基于多步聚类的汉语命名实体识别和歧义消解 *

李广一，王厚峰

北京大学计算语言学教育部重点实验室，北京 100871

北京大学计算语言学研究所，北京 100871

E-mail: {liguangyi, wanghf}@pku.edu.cn

摘　要：命名实体识别和歧义消解是自然语言理解的重要研究内容。针对提供实体知识库情况下的命名实体识别和歧义消解任务，本文提出了一种基于多步聚类的方法。首先通过两轮聚类将命名实体与知识库实体定义链接，然后通过层次聚合式聚类对知识库中未出现的实体进行聚类，最后进行普通词的识别和基于 K-Means 聚类的结果调整。在 CLP-2012 的汉语命名实体识别和歧义消解评测数据上的实验表明，本文的方法表现出良好的性能，在测试集上的 F 值高出评测参赛队伍最好水平 6.46%，达到 86.68%。

关键词：命名实体识别；命名实体消歧；聚类

Chinese Named Entity Recognition and Disambiguation Based on Multi-stage Clustering

Li Guangyi, Wang Houfeng

Key Laboratory of Computational Linguistics (Peking University), Ministry of Education, China

Institute of Computational Linguistics, Peking University, Beijing 100871

E-mail: {liguangyi, wanghf}@pku.edu.cn

Abstract: Named Entity Recognition and Disambiguation is an important research of Natural Language Understanding. For the task of Named Entity Recognition and Disambiguation in the situation of entity knowledge base provided, this paper presents a method based on multi-stage clustering. First, we link the document to the entity definition in the knowledge base by two rounds of clustering. Second, we group entities which don't exist in the knowledge base by Hierarchical Agglomerative Clustering. Finally, we recognize ordinary words and adjust the results by K-Means Clustering. Our experiments on data of CLP-2012 Chinese person name disambiguation task proves our system performs well. The F score on test data is 86.68%, exceeding the best result of the Bake-off by 6.46%.

Keywords: Named Entity Recognition；Name Entity Disambiguation；Clustering

1 引言

命名实体识别和歧义消解是自然语言理解的一项重要研究内容，对信息抽取、信息检索、问答系统等都具有重要作用。有关命名实体识别已有大量研究[1]，近几年来的国际评测进一步体现出对命名实体消歧的关注。UNED组织了三届WePS (Web People Search) 评测[2-4]，在没有命名实体知识库的情况下，将具有相同指称的命名实体聚集到一起。自2009年起，TAC（Text Analysis Conference）的KBP（Knowledge Base Population）评测[5,6]都包含了实体链接（Entity Linking）的任务，与WePS不同的是，KBP提供了关于实体的知识库，需要将某个实体链接到知识库的相应定义，并将无链接关系的实体进行聚类。

与英文不同，汉语命名实体缺少明确的标记形态，这给命名实体识别和歧义消解带来了新的挑战。首先，普通词可以作为命名实体，比如“高超”一词通常用作形容词，但是也可以作为人名出现；其次，一个词可以作为多种类型的命名实体出现，例如“华明”一词可能是人名、公司名或者地名；另外，重名现象也在汉语中也大量存在并十分严重。

为了探索解决这些问题的方法，第二届CIPS-SIGHAN中文处理国际会议（CLP-2012)举办了汉语命名实体识别与歧义消解评测，评测的参加单位提出了不同的方法，取得了较好

* 本文受国家社科基金重大项目（12&ZD227）、国家 863 计划资助项目（2012AA0111101）和国家自然科学基金资助项目（91024009）资助

的结果。本文基于CLP-2012的评测数据，探究了命名实体识别和歧义消解方法，构建了一种基于多步聚类的命名实体识别和歧义消解框架。在评测的测试数据的F值达到86.68%，高出参评单位最好结果6.46%。

2 相关工作

命名实体识别早期主要使用基于规则的方法[7]。近几年大多采用机器学习方法，包括：隐马尔科夫模型[8]、最大熵模型[9]、条件随机场模型[10]等。

命名实体消歧的方法大致可分为基于文本向量空间模型的聚类方法[11]、基于社会网络的方法[12]、基于分类的方法[13]等。在KBP中出现的方法更加丰富多样[6]，包括无指导相似度计算、有指导分类和排序、基于图的排序、层次聚合式聚类、谱图聚类、主题模型等等。

