内蒙古自治区重点实验室 2018年度工作报告

实验室名称：内蒙古自治区蒙古文信息处理技术重点实验室

实验室主任：高光来

主管部门：内蒙古自治区科技厅

依托单位名称：内蒙古大学

通讯地址：呼和浩特市大学西路 235 号

邮政编码：**010021**

联系人：飞龙

联系电话：**13947124377**

E-mail 地址：**csfeilong@imu.edu.cn**

2018 年 12 月 23 日 填报

2018 年制

一、基本信息

实验室名称	中文：内蒙古自治区蒙古文信息处理技术重点实验室			
	英文：Inner Mongolia A.R. Key Laboratory of Mongolian Information Processing Technology			
研究方向（据实增删）	研究方向 1	蒙古文信息处理		
	研究方向 2	人工智能与模式识别		
	研究方向 3	图像处理与虚拟现实		
实验室主任	姓名	高光来	出生年月	1964 年 2 月
	职称	教授	专业领域	人工智能与模式识别
	任职时间	2016.1	在依托单位职务	内蒙古大学副校长
学术委员会主任	姓名	倪光南	出生年月	1939 年
	职称	研究员 院士	专业领域	电子
	任职时间	2007.5	所在单位及职务	中国科学院计算所

二、重点实验室年度情况

实验室经费（合计：万元）	经费构成	运行费（万元）	科研经费（万元）	仪器设备购置费（万元）	人员费（万元）
	国家		523		
	部门（地方）		866	300	
	依托单位		100		
	合计		1489	300	

科研条件（当前情况）	实验室面积	1100 平方米	
	科研仪器、设备累计	395 台（套）	900 万元（原值）
	大型仪器、设备（50 万元以上）累计	台（套）	万元（原值）

科研情况	项目（课题）	29 项	经费合计	1533 万元
	承担国家级项目（课题）	12 项	经费合计	523 万元
	承担省部级项目（课题）	15 项	经费合计	866 万元

	承担地市级项目（课题）			1 项		经费合计		100 万元		
	承担横向项目（课题）			1 项		经费合计		44 万元		
人才队伍	固定人员			17 人						
	高级职称	10 人	中级职称	7 人	初级职称	人				
	流动人员			人						
	高级职称	人	中级职称	人	初级职称	人				
	院士		固定	人	千人计划		固定	人		
			流动	人			流动	人		
	万人计划		固定	人	青年千人		固定	人		
			流动	人			流动	人		
	百千万人才		固定	人	杰青或优青		固定	人		
			流动	人			流动	人		
	省部级人才计划		固定	4 人						
			流动	人						
运行管理	管理制度		3 项		是否全部实施		是√否□			
	组建学术委员会		是√否□		召开会议次数		1 次			
开放共享	开放课题		项		经费合计		万元			
	仪器设施对外开放机时		小时		开展科普活动		次			

三、成果统计

获奖情况	国家级奖励	一等奖	项		二等奖		项	
	省、部级科技奖励	一等奖	项	二等奖	项	三等奖	项	
	行业科技奖励	一等奖	项	二等奖	项	三等奖	项	
论文专著	发表论文	共计	24 篇	SCI	1 篇	EI	17 篇	
	专著	国内出版	部		国外出版		部	
知识产权	发明专利	国际	项		国内		1 项	
	其它专利	国际	项		国内		项	

	标准规范	国际标准	个	国家标准	个
		行业标准	个	团体标准	个
产学研合作	与高校、院所合作	项	合作经费		万元
	与企业合作	项	合作经费		万元
行业支撑	成果转移转化	项	转移转化收入		万元
	行业技术服务	项	服务收入		万元

注：以上各表中所有数据指截止到统计年度所得数据或统计年度当年情况，项目经费指每个项目的总经费。

四、实验室本年度建设情况

简要介绍实验室本年度研发条件与能力、科研水平与贡献、团队建设与人才培养、开放交流与运行管理等情况。

内蒙古自治区蒙古文信息处理技术重点实验室于 2007 年 5 月获得内蒙古自治区科技厅批准，现任实验室主任为高光来教授。实验室主要的研究方向包括蒙古文信息处理、人工智能与模式识别、图像处理与虚拟现实。团队现有成员 17 人。从国内外引进学术骨干和优秀博士 8 人；现有博士学位成员 15 人，占总数的 88%；现有高级职称人员 10 人，占总数的 59%；现有 45 岁以下人员 14 人，占总数的 82%。成员中目前有享受国务院政府特殊津贴专家 1 人、自治区有突出贡献的中青年专家 2 人、入选自治区“草原英才”计划 2 人、自治区“新世纪 321 人才”一层次 1 人、入选自治区“草原英才”工程青年创新创业人才一层次 1 人。以本实验室为主要力量组成的“蒙古文软件研究与开发团队”入选 2012 年度自治区草原英才产业创新人才团队；本实验室和无线网络与移动计算重点实验室联合组成“内蒙古自治区网络协议工程与智能信息处理创新团队”。

