中文语义组块自动抽取方法

作者	钟茂生，荆佳琦，陈晓霞，余美华，陶兰
机构	华东交通大学 信息工程学院；江西水利职业学院
发表期刊	《计算机应用研究》
预排期卷	2018 年第 35 卷第 2 期
访问地址	http://www.arocmag.com/article/02-2018-02-058.html
发布日期	2017-03-15 09:22:19
引用格式	钟茂生，荆佳琦，陈晓霞，余美华，陶兰. 中文语义组块自动抽取方法[J/OL]. [2017-03-15]. http://www.arocmag.com/article/02-2018-02-058.html.
摘要	句子语义表述是当前自然语言处理领域亟待解决的重要问题，是制约自然语言能否取得深度应用的重要因素。根据中文文本的特点，摈弃以前自然语言处理语义与句法相分离的观点，提出语义组块概念，并利用深度信念网络的深度学习方法构建对中文语义组块进行自动抽取的模型，模型以句子中名词为核心，将名词与其前后词语进行组合后构成中文语义组块，之后分别使用神经网络、支持向量机和深度信念网络三种抽取方法构建抽取模型，进行了三组实验，最终结果显示在高维大数据背景下，深度信念网络的方法与支持向量机和神经网络相比较具有更好的抽取效…
关键词	语义表述，深度信念网络，深度学习，中文语义组块
中图分类号	TP391.47
基金项目	国家自然科学基金资助项目（61462027，61363072）

中文语义组块自动抽取方法 *

钟茂生[1]，荆佳琦[1]，陈晓霞[1]，余美华[2]，陶　兰[1]

(1．华东交通大学 信息工程学院，南昌 330013；2．江西水利职业学院，南昌 330013)

摘　要：句子语义表述是当前自然语言处理领域亟待解决的重要问题，是制约自然语言能否取得深度应用的重要因素。根据中文文本的特点，摈弃以前自然语言处理语义与句法相分离的观点，提出语义组块概念，并利用深度信念网络的深度学习方法构建对中文语义组块进行自动抽取的模型，模型以句子中名词为核心，将名词与其前后词语进行组合后构成中文语义组块，之后分别使用神经网络、支持向量机和深度信念网络三种抽取方法构建抽取模型，进行了三组实验，最终结果显示在高维大数据背景下，深度信念网络的方法与支持向量机和神经网络相比较具有更好的抽取效果。

关键词：语义表述；深度信念网络；深度学习；中文语义组块

中图分类号：TP391.47

Research on automatic extraction of chinese semantic clustering unit

Zhong Maosheng[1], Jing Jiaqi[1], Chen Xiaoxia[1], Yu Meihua[2], Tao Lan[1]

(1. School of Information Engineering, East China Jiaotong University, Nanchang 330013, China; 2. Jiangxi Water Resources Institute, Nanchang 330013, China)

Abstract: Sentence semantic representation is not only a key problem in natural language processing to be solved at present, but also an important restriction factor whether nature language processing is ability to make deep application. In this article, we based on the characteristics of the Chinese text, abandoned the point of separating semantic and grammar, and then put forward a new style of Semantic Clustering Unit, and do a research on information extraction with a deep learning model based on deep belief net, the model takes the noun as the core in the sentence and combines the noun with its before and after words to from Semantic Clustering Unit.Then, three extraction methods, neural network, support vector machine and depth belief network, are used to construct the extraction model . Experimentally, there are three groups of experimental in this paper, finally results show that under the conditions of large data, Deep belief network methods compared with support vector machines and neural networks, which has better effect.

