基于潜在语义分析（LSA）的新蒙古文命名实体识别的歧义消解

蒋玉鹏[1]，侯宏旭[1]，杨萍[1,2],杜健[1],申志鹏[1],李金廷[1]
（1.内蒙古大学计 算机学院,内蒙古 呼和浩特 010021；2.临汾职业技术学院 计算机系,山西 临汾 041000）

摘要：命名实体是承载文本重要信息的语言单位。命名实体识别、消歧是自然语言处理的重要研究内容。针对新蒙古文中的命名实体与普通名词不易区分（如：巴特尔既可以指普通的英雄，也可以指具体的一个人），且知识库匮乏、覆盖面小的问题，本文提出了基于潜在语义分析（LSA）的新蒙古文命名实体消歧的方法。首先对新蒙古文词进行词缀切分，只对词根进行向量空间的构建，通过奇异矩阵分解得到实体之间的潜在语义关系；通过上下文的知识来弥补知识库匮乏的问题，进而得到语义相关的实体。通过结合词性相关度的信息和语义相关词对命名实体类别的贡献度进行加权来得到命名实体的真实类别指向。使用该方法进行命名实体的消歧后，命名实体识别的平均 F 值比未消歧之前高出了 3.11%。

关键词：命名实体识别；命名实体消歧；新蒙古文；潜在语义分析（LSA）

Research on Chinese-Slavic Mongolian Named Entity Recognition Disambiguation Based on Latent Semantic Analysis

Jiang Yupeng[1], Hou Hongxu[1], Yang Ping[1] ,Du Jian[1],Shen Zhipeng[1],Li Jinting
(1. Inner Mongolia University College of Computer, Hohhot, Inner Mongolia, 010021,China ;
2. Linfen Vocational and Technical College, Lin Fen, Shan Xi Province, 041000,China)

Abstract: Named Entities are important meaningful units in texts. The recognition and disambiguation of Named Entities is of great significance in the field of natural language processing. It is difficult to distinguish between Named Entities and common nouns in Slavic Mongolian. For example, “巴特尔” which can refer to the ordinary heroes, it can also refer to a specific person. Meanwhile, The Slavic Mongolian is lack of the repository. To solve the above problem, we propose a new method which based on the Latent Semantic Analysis (LSA) to do the Named Entity disambiguation. Firstly, we do the stemming on the Slavic Mongolian words to reduce the matrix dimension. Then, we construct the stem matrix to get semantic related words by clustering. We use the context knowledge to make up the deficiency of knowledge base, then we can get the semantic related Named Entity. Finally, combining the parts of speech related information and the popularity of words to disambiguate. Experiment results show that We got a 3.11% F-Score higher than before-disambiguation result.

Keywords: Named Entity Recognition; Named Entity Disambiguation; Slavic Mongolian; Latent Semantic

1 引言

命名实体的识别、消歧是自然语言处理的重要研究内容。命名实体的识别从早期的基于规则的方法[1]逐渐发展到近期机器学习的方法，如隐马尔可夫模型[2]、最大熵模型[3]、条件随机场模型[4]等。

命名实体的歧义是指一个命名实体的指称项可以与多个命名实体概念相对应[5]。命名实体的消歧就是利用文本上下文信息或其它外部知识库来确定此指称项的具体指向。在过去的几年中，一些处理命名实体消歧的方法被先后提出。例如，基于文本向量空间模型的方法[6]，基于社会网络的方法[7]、基于分类的方法[8]以及在 KBP 中出现的无指导相似度计算、基于图的排序、谱图聚类等等。

从目前已有的文献来看，关于新蒙古文（以下简称蒙古文）命名实体的研究仍相对较少，在传统蒙古文的命名实体识别方面，那顺乌日图等人采用直接标注、词典匹配及

基于上下文的算法等方式进行蒙古文人名的自动识别 [0]。通拉嘎将最大熵的数学模型应用于蒙古文命名实体的识别当中,实现了蒙古文人名自动识别系统 [0]。这些研究都还只是针对传统蒙古文的人名识别，未涉及到传统蒙古文地名以及机构名的识别以及命名实体的消歧工作。而新蒙古文的命名实体识别、消歧还未有过相关方面的论述。传统的利用待消歧实体上下文的向量空间聚类方法只考虑了命名实体上下文之间的词语共现情况，而忽略了文本间词与词之间的关联度以及相互的语义关系。针对上述问题，本文采用LSA对命名实体进行潜在的语义分析以及对命名实体进行空间向量模型（VSM）的构造，采用基于向量距离的词序相似度算法对实体及其相关的词语进行聚类分析，找出语义上最相关的词确定其命名实体的真实属性，从而完成命名实体的消歧，提高命名实体识别的结果。实验结果表明，使用本文的方法进行命名实体消歧后，机构名识别的 F 值均高出约 4%，人名与地名识别的 F 值比未消歧时高出约 1.5%。