2018 年，实验室共承担科研项目 29 项，包括国家自然基金 12 项，新增科研项目经费 547 万元。共发表学术论文 24 篇，其中 SCI 收录 1 篇、CCF B 类 3 篇、CCF C 类 8 篇。国家发明专利“一种蒙古文自动校正方法”获得授权，该发明提供了一种高效的蒙古文自动校正方法。利用该专利技术，实验室研发了蒙古文自动校正系统，该系统主要针对单词的拼写错误，控制符运用错误，格的错误使用，同形异音词的错误使用等四个方面的错误进行校正，进一步奠定蒙古文信息处理的技术基础。在

语音信号处理领域取得标志性成果，3 篇论文分别发表在本领域顶级国际期刊 IEEE Transactions on Audio, Speech and Language Processing 和顶级国际会议 ICASSP 2018。在自然语言处理领域取得重要成果，发表在自然语言处理重要会议 COLING2018 上。

团队建设方面，新增 1 名成员张晖博士，他 2017 年毕业于内蒙古大学，获工学博士学位。主要从事语音信号处理，语音分离、语音识别的研究，在国内外知名期刊会议上发表 10 余篇，主持国家自然科学基金一项，参与国家自然科学基金若干项。2018 年 7 月以青年英才身份引进到实验室工作。

人才培养方面，培养了 10 余名博士研究生，其中王炜华博士毕业；培养毕业了学术型硕士研究生 23 人、专业学位硕士研究生 18 人，毕业后大多在国内著名 IT 企业、国有企业、事业单位就职；高光来教授、飞龙副教授指导的硕士研究王洪伟获得内蒙古自治区优秀学位论文奖，博士研究生张晖获得内蒙古大学优秀学位论文奖。

开放交流方面，在 2018 年，实验室承办了第七届 CCF 计算语言学与中文计算国际会议（The Seventh CCF Conference on Natural Language Processing and Chinese Computing，NLPCC 2018）。NLPCC 是由计算机学会主办的自然语言处理旗舰会议，此次会议吸引了来自国内外 500 多名专家学者参加，大会主席是著名自然语言处理专家美国宾夕法尼亚大学教授 Dan Roth 和中科院自动化所宗成庆研究员。实验室主任高光来教授是大会的组织主席，全体老师和研究全程参与了会议的组织、服务工作，并在会议期间报告最新研究成果，展示了实验室最新开发的蒙古文相关系统与软件，并与国内外专家学者进行了广泛而深入的交流。此次会议还邀请到了加拿大西安大略大学终身教授 Charles Ling、美国密执安州立大学 Joyce Chai 教授、康奈尔大学 Cristian Danescu-Niculescu-Mizil、阿里巴巴达摩院自然语言处理首席技术官司罗进行了大会报告。派出博士生刘瑞参加自然语言处理重要会议 COLING2018,硕士生李庆龙参加语音信号处理顶级会议 ICASSP2018。另外，实验室还有 20 人次参加了其它国际和国内学术会议。

五、审核意见

实验室承诺所填内容属实，数据准确可靠。 实验室主任: （单位公章） 年　月　日
依托单位审核意见 依托单位负责人签字: （单位公章） 年　月　日
主管部门审核意见 主管部门负责人签字: （单位公章） 年　月　日

