练数据中的一个 Token，一维特征模板的基本格式为 %X[row，col]。其中，row 表示与当前 Token 的相对行数，col 表示绝对列数。本实验根据医疗记录的实体长度差异，为 4 类医疗记录分别设计上下文本特征。然后，结合其他文本特征构成组合特征集，并结合外部实体词典分析实验结果。本实验设定的基本特征模板如表 8 所示。

表 8 基本特征模板

Table 8 Template of basic features

特征标识	特征描述
U00:%x[-2, 0]	当前字的前两个字
U01:%x[-1, 0]	当前字的前一个字
U02:%x[0, 0]	当前字
U03:%x[0, 1]	当前字的后一个字
U04:%x[0, 2]	当前字的后两个字
U05:%x[-1, 0]/%x[0, 0]	当前字和前一个字的组合
U06:%x[0, 0]/%x[0, 1]	当前字和后一个字的组合
U07:%x[-1, 0]/%x[0, 0]/%x[1, 0]	当前字和前后一个字的组合
U08:%x[-2, 0]/%x[-1, 0]/%x[0, 0]	当前字和前两个字的组合
U09:%x[0, 0]/%x[1, 0]/%x[2, 0]	当前字和后两个字的组合

4.3　实验设计

本实验采用 CCKS 提供的电子病历命名实体识别语料合并训练集和测试集，并且按照 7:3 的比例重新对其进行划分。最后，以 conlleval.pl[10] 脚本文件比较真实结果与生成序列之间的差异，并对识别结果进行评测。

为了验证医疗类别对电子病历命名实体识别效果的影响，本研究设计了 5 组实验，通过对训练集中实体的统计分析可以得知，总体数据集中所有实体的平均长度约为 4，故本文将所有训练数据不区分类别，选取以 4 为上下文窗口特征，结合其余特征进行实验，作为 Baseline。对照组实验设计为各医疗类别分别选取不同的上下文特征，并结合其余特征构成特征集进行实验。本文 5 次实验设计如表 9 所示。

表 9　实验设计说明

Table 9　Experimental design description

编号	实验名称	实验说明
实验 1	Baseline	合并所有训练语料，选取本文特征及窗口为 4 的上下文特征
实验 2	门诊记录	选择门诊记录类别，选取本文特征及窗口为 3 的上下文特征
实验 3	病史记录	选择病史记录类别，选取本文特征及窗口为 3 的上下文特征
实验 4	诊疗记录	选择诊疗记录类别，选取本文特征及窗口为 4 的上下文特征
实验 5	出院记录	选择出院记录类别，选取本文特征及窗口为 3 的上下文特征

4.4　实验结果分析

4.4.1　评价标准

本研究的评价标准采用信息检索通用评价方法，评价指标包括准确率 P、召回率 R 和 F 值，各指标具体定义如下：

$$\begin{cases} P=\dfrac{\text{正确识别的实体个数}}{\text{识别出的实体总数}}\times 100\%, \\ R=\dfrac{\text{正确识别的实体个数}}{\text{标准结果中的实体个数}}\times 100\%, \\ F=\dfrac{2\times P\times R}{P+R}。 \end{cases} \tag{5}$$

4.4.2　结果分析

在实验 1 中，合并所有训练语料，根据实体平均长度，取窗口为 4 的上下文特征、语言特征、实体边界特征及实体词典构成特征集，将得出的实验结果作为 Baseline。在实验 2~5 中，由表 3 中实体长度信息，分别为 4 组实验选取各自的上下文窗口特征，并结合其余特征构成特征集。所得实验结果如表 10 所示。

表 10　不同医疗类别各类实体指标

Table 10　Physical indicators of different medical categories

%

医疗类别	指标	身体部位	检查与检验	疾病与诊断	症状与体征	治疗	平均
Baseline	P	84.02	93.51	72.65	90.77	75.66	87.77
	R	78.36	92.10	73.28	94.19	57.58	84.10
	F	81.09	92.80	72.96	92.45	65.39	85.89
门诊记录	P	92.68	0.00	93.11	96.89	0.00	94.29
	R	87.36	0.00	86.04	88.00	0.00	87.13
	F	90.03	0.00	89.44	92.23	0.00	90.57
病史记录	P	89.74	97.27	87.16	99.23	76.00	94.56
	R	79.50	91.16	77.45	94.50	33.04	86.13
	F	84.31	94.11	82.01	96.80	46.06	90.15
诊疗记录	P	87.04	89.69	94.88	90.16	97.80	90.69
	R	85.49	90.89	93.88	87.30	97.80	89.33
	F	86.26	90.29	94.38	88.71	97.80	90.04
出院记录	P	91.11	95.95	0.00	94.78	0.00	93.90
	R	88.09	95.07	0.00	97.39	0.00	93.12
	F	89.57	95.51	0.00	96.07	0.00	93.51