CLP-2012的命名实体识别与消歧任务共有8支队伍参加，参评单位提出了很多有效的方法。文献[14]使用了分类-聚类的两步模型，并利用文档集合和互联网信息构造了Out类和Other类的知识库定义；文献[15]应用了关键词提取算法来构建特征;文献[16]抽取了人名实体的19种属性，并使用了支持向量机(SVM)训练分类器来为难以通过相似度进行判断的文档分类；文献[17]使用了模糊聚类。

3 系统构架

CLP-2012的命名实体识别与消歧任务融合了WeSP和KBP评测的特点。任务对每个待消歧词提供了知识库来表示实体定义，每个定义由一段文字描述。对每个待消歧词，评测任务提供了一个文本集合T，每个文本都包含相应的待消歧词。对于每个文本t∈T，判断t中出现的歧义词是否对应于知识库中的某个定义，如果是，则输出该定义的编号，否则需要判断该待消歧词是否作为一个普通词出现，如果是，则将其归入Other集合，否则表明该词作为命名实体出现，但是不指向知识库中的任何一个定义，则将其归入Out集合。最后需要对Out集合中的文本进行进一步划分，将指向同一实体的文本归入同一集合，划分结果表示为Out_01,Out_02,……。

CLP-2012 的命名实体识别和歧义消解任务提供了知识库和待消歧文本两组语料，其中知识库的规模较小，因而文献[14,18]不同程度地使用了互联网资源对知识库进行扩充。本文提出的方法表明，充分利用知识库和待消歧文本便可以取得理想的结果。

本文提出的命名实体识别和歧义消解方法流程如下：首先，依据文档和实体定义之间的相似度，进行第一轮聚类；再依据文档与类簇之间的相似度，进行第二轮聚类。通过两轮聚类，将文档与实体定义之间的链接基本完成，剩余的未链接文档主要由 Other 类和 Out 类文档组成。对未链接文档，使用层次聚合式聚类（HAC）算法将 Out 类文档进行聚类，再基于相似度和规则对 Other 类进行标记。最后，使用 K-Means 算法对结果进行微调。本章余下部分将详细介绍本文提出的基于多步聚类的命名实体识别和歧义消解方法。

3.1 预处理

3.1.1 分词和词性标注

评测提供的知识库和待消歧文本两组语料都是未经处理的原始文本，因而需要对其进行分词和词性标注预处理。本文使用了由条件随机场模型（CRF++[1]工具包）设计并实现的分词系统，以SIGHAN2005中文分词评测的北大语料作为训练语料，在测试语料的分词结果F

[1] http://crfpp.googlecode.com/svn/trunk/doc/index.html

值为95.97%。词性标注系统采用了文献[19]的方法，使用最大熵模型（MaxEnt[2]工具包）实现，利用人民日报1998年1月语料进行训练，交叉验证显示词性标注准确率为91.14%。词性标注采用了北大标注标准，其中nr、ns标注分别代表了人名和地名，因此没有再单独对语料进行命名实体识别。

3.1.2 特征选取

上下文信息是实体消歧的重要信息，命名实体消歧方法大多选取上下文信息为特征。文献[16]对命名实体的属性进行了细致的抽取，选择了19种特征。但由于文本的局限性和抽取技术的限制，过于细致的特征抽取难以保证准确和完整。本文选择的特征如下：

作品名：包括书名、影视剧名等，以书名号作为选取界限。

人名：选择词性标注为nr的词。由于北大分词标准将人名中的姓和名划分开，所以利用简单的规则将其还原成完整的人名。

地名：选择词性标注为ns的词。

职业名称：文献[14]将职业名称选作特征，取得了良好的实验效果。本文同样通过互联网构建了表示职业的词表，共计233个名词。另外，由于知识库中一定比例的实体定义为运动员，对于这些定义来说，运动项目名称对消歧会有显著的帮助。因此，本方法在词表中增加了64个表示运动项目的名词。将文本中出现的包含在该词表中的词作为职业名称类的特征。

其他名词特征：选取文档中所有未被选取的名词以及名词性动词。需要说明的是，由于待消歧词在每篇文档中都会出现，对歧义消除没有帮助，反而可能因大量出现而导致相似度的偏差，因此，特征中没有包含待消歧词。