Key Words: Semantic representation; Deep belief net; Deep learning; Semantic Clustering Unit

0　引言

近几年来，中文自然语言处理的技术在分词与词性标注方面取得了很大的进步，但自然语言处理的最终问题是实现机器对自然语言的理解，即寻找一种文本表示方式或者模型，实现机器对自然语言语义的理解[1]。

在文本表示方面，目前主要采用的方法是依存句法[2]的方法，以及新兴的一种基于概念图[3]的表示方法进行表示，这些方法对于文本的分析与表示都是以词为基本单位的来进行的，构建方法也主要采用手工构建以及一些浅层学习方法进行构建，如 WorldNet[4]和 Cyc[5]。这些语义表述的方法存在以下两点问题：a)将词语作为基本的处理单位，与人们正常理解语句的方法不同，并且以词语为单位对文本进行处理，会很大程度的提高要处理单位的数量，提高问题处理复杂性;b)依存句法与概念图的构建需要构建者具有比较专业的语言学基础，这使得通过这种方式来对大规模文本进行处理变得特别困难[6]。此外，在自然语言处理的信息抽取和模式识别方面，占统治地位的依然是基于规则和统计的方法，如隐马尔可夫模型、最大熵模型、决策树模型等，近几年兴起的支持向量机[7]的方法以及神经网络[8]的方法也取得了一定成功。但这些方法也都存在着一系列问题，尤其是在现如今信息爆炸的大数据时代，在面对大规模复杂多变的自然语言语料（主要是文本），这些处理方法的问题也就更加凸显，比如对文本特征及标注要求较高，模型适应性比较差等[9]。

本文提出了一种介于词与句子聚合体之间的称为中文语义组块的概念，将词语按照人们正常的理解方式进行结合，并利

基金项目：国家自然科学基金资助项目（61462027，61363072）

作者简介：钟茂生（1974-），男，主要研究方向为自然语言处理、数据挖掘；荆佳琦（1992-），男，硕士研究生，主要研究方向为自然语言处理（jjq153287083@163.com）；陈晓霞（1987-），女，硕士研究生，主要研究方向为数据挖掘.

用了深度信念网络对中文语义组合单元进行自动抽取。文章使用分别采用支持向量机，神经网络，深度信念网络三种方法进行实验，对目标对象进行识别，并对实验结果进行了对比分析。

1　中文语义组块

中文语义组块是一种在依存句法树的基础上，结合概念图的思想，按照人对自然语言理解方式，通过将词语与其前后词语进行结合，将单个词语组合成比较大的一个单元，即中文语义组块。

以图 1 的一个依存句法树为例。

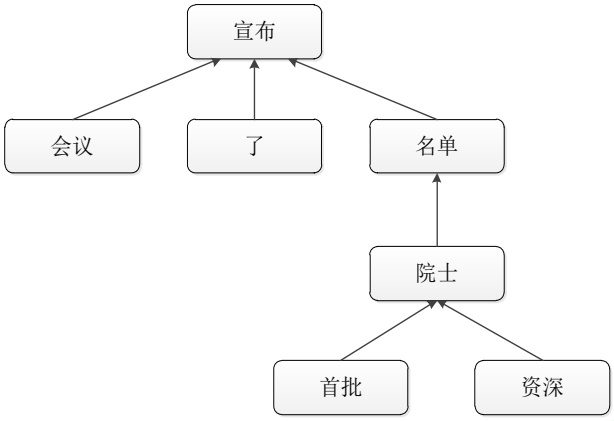

图 1 依存句法树

该依存句法树将句子“会议宣布了首批资深专家名单”分为 7 个词语表示，但对于人们理解而言，如果能将“首批资深专家名单”这四个词进行归并，作为一个整体理解处理，则更符合人们对语言尤其是汉语中文的理解习惯。这种结构介于词语与句子之间，与 E-A-V 概念图[10]中的实体定义比较相似，又较实体定义的范围更加广泛，本文称之为中文语义组块。通过将语句中的词语归并为中文语义组块，可以在不对句子语义造成影响的前提下有效的呈现句子结构，简化语义表述的复杂性，最终为实现语义表述创造条件。

2　基于深度信念网络的中文语义组块抽取方法

2.1　深度信念网络

深度信念网络是一种深度神经网络，由多层 RBM 和一层 BP 组成[11]，这些神经元分为显性神经元和隐性神经元，其中显性神经元用于接受数据的输入，隐性神经元则用于特征提取。典型结构如图 2 所示。