分析表 10 可知，4 个医疗类别实体的识别结果相较于 Baseline 均有一定程度的提高。而门诊记录与出院记录中存在指标为 0.00 的情况。分析训练语料，可知是由于门诊记录中包含较少的检查与检验实体和治疗实体；出院记录中疾病与诊断、治疗也接近于 0，这是因为测试集中没有对应的实体类别，导致出现指标为 0.00 的实验结果。各个医疗类别针对实体平均长度选取合适的上下文窗口，均可得到较好的结果。由此说明：当上下文窗口特征的选择与实体长度相近时，实体识别效果优于统一选取上下文窗口的。在诊疗记录中，由于选取了实体窗口为 4 的上下文特征，所以该模型对疾病与诊断和治疗识别率较其他 3 类实体相对较高。这一结果进一步验证了当实体窗口上下文特征与实体长度相近时，识别效果最好。

综合上述结论可知，先针对电子病历中实体特点进行分析，对电子病历进行医疗记录分类，再设计特征集进行实体抽取，可以有效地提高电子病历命名实体识别率。

5　结语

本文通过进一步挖掘电子病历语料中不同医疗类别之间的实体特点、构成等信息，提出一种基于医疗类别的电子病历命名实体识别方法。实验结果表明，对医疗类别的实体信息进行深度挖掘可以有效提高电子病历实体识别效果。但是本文的研究还存在一些不足之处：

1）电子病历实体词典仅搜集互联网上现存的词汇，还应进一步扩充；

2）基于深度学习的命名实体识别方法已被证明优于传统机器学习算法。

接下来将对算法进行深入研究，以进一步提高电子病历命名实体识别的效果。

参考文献：

[1] 杨锦锋，于秋滨，关 毅，等 . 电子病历命名实体识别和实体关系抽取研究综述 [J]. 自动化学报，2014，40(8)：1537−1562.
YANG Jinfeng，YU Qiubin，GUAN Yi，et al. An Overview of Research on Electronic Medical Record Oriented Named Entity Recognition and Entity Relation Extraction[J]. Acta Automatica Sinica，2014，40(8)：1537−1562.

[2] I2B2. The Center of Informatics for Integrating Biology and the Bedside[EB/OL].（2010−05−25）[2017−11−15]. https://www.i2b2.org/events/index.html.

[3] 肖仰华 . CCKS 2017− 全国知识图谱与语义计算大会 [EB/OL].（2017−04−17） [2017−11−15]. http://www.ccks2017.com/?page_id=51.
XIAO Yanghua. CCKS: China Conference on Knowledge Graph and Semantic Computing in 2017[EB/OL].（2017−04−17）[2017−11−15]. http://www.ccks2017.com/?page_id=51.

[4] 王云吉 . 基于层叠条件随机场的电子病历命名实体识别 [D]. 长春：吉林大学，2011.
WANG Yunji. Recognition of Name Entity in Electronic Medical Records Based on Cascaded Random Fields[D]. Changchun：Jilin University，2011.

[5] 王世昆，李绍滋，陈彤生 . 基于条件随机场的中医命名实体识别 [J]. 厦门大学学报（自然科学版），2009，48(3)：359−364.
WANG Shikun，LI Shaozi，CHEN Tongsheng. Research on TCM Named Entity Recognition Based on Conditional Random Fields[J]. Journal of Xiamen University（Natural Science），2009，48(3)：359−364.

[6] 曲春燕 . 中文电子病历命名实体识别研究 [D]. 哈尔滨：哈尔滨工业大学，2015.
QU Chunyan. Chinese Electronic Medical Record Named Entity Recognition[D]. Harbin：Harbin Institute of Technology，2015.

[7] 李 航 . 统计学习方法 [M]. 北京：清华大学出版社，2012：191−200.
LI Hang. Statistical Learning Methods[M]. Beijing：Tsinghua University Press，2012：191−200.

[8] 医学医药词库 . 搜狗词库 [EB/OL].（2017−10−13）[2017−11−15]. http://pinyin.sogou.com/dict/cate/index/132?rf=dictindex.
Medical Medicine Thesaurus. Sougou Thesaurus[EB/OL].（2017−10−13）[2017−11−15] . http://pinyin.sogou.com/ dict/cate/index/132?rf=dictindex.

[9] Anon. CRF++:Yet Another CRF Toolkit[EB/OL].（2013−02−13） [2017−11−15]. https://taku910.github.io/crfpp/.

[10] Anon. Conlleval.pl[EB/OL].（2013−01−16）[2017−11−15]. https://github.com/tpeng/npchunker/blob/master/conlleval.pl.

（责任编辑：廖友媛）