项目批准号	61866029
申请代码	F060405
归口管理部门	
依托单位代码	01002108A1582-1273

国家自然科学基金委员会

资助项目计划书

资助类别：地区科学基金项目

亚类说明：

附注说明：

项目名称：基于无标度复杂网络的文本内容分析与检索

直接费用：37万元　　执行年限：2019.01-2022.12

负 责 人：闫蓉

通讯地址：内蒙古呼和浩特市赛罕区大学西路235号内蒙古大学计算机学院

邮政编码：010021　　电　　话：04714993132

电子邮件：csyanr@imu.edu.cn

依托单位：内蒙古大学

联 系 人：牛一丁　　电　　话：0471-4994861

填表日期：2018年08月22日

国家自然科学基金委员会制

Version: 1.004.482

项目批准号	61866030
申请代码	F060306
归口管理部门	
依托单位代码	01002108A1582-1273

国家自然科学基金委员会

资助项目计划书

资助类别：地区科学基金项目

亚类说明：

附注说明：

项目名称：噪声环境下鲁棒性蒙古语语音识别技术研究

直接费用：37万元　　执行年限：2019.01-2022.12

负 责 人：张晖

通讯地址：内蒙古呼和浩特市大学西路235号

邮政编码：010021　　电　　话：15004712161

电子邮件：alzhu.san@163.com

依托单位：内蒙古大学

联 系 人：牛一丁　　电　　话：0471-4994861

填表日期：2018年08月26日

国家自然科学基金委员会制

Version: 1.005.530

项目批准号	61876214
申请代码	F060306
归口管理部门	
依托单位代码	01002108A1582-1273

国家自然科学基金委员会

资助项目计划书

资助类别：面上项目

亚类说明：

附注说明：常规面上项目

项目名称：知识引导的深度学习语音降噪研究

直接费用：62万元　　执行年限：2019.01-2022.12

负 责 人：张学良

通讯地址：内蒙古自治区呼和浩特市赛罕区大学西路235号

邮政编码：010021　　电　　话：0471-4992341

电子邮件：cszxl@imu.edu.cn

依托单位：内蒙古大学

联 系 人：牛一丁　　电　　话：0471-4994861

填表日期：2018年08月22日

国家自然科学基金委员会制

Version: 1.010.749

中华人民共和国国家版权局

计算机软件著作权登记证书

证书号：软著登字第2949001号

软 件 名 称：蒙古文依存树库显示和管理软件
V1.0

著 作 权 人：内蒙古大学

开发完成日期：2018年03月20日

首次发表日期：2018年03月20日

权利取得方式：原始取得

权 利 范 围：全部权利

登 记 号：2018SR619906

根据《计算机软件保护条例》和《计算机软件著作权登记办法》的规定，经中国版权保护中心审核，对以上事项予以登记。

中华人民共和国国家版权局
计算机软件著作权登记专用章
2018年08月06日

No. 02877704

证书号第3076057号

北京市海淀区地方税务局 印花税 代扣专用章

国家知识产权局

发明专利证书

发明名称：一种蒙古文自动校正方法

发明人：飞龙;路敏;高光来

专利号：ZL 2016 1 0706212.0

专利申请日：2016年08月22日

专利权人：内蒙古大学

地址：010021 内蒙古自治区呼和浩特市大学西街235号

授权公告日：2018年09月18日　　授权公告号：CN 106339367 B

本发明经过本局依照中华人民共和国专利法进行审查，决定授予专利权，颁发本证书并在专利登记簿上予以登记。专利权自授权公告之日起生效。

本专利的专利权期限为二十年，自申请日起算。专利权人应当依照专利法及其实施细则规定缴纳年费。本专利的年费应当在每年08月22日前缴纳。未按照规定缴纳年费的，专利权自应当缴纳年费期满之日起终止。

专利证书记载专利权登记时的法律状况。专利权的转移、质押、无效、终止、恢复和专利权人的姓名或名称、国籍、地址变更等事项记载在专利登记簿上。

局长
申长雨

国家知识产权局

2018年09月18日

中华人民共和国国家版权局

计算机软件著作权登记证书

证书号：软著登字第2444135号

软 件 名 称：基于Web的印刷体蒙古文字识别系统 V1.0

著 作 权 人：内蒙古大学

开发完成日期：2018年01月01日

首次发表日期：2018年01月01日

权利取得方式：原始取得

权 利 范 围：全部权利

登 记 号：2018SR115040

根据《计算机软件保护条例》和《计算机软件著作权登记办法》的规定，经中国版权保护中心审核，对以上事项予以登记。

中华人民共和国国家版权局
计算机软件著作权登记专用章
2018年02月22日

No. 02326790

中华人民共和国国家版权局

计算机软件著作权登记证书

证书号：软著登字第2948101号

软 件 名 称：蒙古文短语树库转换软件 V1.0

著 作 权 人：内蒙古大学

开发完成日期：2018年03月20日

首次发表日期：2018年03月20日

权利取得方式：原始取得

权 利 范 围：全部权利

登 记 号：2018SR619006

根据《计算机软件保护条例》和《计算机软件著作权登记办法》的规定，经中国版权保护中心审核，对以上事项予以登记。

中华人民共和国国家版权局
计算机软件著作权登记专用章
2018年08月06日

No. 02887229

中华人民共和国国家版权局

计算机软件著作权登记证书

证书号：软著登字第2947598号

软 件 名 称：蒙古文手写体数据采集软件 V1.0

著 作 权 人：内蒙古大学

开发完成日期：2018年03月20日

首次发表日期：2018年03月20日

权利取得方式：原始取得

权 利 范 围：全部权利

登 记 号：2018SR618503

根据《计算机软件保护条例》和《计算机软件著作权登记办法》的规定，经中国版权保护中心审核，对以上事项予以登记。

中华人民共和国国家版权局
计算机软件著作权登记专用章
2018年08月06日

No. 02887179

中华人民共和国国家版权局

计算机软件著作权登记证书

证书号：软著登字第2965053号

软 件 名 称：蒙古文依存句法分析软件 V1.0

著 作 权 人：内蒙古大学

开发完成日期：2018年03月20日

首次发表日期：2018年03月20日

权利取得方式：原始取得

权 利 范 围：全部权利

登 记 号：2018SR635958

根据《计算机软件保护条例》和《计算机软件著作权登记办法》的规定，经中国版权保护中心审核，对以上事项予以登记。

中华人民共和国国家版权局
计算机软件著作权登记专用章

No. 02884318

2018年08月09日

A LSTM Approach with Sub-word Embeddings for Mongolian Phrase Break Prediction

Rui Liu, Feilong Bao ✉, Guanglai Gao, Hui Zhang, Yonghe Wang
College of Computer Science, Inner Mongolia University,
Inner Mongolia Key Laboratory of Mongolian Information Processing Technology,
Hohhot 010021, China
liurui_imu@163.com;csfeilong@imu.edu.cn

Abstract

In this paper, we first utilize the word embedding that focuses on sub-word units to the Mongolian Phrase Break (PB) prediction task by using Long Short-Term Memory (LSTM) model. Mongolian is an agglutinative language. Each root can be followed by several suffixes to form probably millions of words, but the existing Mongolian corpus is not enough to build a robust entire word embedding, thus it suffers a serious data sparse problem and brings a great difficulty for Mongolian PB prediction. To solve this problem, we look at sub-word units in Mongolian word, and encode their information to a meaningful representation, then fed it to LSTM to decode the best corresponding PB label. Experimental results show that the proposed model significantly outperforms traditional CRF model using manually features and obtains 7.49% F-Measure gain.