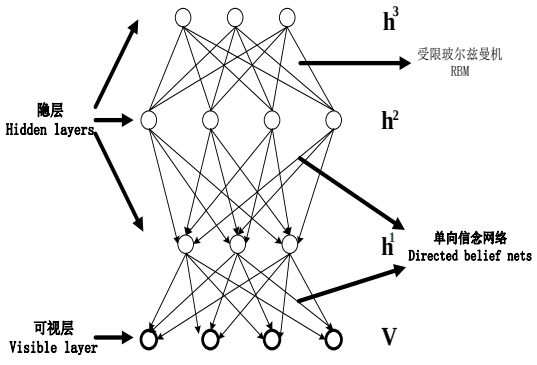

图 2　深度信念网络网络结构图

深度信念网络的一个重要组成元件是受限玻尔兹曼机，训练深度信念网络的过程是先一层一层的训练受限玻尔兹曼机，每一层使用上一层的隐层输出作为新一层的输入，再把新一层的隐层作为更高层的输入向量，最后使用反向传播网络对整个网络权值进行微调。

由式（1）可以得到 RBM 网络的能量函数：

$$E(v,h) = -\sum_{i \in visible} a_i v_i - \sum_{j \in hidden} b_j h_j - \sum_{i,j} v_i h_j w_{ij} \quad (1)$$

其中：v_i 和 h_j 表示隐含层和可视层的值，a_i 和 b_j 代表隐含层和可视层的偏置值，最后的 W 是指隐含层与可视层相互映射的权重。由式（2）表示了从可视层得到隐含层的关系。

$$P(h_{j=1} \mid v) = sigm(W_{ij} v + b_j) \quad (2)$$

由于 RBM 网络的对称性，所以从隐含层得到可视层的函数表示如式（3）所示。

$$P(v_{i=1} \mid h) = sigm(W_{ji} h + a_i) \quad (3)$$

因此得到 v 和 h 的联合概率如式（4）所示。

$$p(v,h) = (1/Z)e^{-E(v,h)} \quad (4)$$

其中 Z 的定义如式（5）所示。

$$Z = \sum_{v,h} e^{-E(v,h)} \quad (5)$$

最终得到概率分布函数为式（6）。

$$P(v) = \sum_{h} e^{-E(v,h)} / Z \quad (6)$$

网络训练通过确定 $\theta = (W,a,b)$，使得联合概率分布 $P(v,h)$ 取得最大值。由于网络的复杂性，最大似然法并不能求得满足条件的参数 θ，常用的做法是使用马尔可夫链蒙特卡罗的特性，可视层与隐含层相互不断更新，最终趋于平稳，此时的参数 θ 便是参数目标值。

使用深度信念网络，可以有效的减少神经网络陷入局部最优解的现象，也能极大地缩短神经网络的训练时间，使其尽快接近最优解。

2.2　中文语义组块自动抽取模型

中文语义组块模型主要完成两个任务，第一完成语料处理，第二从文本语料中自动抽取中文语义组块。整体结构如图 3 所示。

模型在前期数据处理之后，在中文语义组块抽取阶段，将语义组块抽取工作转变为词语不同组合形式的分类问题，分别支持向量机，神经网络和深度信念网络进行语义组块的识别抽取，对比三种方法的抽取结果，最终抽取阶段采用深度信念网络的方法构建抽取模型。

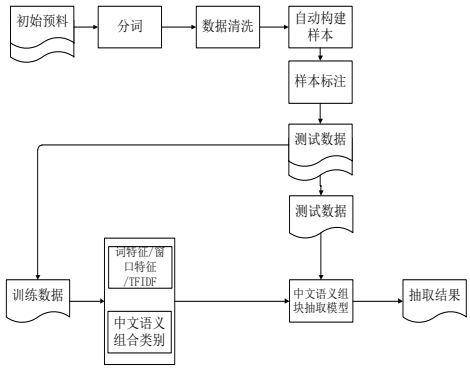

图 4 中文语义组块自动抽取模型

3 实验设置

3.1 标注语料及标注符号

本文选用的数据为江西省 2010 年科技评审项目的项目简介，共 101 篇文章，采用中科院 ICTCLAS2015 分词系统，分词去掉停用词结果共有 58 576 个，去掉重复项共 4213 个，其中名词 18126 个。实验以名词为核心，设立大小为 5 的窗口，标点符合默认为“Null”，不同的结合方式标注不同的值，假设词语的前两个词分别为 T1 和 T2，后两个词为 L1 和 L2，标注方式如表 1 所示。