2 LSA 潜在语义分析

LSA（Latent Semantic Analysis）潜在语义分析，即对文本中的浅层语义进行分析提取。LSA 使用统计计算的方法对大量的文本集进行分析，将词和文档映射到潜在语义空间，从而提取和表示出词的语义。在很大程度上决定了词语之间语义上的相关性。

在对新蒙古文构造向量空间时，由于新蒙古文存在构词词缀的问题，即同一种事物在句子中充当不同的结构时会存在不同的词缀表达；直接对新蒙文进行向量空间构造时会存在大量的矩阵稀疏的问题。例如“国务院”一词，在新蒙语中用作主语的时候为“Төрийн Зөвлөл”，而当其用作谓语或定语的时候，要再缀接一个词缀“ийн”“Төрийн Зөвлөлийн”。所以本文在对新蒙古文文本构造向量空间时，首先对其进行词缀切分。我们使用包含 567 个不重复词缀的词缀词典，包含 7843 个不重复词干的词干词典以及 100 个词的高频词典以及包含 48 个词的外来词词典用于词缀切分。每类词典在系统初始化时都建立索引，以提高查询速度。在切分词缀之后对词干直接进行向量空间的构造。然后通过奇异矩阵分解进行降维得到潜在的相关语义。

降维采用 LSA/SVD 中的 SVD 奇异矩阵分解。矩阵分解的原理是采用原始矩阵 C 分解为 U*Σ*V 三个矩阵的乘积，通过对Σ矩阵进行降阶为 k 维得到Σk，然后重新进行合并 U*Σk *V 得到新的矩阵 Ck，即可得到词与词，文档与文档之间的潜在语义关系。具体的分解过程如图 2.1 所示，

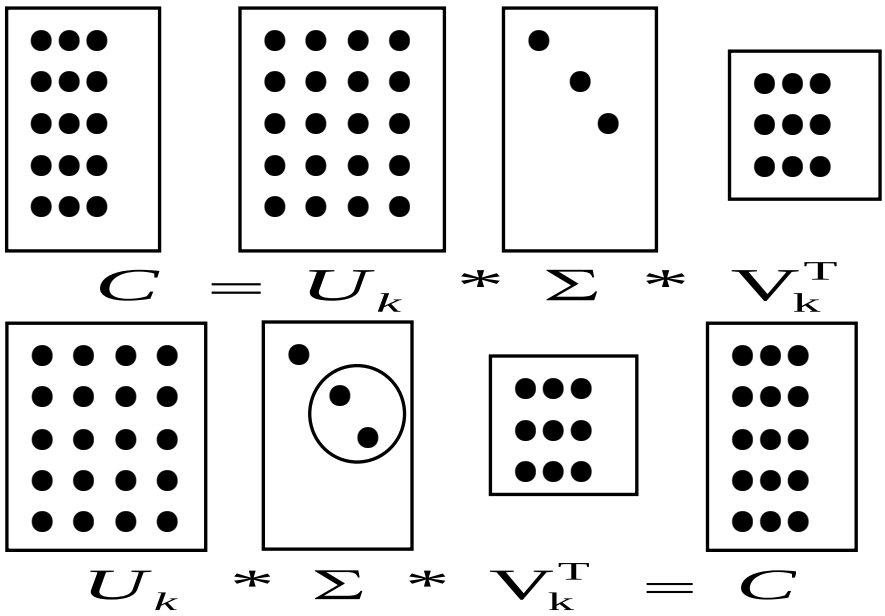

图 2.1 LSA 中的 SVD 语义压缩示意图

3 基于向量距离和词性信息的模式匹　配

我们采用向量空间和词性信息匹配来寻找与待消歧命名实体相关的实体词。通过潜在语义分析（LSA）构造词和文档向量空间模型。再对向量矩阵进行相似度计算找出语义相关词。通过更加细致的切分，挖掘实体内部词汇的语义信息，进行细粒度的词性信息匹配，找出语义和形式相近的命名实体。