1 Introduction

A Text-to-Speech (TTS) system converts the input text into synthetic speech with high naturalness and intelligibility. Naturalness is mainly influenced by the prosody modeling, especially by the Phrase Break (PB) prediction. Because the PB prediction is the first step of TTS, any error in this step will propagate to downstream steps such as intonation prediction and duration modeling. Those errors will result in the synthetic speech which is unnatural and difficult to understand. So that many researchers devote themselves to improving the performance of the PB prediction.

Typically PB prediction methods usually use machine learning models like Hidden Markov Models (HMMs) or Conditional Random Fields (CRFs) which trained with large sets of labeled training data. In these PB prediction models, the Part-of-Speech (POS) tag have been shown to be an effective feature and usually been included in the input feature set. The POS estimation itself is also a challenging task, and relies on large labeled training corpus, too. Its accuracy is always lower than our expectation, especially for those low-resource languages like Mongolian where the required linguistic resources are not readily available, and manual annotation is expensive and time-consuming.

In recent years, there are many works applying the word embedding techniques to Natural Language Processing (NLP) tasks, such as question answering, machine translation and so on (Bordes, 2014; Xiong, 2017; Devlin, 2014). Previous work has shown that the POS prediction task can be solved with high accuracy only using the word embedding feature as the input (Wang, 2015). POS information is most likely to be included in the word embedding representations. Therefore, some PB prediction systems which don't rely on the POS feature are developed (Watts, 2011; Vadapalli, 2014; Vadapalli, 2016). In (Watts, 2011), the authors obtain continuous-valued word embedding features that summarize the distributional characteristics of word types as surrogates of POS features. In (Vadapalli, 2014), researchers propose a neural network dictionary learning architecture to induce task-specified word embedding representations and show that these features perform better at PB prediction task. (Vadapalli, 2016) presents their investigations of recurrent neural networks (RNNs) for the phrase break prediction task by using word embedding. The above efforts have also been directed toward unsupervised methods of inducing word representations, which can be used as surrogates for POS tags, in the PB prediction task.

Proceedings of the 27th International Conference on Computational Linguistics, pages 2448–2455
Santa Fe, New Mexico, USA, August 20-26, 2018.

Training Supervised Speech Separation System to Improve STOI and PESQ Directly

Conference Paper · April 2018

DOI: 10.1109/ICASSP.2018.8461965

3 authors, including:

Hui Zhang
Inner Mongolia University
25 PUBLICATIONS **50** CITATIONS
SEE PROFILE

Xueliang Zhang
Inner Mongolia University
41 PUBLICATIONS **135** CITATIONS
SEE PROFILE

Online Direction of Arrival Estimation Based on Deep Learning

Conference Paper · April 2018

DOI: 10.1109/ICASSP.2018.8461386

CITATION
1

READS
405

3 authors, including:

Xueliang Zhang
Inner Mongolia University
41 PUBLICATIONS **136** CITATIONS
SEE PROFILE

Hao Li
Inner Mongolia University
9 PUBLICATIONS **13** CITATIONS
SEE PROFILE

Interspeech 2018
2-6 September 2018, Hyderabad

Improving Mongolian Phrase Break Prediction by Using Syllable and Morphological Embeddings with BiLSTM Model

Rui Liu, Feilong Bao ✉, Guanglai Gao, Hui Zhang, Yonghe Wang

College of Computer Science, Inner Mongolia University,
Inner Mongolia Key Laboratory of Mongolian Information Processing Technology,
Hohhot 010021, China

liurui_imu@163.com, csfeilong@imu.edu.cn

Abstract

In the speech synthesis systems, the phrase break (PB) prediction is the first and most important step. Recently, the state-of-the-art PB prediction systems mainly rely on word embeddings. However this method is not fully applicable to Mongolian language, because its word embeddings are inadequate trained, owing to the lack of resources. In this paper, we introduce a bidirectional Long Short Term Memory (BiLSTM) model which combined word embeddings with syllable and morphological embedding representations to provide richer and multi-view information which leverages the agglutinative property. Experimental results show the proposed method outperforms compared systems which only used the word embeddings. In addition, further analysis shows that it is quite robust to the Out-of-Vocabulary (OOV) problem owe to the refined word embedding. The proposed method achieves the state-of-the-art performance in the Mongolian PB prediction.