表 1 标注方法

标注结果	标注值
N	9
（T2，N）	1
（N，L1）	2
（T2，N,L1）	3
（T1，T2，N）	4
（N,L1，L2）	5
（T1，T2，N,L1）	6
（T2，N,L1，L2）	7
（T1，T2，N,L1，L2）	8

例如对于“本项目采用双涂层工艺，改传统的三层涂漆为二层涂漆。”。 分词结果为“/r 项目/n 采用/v 双/m 涂层/n 工艺/n，/w 改/v 传统/n 的/u 三/m 层/q 涂/v 漆/n 为/p 二/m 层/q 涂/v 漆/n。/w”。以名词为核心，最终构建窗口如表 2 所示。

表 2 上下文窗口及标注示例

词(N)	上文窗口 1（T1）	上文窗口 2（T2）	下文窗口 1 (L1)	下文窗口 2 (L2)	标注结果	标注值
项目	Null	本	采用	双	（T2,N）	1
涂层	采用	双	工艺	Null	（T2，N，L1）	4
工艺	双	涂层	Null	Null	（T1，T2，N）	3
传统	Null	改	的	三	（N）	9
漆	层	涂	为	二	（T2，N）	1

实验数据采用上述预处理完成后，对 18126 核心名词进行组合得到 18126 个数据样本，随机选取其中 12000 个样本作为训练样本。剩余的 6126 个数据作为测试样本。

3.2 特征向量

本次实验选取特征包含以下四个部分：

a)词语的 IFIDF 值。

b)五个词组成的窗口词向量。

c)五个词组成的词性向量。

d)中心词在文章中所处的位置。

词典中共有 4213 个词语，因此在将词向量中置“1”处由该出的词的 TFIDF 值替换后，第一个特征与第二个特征组成的向量维数为 4 213 维，词性共有 23 种词性，故第三个特征的特征向量应有 23*5 即 115 维的词性向量，统计得到 101 篇文本中，分词最多为 1 087 个词，因此第四个特征特征向量为 1 087 维，最终的到特征向量 5 415 维。

实验数据维度在 5 415 维，有效句子总数超过 30 000 条，这种数据就以前的文本抽取模型来说，都是比较巨大的。

3.3 标签向量

标签值为 1~9，设置 9 维向量，将与标签值相应位置为“1”。设置结果如表 3 所示。

表 3 标签向量

标签值	标签向量
1	000000001
2	000000010
3	000000100
4	000001000
5	000010000
6	000100000
7	001000000
8	010000000
9	100000000

3.4 实验评价指标

实验的每类结果使用准确度（precision），召回率(recall)和 F 值进行评价。对于深度信念网络和神经网络两种模型，加入训练时间及收敛速度和结果的评价。

指标定义如下：

$$Precision^i = \frac{|\{relevent\ \ number^i\} \cap \{retrieved\ number^i\}|}{|\{retrieved\ number^i\}|}$$

$$Recall^i = \frac{|\{relevent\ \ number^i\} \cap \{retrieved\ number^i\}|}{|\{relevent\ \ number^i\}|}$$

$$F_1^i = 2 \cdot \frac{precision^i \cdot recall^i}{precision^i + recall^i}$$

其中 i 的取值范围为 1-9，分别代表第一类到第九类各自的相关指标。

4　实验结果与分析

4.1　实验效果与分析

表 4 是对整个实验语料词语前后组合的九类情况数目汇总情况。

将数据的特征向量作为输入参数分别输入三种模型进行训练，并将模型输出结果与标注结果进行对比，计算各个方法不同类别的准确度，召回率和 F 值，所得结果如表 5 所示。