3.1.3 相似度计算

文本间相似度计算采用了基于向量空间模型的余弦相似度，特征权重使用了加权的*tf-idf*值。由于知识库中的实体定义与待消歧文本之间在文本长度上存在显著差异，为了缓解这种不平衡性带来的误差，本文对待消歧文本的*tf*值进行了调整，定义调整函数*f*如下：

$$f(tf_{\text{word}}) = \begin{cases} tf_{\text{word}} & \text{如果该文档属于知识库} \\ \text{ceil}\left(\sqrt{tf_{\text{word}}}\right) & \text{如果该文档是待消歧文档} \end{cases} \tag{1}$$

其中ceil表示向上取整函数。不同类型的特征在歧义消解时的影响是不同的，为此，本文通过实验，为不同特征设定了不同的权重，权重值如表1所示。

特征类型	权重
作品名	2
人名	1.5
地名	1.3
职业名称	2
其他名词特征	1

表1　特征类型权重表

于是，特征词的权重w_{word}按如下方式计算：

$$w_{\text{word}} = f(tf_{\text{word}}) \times idf_{\text{word}} \times \text{weight}(\text{type}_{\text{word}}) \tag{2}$$

向量A与向量B的相似度定义为余弦相似度，即：

$$\text{Sim}(A,B) = \frac{\sum_{i=0}^{n} a_i b_i}{\sqrt{\sum_{i=0}^{n} a_i^2}\sqrt{\sum_{i=0}^{n} b_i^2}} \tag{3}$$

[2] http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html

3.2 基于知识库的聚类

首先，将文档与知识库的定义之间进行链接。对待消歧文档$t \in T$，计算其与知识库中所有定义的相似度，按相似度从大到小进行排序，依据排序结果，可以找到与文档t相似度最高的定义x，如果x与文档t的相似度满足显著条件，则将文档t链接到定义x。本文定义显著条件为，若t与知识库中所有定义的相似度中，最高值与次高值的差值达到某一阈值$threshold_1$，则认为结果显著。之所以设定显著条件，而不是将每个文档分配给最高相似度的那个定义，主要是为了保证聚类结果的准确度。

上述聚类作为第一轮聚类。通过第一轮聚类后，一部分文档被链接到知识库的定义上。假定知识库中的定义数为n，将所有与知识库第i个定义链接的文档都归入集合S_i，于是，第一轮聚类的结果可以表示为n个集合$S_1,S_2,\ldots,S_n$。我们发现，即便有显著条件限制，聚类结果中还是会存在部分错误。为了后续处理达到更好效果，还需要对第一轮聚类的结果进行调整，以尽可能将错误的结果从集合中剔除。调整主要利用了待消歧文本之间的相似度。本文假定，若同一集合中的文本都含有指向同一定义的同名实体，这些文本之间也存在密切的关联。由于文档的长度相较实体定义更长，词汇分布更加均匀，因而相似度的可靠性更高。因此，本文采用的调整策略为，对于集合S_i中的文档t，如果t与集合中其他文档的平均相似度大于t与第i个实体定义的相似度，则S_i保留t，否则从S_i中剔除t，t重新归入未链接文档集合。

3.3 基于文档集合的聚类

经过第一轮聚类和结果调整，得到了聚类结果$S_1,S_2,\ldots,S_n$。在第二轮聚类中，本文使用一轮聚类的结果来对剩余的文档进行链接。聚类过程如下：对于每个未链接文档t，分别计算t与$S_1,S_2,\ldots,S_n$的相似度，定义t与文档集合S的相似度如下：

$$\mathrm{Sim}(t,S) = \frac{1}{|S|}\sum_{k\in S}\mathrm{sim}(t,k) \tag{4}$$

得到n个相似度后，同样采用第一轮聚类中的显著条件，即如果相似度中最高值与次高值的差大于阈值$threshold_2$，则将该文档t加入与之相似度最高的文档集合，即将t链接到该集合对应的实体定义。