在向量距离的相似度计算中，我们用到了表 3.1 词性相似度权值表。通过表 3.1 中各词性间相似度的权值以及由我们在 5 万句语料库上根据互信息统计出的各个词在命名实体中的贡献度，进行线性插值而得到最相近的前 N 个词语。

表 3.1 词性相似度权值表

	名词	代词	数词	形容词	副词	其它词
名词	1	0.8	0.8	0.5	0.5	0.3
代词	0.8	1	0.8	0.5	0.5	0.3
数词	0.8	0.8	1	0.5	0.5	0.3
形容词	0.5	0.5	0.5	1	0.5	0.3
副词	0.5	0.5	0.5	0.5	1	0.3
其它词	0.3	0.3	0.3	0.3	0.3	1

区别于前面的研究者进行消歧仅仅寻找语义相关的词语。我们在实验过程不仅寻找语义最相关词，而且也通过模式匹配找到模式相近的词。综合语义和模式相关词，进行歧义消解。

所谓的模式相关词，是指将组合性的命名实体进行细粒度的切分，通过词性信息逐一匹配词性而得到的模式相近的组合词。而命名实体一般是由一个或多个词的组成的组合性名词。在进行命名实体的消歧过程中，将命名实体进行细粒度的再切分。将细粒度的词性信息和相关度信息加入到消歧任务中来会得到一个很好的效果。所以我们在命名消歧的过程中将命名实体进行细粒度的拆分。例如，“Күнзийн Институт”，我们将它进行划分，Күнзийн\n Институт\n（或者例如，“北京\n 自然语言\n 处理\v 实验室\n”）。这样将命名实体的内部信息进行挖掘，来得到命名实体的真实类别。

所以，我们设计了以下两个部分进行命名实体的消歧。

首先，寻找跟待消歧实体相近的命名实体。这些命名实体分类两类，第一类是语义相关词，通过潜在语义分析进行词义聚类，得到语义相关词；第二类是与待消歧实体具有相同词性结构的同形实体，这些命名实体通过下面的公式（1）进行形式匹配得到。公式（1）如下所示，

$$match_d = \sum_{i=1}^{n}\sum_{j=1}^{m} 2 * w_i / (len_i + len_j) \quad (1)$$

其中，$match_d$ 表示实体跟待消歧实体的相似度，w_i 表示词性相似度，len_i 表示待消歧的命名实体切分后的长度，len_j 表示寻找的词的长度。

结合上面得到的两类词汇，得到一个与待消歧命名实体相关的 *N-Best* 列表，其中 N 取值为 10，语义相关词 8 个，模式相关词 2 个。在 *N-Best* 的基础上结合词性相似度权值表，融合词在命名实体中的贡献度信息进行线性插值，通过公式（2）的得到待消歧命名实体的真实属性类别。公式（2）如下所示，

$$P_{LOC}(W)=\sum_{i}^{n}\left\{\sum_{j}^{m}P(LOC\mid N_j)*R(W_i,N_j)\right\}$$
$$P_{ORG}(W)=\sum_{i}^{n}\left\{\sum_{j}^{m}P(ORG\mid N_j)*R(W_i,N_j)\right\} \quad (2)$$
$$P_{PER}(W)=\sum_{i}^{n}\left\{\sum_{j}^{m}P(PER\mid N_j)*R(W_i,N_j)\right\}$$

其中 W 表示待消歧的命名实体，N_j 是通过语义分析得到的 m 个与命名实体词 W_i 相关的词，$R(W_i,N_j)$是命名实体 W_i 和相关词 N_j 通过表 3.1 得到的词性相似度，$P(LOC|N_j)$ 是相关词 N_j 对命名实体识别为地名的贡献度，$P(ORG|N_j)$是相关词 N_i 对命名实体识别为机构名的贡献度，$P(PER|N_j)$是相关词 Ni 对命名实体识别为人名的贡献度，$P_{LOC}(W)$，$P_{ORG}(W)$，$P_{PER}(W)$是在结合词性信息表的基础上重新得到命名实体的真实类别的概率。