Index Terms: Mongolian, Syllable, Morphological, Phrase Break Prediction, BiLSTM

1. Introduction

Phrase break (PB) prediction is a crucial step in speech synthesis [1, 2]. It breaks long utterances into meaningful units of information and makes the speech more understandable. More importantly, in the context of speech synthesis, phrase breaks are often the first step for other models of prosody, such as intonation prediction and duration modeling [3, 4, 5]. Any errors made in the initial phrasing step are propagated to other downstream prosody models. Ultimately resulting in synthetic speech that is unnatural and difficult to understand.

Traditional PB prediction methods use machine learning models like Hidden Markov Models (HMMs) [6] or Conditional Random Fields (CRFs) [7, 8] which trained with large sets of labeled training data. Work in this area has traditionally involved linguistic features - for example, part-of-speech (POS), length of word etc [9, 10]. However, the linguistic features are discrete linguistic representations of words, which don't take into account the distributional behavior of words. Recent developments in neural architecture and representation learning have opened the door to models that can discover useful features automatically from the unlabelled data. With this development, word embedding [11] was proposed to learn distributed representation of word, which encodes a word as a real-valued low-dimensional vector. There are many works applying the word embedding techniques to Natural Language Processing (NLP) tasks, such as question answering, machine translation and so on [12, 13, 14]. Related ideas have been successfully applied to statistics parameter based and unit selection based speech synthesis system [15, 16]. Furthermore, for PB prediction task, some systems which do not rely on the linguistic feature are developed [17, 18, 19, 20, 21]. In [19], the authors obtain continuous-valued word embedding features that summarize the distributional characteristics of word types as surrogates of POS features. In [20], researchers utilize deep neural networks (DNNs) and recurrent neural networks (RNNs) to model PB by using word embeddings. Some further work can be found in [22] and [23]. In [22], authors obtain useful character embedding features to prediction PB in Chinese. In [23], a character-enhanced word embedding model and a multi-prototype character embedding model are proposed for Mandarin PB prediction.

All the methods mentioned has made great contributions, while they are not directly applicable to highly agglutinative languages such as Mongolian, Korean and Japanese for two reasons. First, sufficient training corpus is necessary for these methods to achieve such great performance, while the Mongolian training corpus is not very abundant; Second, such embeddings learned from these methods is unaware of the morphology of words. Mongolian is agglutinative in its morphology, words mainly contain different morphemes to determine the meaning of the word [24, 25, 26] hence increasing the vocabulary size for word embedding training and bring a considerably great challenge to train entire word-level distributed representation. Specifically, many suffixes can be in addition to word-stem to generate many new words. Its suffixes often serve as a positive signal which implies the POS of the word. It's like that the word implied by the suffix '-ly' is an adverb in English. For example: ᠬᠦᠮᠦᠨ ᠲᠤ᠋, ᠬᠦᠮᠦᠨ ᠳᠤ᠋, ᠬᠦᠮᠦᠨ ᠤ᠋, ᠬᠦᠮᠦᠨ ᠪᠦ᠋, ᠬᠦᠮᠦᠨ ᠪᠢ᠋. These words share the same word-stem "ᠬᠦᠮᠦᠨ" (Latin: "sandali", means: "chair"). In addition, a sequence of syllables forms a Mongolian word, and the composition of 2 or 3 characters forms a syllable. A single syllable possess a semantic meaning similar to morpheme. For instance, representation of "qihirag-tv", "qihiju", "qihitai" are constructed by the same syllables "qi" and "hi".

However, the Mongolian PB prediction research is at its initial stage compared with Chinese and English [27]. There are many works on Mongolian Text-to-Speech (TTS) which have made great contributions [28, 29], but the naturalness of synthetic speech is less than satisfactory especially without a good rhythm.

In this work, we leverage morphologic and syllable features to model Mongolian PB. We first use Bidirectional Long Short Term Memory (BiLSTM) networks to encode syllable and morphologic level information to capture the semantics of the word. Then we combine syllable, morphologic level representation and word level representation to an improved representation and

10.21437/Interspeech.2018-1706

Using Shifted Real Spectrum Mask as Training Target for Supervised Speech Separation

Yun Liu, Hui Zhang, Xueliang Zhang*

Inner Mongolia Key Laboratory of Mongolian Information Processing Technology,
Inner Mongolia University, Hohhot, China

liuyun.nogizaka@qq.com, alzhu.san@163.com, cszxl@imu.edu.cn

Abstract

Deep learning-based speech separation has been widely studied in recent years. Most of these kind approaches focus on recovering the magnitude spectrum of the target speech, but ignore the phase estimation. Recently, a method called shifted real spectrum (SRS) is proposed. Unlike the short-time Fourier transform (STFT), the SRS contains only real components which encode the phase information. In this paper, we propose several SRS-based masks and use them as the training target of deep neural networks. Experimental results show that the proposed target outperforms the commonly used masks computed on STFT in general.