3.4 层次聚合式聚类

经过前两轮聚类，大部分与知识库中的定义相关联的文档已经被链接，剩余的未链接文档集合主要由Out类和Other类组成。这些文档与已链接的文档集合$S_1,S_2,\ldots,S_n$的相似度都不符合显著条件，但其中Out类的文档可以进一步形成多个集合，这些集合内的文档间相似度较高。本文使用层次聚合式聚类(Hierarchical Agglomerative Clustering, HAC)算法对剩余文档进行聚类，得到Out类文档的聚类结果。聚类方法如下：

(1) 将每个文档作为一个聚类集合
(2) 计算每两个集合之间的相似度
(3) 将相似度最高的两个集合合并为一个集合
(4) 重复(2)和(3)直到任意两个集合之间的相似度小于某个阈值$threshold_{HAC}$

聚类集合之间的相似度采用组平均相似度，即

$$\mathrm{Sim}(S_i,S_j) = \frac{1}{|S_i|\times|S_j|}\sum_{k_i\in S_i}\sum_{k_j\in S_j}\mathrm{sim}(k_i,k_j) \tag{5}$$

因为未链接文档中还包含了Other类文档，以及少数未被准确链接的实体文档，因此层次聚类的停止阈值不应过低，目的是尽可能使得聚类过程仅涉及相似度较高的Out类文档，而不使非Out类文档参与聚类。

对于聚类的结果，本文选择大小超过2的聚类集合作为Out类集合，因为通过实验我们发

现，大小为2的聚类集合是真实Out类集合的概率不高，选取大小超过2的聚类集合作为Out类集合效果最好。

3.5 判断0ther类文档

Other类的识别是评测任务的一个难点，多个参赛单位使用了命名实体识别系统来对Other类进行识别，但由于评测任务中的待消歧词大多在汉语中通常作为普通词出现，所以命名实体识别对这些词的识别效果不佳。文献[14]指出他们使用的命名实体识别系统对于多个待消歧词的识别准确率仅为0，文献[20]介绍了CLP-2010人名消歧任务取得第一的参评单位所采用的人名识别系统，但该系统对“高明”这类通常作为普通词的人名识别却无能为力，因而文献[20]采用了规则来进行这类人名的识别。

所以，本文提出的方法并没有像大部分参评单位那样在第一步进行Other类文档的识别，而是通过前三步的准确聚类，来保证大部分Other类文档在三步聚类之后仍然未被标注，然后在剩余的未标注文档中，通过相似度和规则相结合的方式来确定Other类文档。具体方法是：如果未标注文档t与实体定义文档集合$S_1, S_2, \ldots, S_n$以及层次聚合式聚类结果$Out_1, Out_2, \ldots, Out_m$的相似度都低于0.02，且待消歧词前后大小为2的窗口中未出现命名实体或职业名称类词语，则将其标记为Other类。

3.6 基于K-Means聚类的结果调整

前几步聚类可以得到链接到知识库定义的n个文档集合$S_1, S_2, \ldots, S_n$以及层次聚合式聚类结果$Out_1, Out_2, \ldots, Out_m$，相应地，可以得到k=m+n个聚类中心，使用类似K-Means聚类的方法，可以对除Other类文档之外的聚类结果进行调整。方法是，将每个非Other类文档t（t可能仍未归入到任何一个集合），归入与之相似度最高的集合，重复该过程直到所有集合保持稳定不变。此时的标记结果就是系统输出的最终结果。

4 实验及结果分析

4.1 实验结果

我们使用CLP-2012评测提供的训练数据作为实验数据，训练数据共有16个待消歧词，1634个待消歧文档。基于训练数据的实验表明，表2所示的阈值取值得到了最佳结果，因此我们依据表2设定阈值。

$threshold_1$	$threshold_2$	$threshold_{HAC}$
0.03	0.04	0.07

表2　阈值选择

为了显示每一步聚类的效果，我们对每一步的结果进行了评测。由于中间结果并没有对所有文档完成标注，所以仅对标注结果的文档进行评估，准确率和召回率均为已标注文档的均值。相应地我们增加了标注率指标，来显示已标注文档占所有文档的比例。中间结果的评测数据如表3所示。

实验阶段	P	R	F	标注率
第一轮聚类	90.96%	84.71%	87.61%	60.32%
聚类调整	96.84%	67.31%	78.78%	40.95%
第二轮聚类	93.96%	87.26%	90.21%	61.76%
层次聚合式聚类	93.75%	86.34%	89.67%	80.82%