将系统识别得到的的命名实体依次经过上述两个步骤进行歧义消解，得到待消歧命名实体真实类别属性。

4 实验以及结果分析

4.1 实验设置

我们在新蒙古文命名实体识别工作的基础上，对 5 万句新蒙古文语料进行命名实体识别的消歧处理。新蒙古文的命名实体识别效果如下表所示。

表 4.1 新蒙古文命名实体识别的效果

	准确率	召回率	F 值
新蒙古文地名	90.86%	84.66%	87.65
新蒙古文机构名	81.39%	63.72%	71.48
新蒙古文人名	68.30%	66.04%	67.15

本文采用潜在语义分析/奇异矩阵分解（*LSA/SVD*）对待消歧实体构造的向量空间的构造进行降维寻找实体之间的潜在语义。结合相关的聚类算法（*K-Means*）找出语义相关的命名实体。对相关的命名实体进行细粒度的切分，通过结合词性相似度和相关实体对待消歧命名实体识别的贡献度来确定待消歧的命名实体的真实类别。

4.2 实验结果及分析

我们在5万句标注的新蒙古语语料的基础上进行命名实体的识别和消歧工作。在命名实体识别的任务中，4 万 8 千句作为训练语料进行模型训练，2000 句进行测试。并在这 2000 句的识别的基础上进行命名实体的消歧。实验结果如表 4.2 所示，

表 4.2 实验结果

功能模块	识别类型	准确率	召回率	F1 值
CRF	LOC（地名）	93.04%	88.75%	90.84%

	ORG（机构名）	83.68%	76.34%	79.84%
	PER（人名）	89.35%	91.46%	90.39%
CRF++N-best Clustering	LOC（地名）	93.04%	88.75%	90.84%
	ORG（机构名）	83.68%	76.34%	79.84%
	PER（人名）	90.33%	91.03%	90.68%
CRF+N-best Clustering+LSA	LOC（地名）	***94.01%***	***90.71%***	***92.33%***
	ORG（机构名）	***87.71%***	***80.34%***	***83.86%***
	PER（人名）	***91.44%***	***91.50%***	***91.47%***

通过实验结果可以发现，通过语义分析结合聚类的消歧使命名实体的识别效果取得了一个显著的提升。对于兼类现象明显的机构名识别方面，由于系统加入了实体内部的语义特征信息以及上下文的语义信息所以取得了一个较高的提升。由于人名一般不可再分，不能统计内部信息，所以在在人名的识别方面提升效果不是特别明显。

5 结论与展望

命名实体识别、消歧是信息抽取、信息检索、机器翻译、组块分析、问答系统等多种自然语言处理技术的重要基础。本文在新蒙古语命名实体识别的基础上，对相应的命名实体进行消歧。与基线系统相比，在人名、地名和机构名方面有相应的提升，对机构名的识别效果提升最大。

之后，我们将从以下几个方面对系统进行改进和完善:首先，我们将进一步扩大训练语料的规模，并扩展语料涉及的主题和语料类型。同时可以考虑人工建立歧义消解语料库，对于歧义消解算法的性能进行量化评估和比较。其次，系统目前仍然对人工参与具有一定的依赖性，要实现更高的自动化要求，需要在无监督及半监督学习方法做更多的深入研究。

参考文献

[1]Grishman R, Sundheim B. Design of the MUC-6 Evaluation[J]. In Proceedings of the 6th Conference on Message Understanding (MUC ’95, 1995:1-11.

[2]俞鸿魁, 张华平, 刘群等. 基于层叠隐马尔可夫模型的中文命名实体识别[J]. 通信学报, 2006, 27(2):87-94.

DOI:10.3321/j.issn:1000-436X.2006.02.013.

[3]Borthwick A, Sterling J, Agichtein E, et al. Exploiting Diverse Knowledge Sources via Maximum Entropy in Named Entity Recognition[J]. IN PROCEEDINGS OF THE SIXTH WORKSHOP ON VERY LARGE CORPORA, 1998.

[4]McCallum A, Li W. Early Results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-Enhanced Lexicons[J]. Computer Science Department Faculty Publication Series, 2003.

[5] 龚凌晖. 中文命名实体识别与歧义消解研究

[D]. 复旦大学, 2011.

[6]large-scale named entity disambiguation based on Wikipedia data

[7]Disambiguating web appearances of people in a social network

[8] person name disambiguation based on web-based person mining and categorization

[9]那顺乌日图, 雪艳, 淑琴,等. 蒙古文人名自动识别研究[C]// 全国第七届计算语言学联合学术会议. 2003.

[10]通拉嘎. 基于蒙古文语料库的人名自动识别[D]. 中央民族大学, 2013.