Index Terms: shifted real spectrum, deep neural networks, training targets, speech separation

1. Introduction

Many speech applications, such as robust automatic speech recognition (ASR) and voice communication, need to acquire speech signals for further processing, but the signal of interest may be corrupted by additive background noise sometimes. To fight against the noise, speech separation aims to extract the target speech signal from a noisy speech. However, in the real environment, the speech separation performance is far from satisfactory, especially in the case of non-stationary noise and monaural conditions. This study focuses on monaural speech separation and non-stationary noise.

It is challenging that monaural speech separation uses only a single microphone to capture speech signal, while there are also many valuable methods have been proposed. Speech enhancement approaches [1, 2], such as spectral subtraction [3], estimate clean speech from noisy speech by estimating the noise firstly. In order to estimate the noise, speech enhancement approaches typically assume that the noise is stationary, therefore these methods cannot deal with the non-stationary noise. Computational auditory scene analysis (CASA) [4] tries to simulate the processing of the human auditory system which can solve the speech separation problem easily. CASA uses the ideal binary mask (IBM) [5] as the basic computational target. IBM performs well in both of the stationary and non-stationary noise conditions. Taking the IBM as the computational target leads the speech separation to be formalized as a supervised learning problem which could be solved with deep learning algorithms. Recently, with the development of supervised methods, the speech separation systems have achieved considerable performance improvements [6].

A typical supervised speech separation system usually learns a mapping function from noisy features to a training target with a supervised model, such as a deep neural network (DNN). The input features have been well-studied [7]. Amplitude modulation spectrogram (AMS) [8], mel-frequency cepstral coefficient (MFCC) [9], gammatone frequency cepstral coefficient (GFCC) [10], perceptual linear prediction (PLP) [11], and relative spectral transforms PLP (RASTA-PLP) [12] are commonly used features, and a complementary feature set included AMS, MFCC, GFCC and RASTA-PLP has been recommended in [7] and then been applied in many studies. The training targets have also been well-studied [13]. There are mainly two groups of training targets: mapping-based targets and masking-based targets. Mapping-based targets are the spectral representations of clean speech, such as the short-time Fourier transform (STFT) magnitude spectrum. Masking-based targets describe the relationships between clean speech and background interference in the time-frequency (T-F) domain, such as the IBM, ideal ratio mask (IRM) [14], FFT-mask [13], target binary mask (TBM) [15].

Most of these training targets only focus on the magnitude spectrum and ignored phase spectrum, because the early studies suggested that the phase is unimportant [16, 17], while recent studies suggest that phase is also important for perceptual quality [18]. Ignoring the phase will lead to degradation in the speech separation performance. The phase-sensitive mask (PSM) [19] and the complex ideal ratio mask (cIRM) [20] take the phase into their consideration and show better separation performance than the training targets without phase information. The cIRM is a complex mask, whose elements are complex numbers. Note that PSM is the real part of the cIRM.

In this study, we proposed a new training target with phase information. It is mainly inspired by the shifted real spectrum (SRS) [21] which is a spectral representation method in the real number field instead of STFT in the complex number field. On the basis of SRS, we proposed SRS-masks. Because SRS is in the real number field, all of the elements of SRS-mask are real numbers. Following the definition of cIRM and IRM, we define two versions of the SRS-mask. One is cIRM-like SRS-mask, called $cIRM_{srs}$. It contains the same information as the cIRM. Therefore, as the cIRM, $cIRM_{srs}$ is an optimal mask, we can perfectly reconstruct the speech signal from the $cIRM_{srs}$ and the noisy speech. Another version is IRM-like SRS-mask, called IRM_{srs}. As IRM, IRM_{srs} assumes the noise and speech are independent, so that the IRM_{srs} varies from 0 to 1, which makes it easy to model. Experimental results show that using the proposed SRS-mask can achieve better performance than the IRM which lacks of phase information, and achieve comparable performance compared to the cIRM and PSM which contain phase information. Further analysis indicates the IRM-like IRM_{srs} is a good tradeoff between the accuracy and modeling difficulty.

This research was supported in part by the China national nature science foundation (No. 61365006).

Mongolian Grapheme to Phoneme Conversion by Using Hybrid Approach

Zhinan Liu[1], Feilong Bao[1(✉)], Guanglai Gao[1], and Suburi[2]

[1] College of Computer Science,
Inner Mongolia University, Huhhot 010021, China
lzn_bung@163.com, {csfeilong,csggl}@imu.edu.cn
[2] Inner Mongolia Public Security Department, Huhhot 010021, China
sunbuer@163.com

Abstract. Grapheme to phoneme (G2P) conversion is the assignment of converting word to its pronunciation. It has important applications in text-to-speech (TTS), speech recognition and sounds-like queries in textual databases. In this paper, we present the first application of sequence-to-sequence (Seq2Seq) Long Short-Term Memory (LSTM) model with the attention mechanism for Mongolian G2P conversion. Furthermore, we propose a novel hybrid approach of combining rules with Seq2Seq LSTM model for Mongolian G2P conversion, and implement the Mongolian G2P conversion system. The experimental results show that: Adopting Seq2Seq LSTM model can obtain better performance than traditional methods of Mongolian G2P conversion, and the hybrid approach further improves G2P conversion performance. The word error rate (WER) relatively reduces by 10.8% and the phoneme error rate (PER) approximately reduces by 1.6% through comparing with the Mongolian G2P conversion method being used based on the joint-sequence models, which completely meets the practical requirements of Mongolian G2P conversion.