判断Other类	91.08%	83.87%	87.10%	87.02%
K-Means调整	86.83%	89.93%	88.35%	100%

表3 分步标注结果评测

从表3中可以看出，第一轮聚类标注了60%的文本，并且准确率已经达到了90%，第一轮聚类总体效果良好。第一轮聚类后的调整有效地提高了准确率，使得调整后的聚类集合保持了较高的纯度，但是召回率以及标注率都有所下降，这说明部分正确标注从结果中被剔除，但调整的主要目的是提高准确率，第二轮聚类仍然有可能保证这部分正确的链接重新被加入结果中。第二轮聚类的结果很好地弥补了第一轮聚类的问题，标注文档比例较第一轮聚类有所上升，准确率和召回率都显著提高。层次聚合式聚类后，标注率提高了20%，准确率和召回率仅稍有下降，说明对Out类的聚类结果比较准确。判断Other类后，准确率和召回率有所下降，说明标记Other类的准确性比知识库和Out类低。经过K-Means聚类调整后，F值最终为88.35%。

K-Means聚类调整的迭代过程如表4所示。从表4可以看出，基于K-Means聚类的调整对结果有小幅度的提升，由于调整前聚类结果较好，所以调整在4轮迭代后就达到稳定。

迭代轮次	P	R	F
1	86.56%	89.43%	87.97%
2	86.59%	89.46%	88.00%
3	86.76%	89.80%	88.25%
4	86.83%	89.93%	88.35%
5	86.83%	89.93%	88.35%

表4 K-Means调整的迭代过程

使用在训练数据上取得最优效果的设定，我们在CLP-2012的测试数据上进行了实验，实验结果如表5所示。

P	R	F
87.77%	85.62%	86.68%

表5 测试数据实验结果

我们将实验结果与参与评测的前三名系统结果进行了比较，如表6所示。可以看出，本文的方法无论在训练集还是测试集上，都优于评测前三名的系统。其中测试集F值与评测第一名相比，提高了6.46%。

系统	训练集F值	测试集F值
本文	**88.35%**	**86.68%**
文献[14]	85.55%	80.22%
文献[15]	76.32%	75.75%
文献[16]	-	75.29%

表6 与其他系统结果比较

4.2 实验结果分析

我们对知识库实体类、Out类、Other类分别进行了评价，结果如表7所示。

类别	P	R	F
知识库实体类	89.93%	95.37%	92.57%

Out类	86.77%	85.97%	86.37%
Other类	64.78%	48.97%	55.78%

表7 结果分类评价

从表7可以看出，对知识库链接以及Out类聚类的结果较好，这说明本文构建的基于向量空间相似度的聚类算法体现出了良好的消歧性能。但基于文本相似度的方法也存在局限性，比如“高峰”一词的文档中，有多篇文章涉及了德云社的演员高峰调侃北京国安足球队引发风波的消息，由于国安、足球等词汇大量出现，错误地将相声演员高峰判断为曾在北京国安队效力的球员高峰。对于这种情况，需要更深层次的语义信息来帮助判断。

表7还显示，Other类的整体F值仅有55.78%。这说明当普通词作为命名实体时，辨识普通词的效果不尽如人意。普通词作为命名实体是汉语的一种常见现象，现有的基于规则和机器学习的方法尚不能很好地解决这类识别问题，还需要从语义理解的角度获取更多可靠信息来提高该任务中普通名词的识别效果。

5. 结语

本文基于向量空间相似度，使用多步聚类的方法，实现了命名实体识别与歧义消解的模型。在 CLP-2012 评测语料上的实验结果表明，本文所采用的多步聚类方法是有效的，将评测的结果提高了 6.46%。同时，本文的方法不需要借助其他语料或者人工构造、修改语料，具有良好的适用性。但仍有不足之处，对于普通词的识别效果较差。

下一步，我们将进一步利用和融合更多信息，包括互联网搜索结果及百科信息等，并从更深层次的语义层面入手，挖掘文本中蕴含的语义信息，来进一步提高命名实体识别和歧义消解的效果。

参 考 文 献

[1] 赵军. 命名实体识别、排歧和跨语言关联[J]. 中文信息学报, 第 23 卷第 02 期, 2009 : 3 - 17.