表 4 实验语料各类别数量

测试类别	训练数据数目	测试数据数目
第 1 类	4646	2638
第 2 类	2150	1079
第 3 类	1180	597
第 4 类	450	203
第 5 类	274	137
第 6 类	85	44
第 7 类	89	53
第 8 类	27	19
第 9 类	2595	1356

表 5 实验模型结果

模型	深度信念网络			神经网络			支持向量机		
指标	准确率	召回率	F 值	准确率	召回率	F 值	准确率	召回率	F 值
第 1 类	73.45%	86.62%	79.49%	43.61%	54.09%	48.29%	70.92%	83.78%	76.82%
第 2 类	74.94%	61.54%	67.58%	16.80%	30.03%	21.55%	60.41%	59.68%	60.05%
第 3 类	32.89%	49.41%	39.49%	0.00%	0.00%	0.00%	45.02%	48.41%	46.65%
第 4 类	7.94%	2.46%	3.76%	5.00%	0.49%	0.90%	34.62%	13.30%	19.22%
第 5 类	4.03%	3.65%	3.83%	9.09%	0.73%	1.35%	36.92%	35.04%	35.96%
第 6 类	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	66.67%	22.73%	33.90%
第 7 类	12.50%	18.87%	15.04%	1.44%	3.77%	2.08%	6.25%	1.89%	2.90%
第 8 类	0.00%	0.00%	0.00%	0.00%	15.79%	0.00%	75.00%	15.79%	26.09%
第 9 类	77.20%	54.94%	64.20%	22.58%	12.39%	16.00%	73.94%	57.74%	64.84%

由表 5 中可以看出，普通的神经网络在处理这种高维度的数据时，很难取得令人满意的效果，而与支持向量机的结果相比较，虽然深度信念网络语支持向量机整体的数据结果相差不大，但在第一类、第二类以及第九类这种训练样本比较大的情况下，深度信念网络的准确率分别为 73.45%，74.94%，77.20%，而支持向量机只有 70.92%，60.41%，73.94%，而两种模型对应实验数据的召回率和 F 值则都相差不多，但对有样本值较少的类别，支持向量机模型则依旧展现出它传统的优势。例如对于句子“为我国高压输变电的顺利输送以及高速电气化铁路接触网的正常运行提供强有力的物质保障”，分词结果为“为/p 我国/n 高压/n 输变电/vn 的/ude1 顺利/ad 输送/v 以及/cc 高速/b 电气化/vn 铁路/n 接触网/n 的/ude1 正常/a 运行/vn 提供/v 强有力/bl 的/ude1 物质/n 保障/vn”，本文以“我国”“高压”“铁路”“接触网”以及“物质”五个词为核心，支持向量机的判别结果为“9,2,9,1,2”，深度信念网络将其判别为“9,2,1,1,2”,其中深度信念网络将“电气化铁路”成功结合在一起，而支持向量机则将二者分割开来。由于中文语义组块所要处理的是大规模的文本语句，采用深度信念网络优势明显。

4.2　实验性能分析

实验对深度信念网络与神经网络的训练时间和收敛速度进行了统计，比较了深度信念网络与神经网络在这两方面表现的差异。图 5 表示深度信念网络对神经网络参数自调整的过程方差次数变化图。