Keywords: Mongolian · Grapheme-to-phoneme · Sequence-to-sequence LSTM

1 Introduction

Grapheme-to-phoneme conversion (G2P) refers to the task of converting a word from the orthographic form (sequence of letters/characters/graphemes) to its pronunciation (a sequence of phonemes). It has a wide range of applications in speech synthesis [1–3], automatic speech recognition (ASR) [4–6] and speech retrieval [7, 8].

One of the challenges in G2P conversion is that the pronunciation of any grapheme depends on a variety of factors including its context and the etymology of the word. Another complication is that output phone sequence can be either shorter than or longer than the input grapheme sequence. Typical approaches to G2P involve using rule-based methods and joint-sequence models. While rule-based methods are effective to handle new words, they have some limitations: designing the rules is hard and requires specific linguistic skills, and it is extremely difficult to capture all rules for natural languages. To overcome the above limitations, another called statistics-based method are proposed,

M. Zhang et al. (Eds.): NLPCC 2018, LNAI 11108, pp. 40–50, 2018.
https://doi.org/10.1007/978-3-319-99495-6_4

Phonologically Aware BiLSTM Model for Mongolian Phrase Break Prediction with Attention Mechanism

Rui Liu, FeiLong Bao(✉), Guanglai Gao, Hui Zhang, and Yonghe Wang

Inner Mongolia Key Laboratory of Mongolian Information Processing Technology, College of Computer Science, Inner Mongolia University, Hohhot 010021, China
liurui_imu@163.com, csfeilong@imu.edu.cn

Abstract. Phrase break prediction is the first and most important component in increasing naturalness and intelligibility of text-to-speech (TTS) systems. Most works rely on language specific resources, large annotated corpus and feature engineering to perform well. However, phrase break prediction from text for Mongolian speech synthesis is still a great challenge because the data sparse problem due to the scarcity of resources. In this paper, we introduce a Bidirectional Long Short-Term Memory (BiLSTM) model with attention mechanism which uses the position-based enhanced phonological representations, word embeddings and character embeddings to achieve state of the art performance. The position-based enhanced phonological representations, derived from a separately BiLSTM model, are comprised of phoneme and syllable embeddings which take along position information. By using an attention mechanism, the model is able to dynamically decide how much information to use from a word or phonological component. To handle Out-of-Vocabulary (OOV) problem, we incorporated word, phonological and character embeddings together as inputs to the model. Experimental results show the proposed method significantly outperforms the systems which only used the word embeddings by successfully leveraging position-based phonologically information and attention mechanism.

Keywords: Mongolian · Phrase break · Phonologically
Attention mechanism · Position

1 Introduction

Phrase break plays an important role in both naturalness and intelligibility of speech [1]. It breaks long utterances into meaningful units of information and

This research was supports by the China national natural science foundation (No. 61563040, No. 61773224), Inner Mongolian nature science foundation (No. 2016ZD06) and the Enhancing Comprehensive Strength Foundation of Inner Mongolia University (No. 10000-16010109-23).

X. Geng and B.-H. Kang (Eds.): PRICAI 2018, LNAI 11012, pp. 217–231, 2018.
https://doi.org/10.1007/978-3-319-97304-3_17

Research on transfer learning for Khalkha Mongolian speech recognition based on TDNN

Linyan Shi, Feilong Bao✉, Yonghe Wang and Guanglai Gao
College of Computer Science,Inner Mongolia University,
Inner Mongolian Key Laboratory of Mongolian Information Processing Technology,
Hohhot 010021,China
shilinyan_2016@163.com; csfeilong@imu.edu.cn; cswyh92@163.com; csggl@imu.edu.cn

***Abstract*—Automated speech recognition(ASR) incorporating Neural Networks with Hidden Markov Models (NNs/HMMs) have achieved the state-of-the-art in various benchmarks. Most of them use a large amount of training data. However, ASR research is still quite difficult in languages with limited resources, such as Khalkha Mongolian. Transfer learning methods have been shown to be effective utilizing out-of-domain data to improve ASR performance in similar data-scarce. In this paper, we investigate two different weight transfer approaches to improve the performance of Khalkha Mongolian ASR based on Lattice-free Maximum Mutual Information(LF-MMI). Moreover, the i-vector feature is used to combine with the MFCCs feature as the input to validate the effectiveness of Khalkha Mongolian ASR transfer models. Experimental results show that the weight transfer methods with out-of-domain Chahar speech can achieve great improvements over baseline model on Khalkha speech. And transferring parts of the model performs better than transferring the whole model. Furthermore, the i-vector spliced together with MFCCs as input features can further enhance the performance of the acoustic model. The WER of optimal model is relatively reduced by 10.96% compared with the in-of-domain Khalkha speech baseline model.**

***Keywords*-Mongolian; speech recognition; weight transfer**

I. INTRODUCTION

With the emergence of the deep learning and its successful application in the fields of speech, the performance of the ASR system using neural network has been greatly improved. But the acoustic model of State-of-the-art ASR system has a strong dependence on manually annotated speech data. The lack of corpus is a significant problem in ASR systems on Khalkha Mongolian. Recently, transfer learning, an ability of transferring knowledge between two models, has become a popular method to improve acoustic model performance on lower resource languages in [1-3].