[2] J. Artiles, J. Gonzalo, and S. Sekine. The SemEval-2007 WePS evaluation: Establishing a Benchmark for the Web People Search Task. SemEval '07 Proceedings of the 4th International Workshop on Semantic Evaluations, 2007: 64-69.

[3] J. Artiles, J. Gonzalo, and S. Sekine. WePS 2 Evaluation Campaign: Overview of the Web People Search Clustering Task. In 2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference, 2009.

[4] J. Artiles, A. Borthwick, J. Gonzalo, S. Sekine, and E. Amigó. WePS-3 Evaluation Campaign: Overview of the Web People Search Clustering and Attribute Extraction Tasks. In Conference on Multilingual and Multimodal Information Access Evaluation (CLEF). 2010.

[5] H. Ji, R. Grishman, H. T. Dang, K. Griffitt and J. Ellis. An Overview of the TAC2010 Knowledge Base Population Track. In Proceedings of Text Analytics Conference (TAC2010).

[6] H. Ji, R. Grishman and H. T. Dang. An Overview of the TAC2011 Knowledge Base Population Track. In Proceedings of Text Analysis Conference (TAC2011).

[7] R Grishman , B Sundheim. Design of the MUC-6 evaluation. In Proceedings of 6th Message Understanding Conference , 1995.

[8] J. Sun , J. Gao , L. Zhang , et al. Chinese Named Entity Identification Using Class-based Language Model . Proceedings of the 19th International Conference on Computational Linguistics(COLING 2002):1-7.

[9] A. Borthwick. A Maximum Entropy Approach to Named Entity Recognition . New York : New York University. 1999.

[10] X. Mao, Y. Dong, S. He, et al. Chinese Word Segmentation and Named Entity Recognition Based on Conditional Random Fields. In Sixth SIGHAN Workshop on Chinese Language Processing. 2008: 90-93.

[11] Silviu Cucerzan. Large-Scale Named Entity Disambiguation Based on Wikipedia Data. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP2007): 708-716.

[12] Ron Bekkerman , Andrew McCallum. Disambiguating Web Appearances of People in a Social Network. Proceedings of the 14th International Conference on World Wide Web (WWW2005) : 463-470.

[13] X. Han , J. Zhao. Person Name Disambiguation Based on Web-Based Person Mining and Categorization. Submitted to Second Web People Search Evaluation Workshop in conjunction with WWW2009.

[14] Z. Peng, L. Sun, and X. Han. SIR-NERD: A Chinese Named Entity Recognition and Disambiguation System Using a Two-stage Method. In The 2nd CIPS-SIGHAN Joint Conference on Chinese Language Processing (CLP-2012).

[15] H. Zong, D. F. Wong, and L. S. Chao. A Template Based Hybrid Model for Chinese Personal Name Disambiguation. In The 2nd CIPS-SIGHAN Joint Conference on Chinese Language Processing(CLP-2012).

[16] W. Han, G. Liu, Y. Mao, and Z. Huang. Attribute Based Chinese Named Entity Recognition and Disambiguation. In The 2nd CIPS-SIGHAN Joint Conference on Chinese Language Processing (CLP-2012).

[17] W. Tian, X. Pan, Z. Yu, Y. Xian, X. Yang, Y. Qin, and W. Long. Chinese Name Disambiguation Based on Adaptive Clustering with the Attribute Features. In The 2nd CIPS-SIGHAN Joint Conference on Chinese Language Processing (CLP-2012).

[18] J. Liu, R. Xu, Q. Lu, and J.Xu. Explore Chinese Encyclopedic Knowledge to Disambiguate Person Names. In The 2nd CIPS-SIGHAN Joint Conference on Chinese Language Processing (CLP-2012).

[19] H. T. Ng and J. K. Low. 2004. Chinese Part-of-speech Tagging: One-at-a-time or All-at-once? Word-based or Character-based? In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004):277–284.

[20] 时迎超, 王会珍, 肖桐, 胡明涵. 面向人名消歧任务的人名识别系统[J]. 中文信息学报, 第 25 卷第 03 期, 2009: 17 – 22.