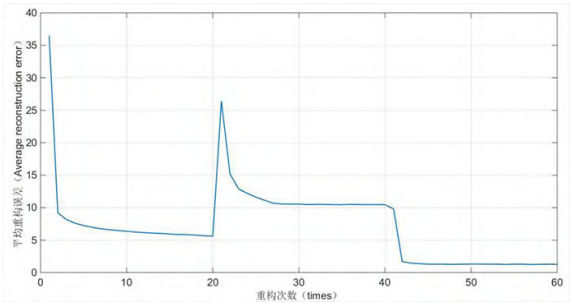

图 5 深度信念网络参数自调整过程

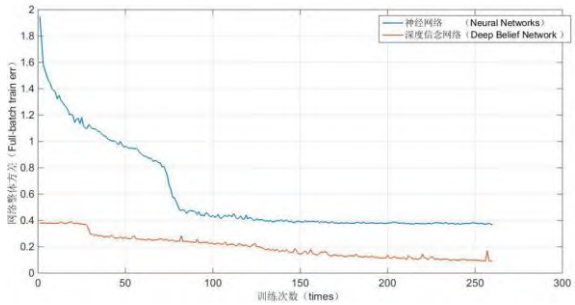

图 6 深度信念网络与神经网络参数方差收敛过程图

由图 5、6 的结果可以看出，与神经网络随机初始化网络权

值参数相比，深度信念网络通过预先的无监督训练来初始化网络参数极大地加快了网络的收敛速度，也同时提高网络对数据的拟合精度，避免了神经网络陷入局部最优的问题，从而提高了模型的准确度。

4.3 误差分析

通过对实验结果进行分析，实验结果的误差主要体现在如下几个方面：

a）实验语料的数据规模对网络权重的训练非常重要，但当前硬件条件下，无法进行更高数量级的训练。

b）人工标注过程中对不同中心词标注会出现一定偏差，对训练结果有影响。

c）实验的判别标注是在窗口下进行的，由于窗口大小一般小于句子长度，因此窗口大小的选取对实验结果有影响。

5 结束语

本文结合语法结构与词法结构，在语义层面上提出了一种介于词语和句子之间的新的语义表述单元，本文称之为中文语义组块，并使用神经网络、支持向量机、深度信念网络三种模型对中文语义组块结构进行了自动抽取，验证了深度信念网络在大数据层面下对中文语义组块抽取的优势。下一步的研究重点有两个方面：一方面是通过改进算法提高对中文语义组块自动抽取结果的准确度，另一方面是使用中文语义组块实现对句子语义的表述，提高中文文本语义表述的准确度。

参考文献

[1] 魏晓宁. 人工智能在自然语言理解技术上的应用[J]. 中国科技信息, 2005, 19 (2):57-57.

[2] Mel′čuk, I A. Dependency syntax : theory and practice[M]. New York:State University of New York Press, 1988.

[3] Sowa J F. Conceptual structures: information processing in mind and machine[M]// Conceptual Structures : Information Processing in Mind and Machine.[S.l.]: Addison-Wesley, 1984.

[4] Fellbaum C, Miller G. WordNet: an electronic lexical database[M]. Cambridge: MIT Press, 1998.

[5] Matuszek C, Cabral J, Witbrock M J, et al. An introduction to the syntax and content of Cyc.[C]//Proc of AAAI Spring Symposium on Formalizing and Compiling Background Knowledge and Its Applications to Knowledge Representation and Question Answering. 2006:44--49.

[6] 李珩, 朱靖波, 姚天顺. 基于 SVM 的中文组块分析[J]. 中文信息学报, 2004, 18(2):1-7.

[7] Chen N S, Kinshuk C, Wei W, et al. Mining e-learning domain concept map from academic articles[C]//Proc of International Conference on Advanced Learning Technologies. 2006:1009-1021.

[8] Mansouri A, Affendy L S, Mamat A. A new fuzzy support vector machine method for named entity recognition[C]//Proc of International Conference on Computer Science and Information Technology.2008: 24-28.

[9] 胡芳槐.基于多种数据源的中文知识图谱构建方法研究[D].上海：华东理工大学, 2015.

[10] 熊李艳, 陈建军, 钟茂生. 基于 E-A-V 结构的概念图匹配算法[J]. 计算机应用研究, 2014, 31(8):2290-2293.

[11] 陈宇, 郑德权, 赵铁军,等. 基于 DeepBeliefNets 的中文名实体关系抽取[J]. 软件学报, 2012, 23(10):2572-2585.