Literature [4] trained a deep neural network(DNN) model based on transfer learning to improve the accuracy of low-resource target language. But the acoustic model based on DNN using a fixed-size context window cannot represent more contextual information. By mining the temporal correlation of short-term features, time delay neural network(TDNN) can model long-term dependency information in [5]. However, being a feed-forward architecture, a large amount of context information is repeatedly calculated in adjacent time steps, which increases the complexity of the model. Later, a sub-sampling technique was proposed in [6] to reduce the computational complexity of the model. More recently, a chain model training strategy in [7] can significantly improve ASR accuracy.

Mongolian is a kind of cross-border language and has many dialects, such as Chahar dialect and Khalkha dialect. It is used in Mongolia, Inner Mongolia of China and some regions of Russia. The Mongolian in China is mainly Chahar dialect, which is the official Mongolian. It is usually written in traditional Mongolian letters, we call it Traditional Mongolian. Meanwhile, in Mongolia, Khalkha dialect is the standard Mongolian and usually written in Cyrillic Mongolian letters, we call it Cyrillic Mongolian. In this paper, we use "Chahar Mongolian" to represent the Chahar dialect written in traditional Mongolian and "Khalkha Mongolian" to represent the Khalkha dialect written in Cyrillic Mongolian. The speech recognition based on Chahar Mongolian started at 2003 in China and was established in [8]. Then some methods were proposed in [9, 10] to optimize the acoustic model. In Literature [11, 12], NN-based acoustic models were widely used in Chahar Mongolian ASR and achieved remarkable promotion. Recently, Hybrid Frame Neural Network in [13] achieved remarkable promotion with the WER score of the best system reaching 6.99%. However, the study of Khalkha Mongolian speech recognition system is still in its infancy.

In this paper, we proposed using Chahar Mongolian to initialize an acoustic model on Khalkha Mongolian and investigated two different weight transfer approaches of transferring knowledge. One approach is to transfer the whole network and share the phone set for both the Chahar Mongolian and the Khalkha Mongolian. Another is to transfer part of the network and then train it with single-stage or two-stage strategy. As a contrast, the randomly initialized models were trained in the Khalkha Mongolian. We find that chain TDNN with LF-MMI criterion can achieve better performance than cross-entropy(CE). We also find weight transfer model can get higher accuracy than the random initialization model. The single-stage strategy can get the optimal WER. Moreover, it was shown in [14] that i-vector features can enhance the adaptability of the model. We combined MFCC input features with i-vector as the input of both training on Khalkha Mongolian and Chahar Mongolian.

The structure of this paper is as follows. Section II detailedly introduces the models we used. Section III is the experimental setup. The results of the experiments is

978-1-7281-1175-9/18/$31.00 ©2018 IEEE

文章编号：1003-0077(2018)09-0028-07

基于 TDNN-FSMN 的蒙古语语音识别技术研究

王勇和，飞 龙，高光来

(内蒙古大学 计算机学院，内蒙古 呼和浩特 010021)

摘 要：为了提高蒙古语语音识别性能，该文首先将时延神经网络融合前馈型序列记忆网络应用于蒙古语语音识别任务中，通过对长序列语音帧建模来充分挖掘上下文相关信息；此外研究了前馈型序列记忆网络“记忆”模块中历史信息和未来信息长度对模型的影响；最后分析了融合的网络结构中隐藏层个数及隐藏层节点数对声学模型性能的影响。实验结果表明，时延神经网络融合前馈型序列记忆网络相比深度神经网络、时延神经网络和前馈型序列记忆网络具有更好的性能，单词错误率与基线深度神经网络模型相比降低 22.2%。

关键词：蒙古语；语音识别；时延神经网络；前馈型序列记忆网络

中图分类号：TP391　　文献标识码：A

Mongolian Speech Recognition Based on TDNN-FSMN

WANG Yonghe, BAO Feilong, GAO Guanglai

(College of Computer Science, Inner Mongolia University, Hohhot, Inner Mongolia 010021, China)

Abstract: In order to improve Mongolian speech recognition, the Time Delay Neural Network (TDNN) and Feed-forward Sequential Memory Network (FSMN) are combined to model the long sequence speech frames. In addition, we investigate the influence caused by the information from the preceding and the subsequent frames in the memory block over FSMN. We compare the performance of the TDNN-LSTM using different hidden layers and nodes. The results show that the fusion of TDNN and FSMN produces better performance than DNN, TDNN and FSMN, reducing the word error rate (WER) by 22.2% compared with the DNN baseline.

Key words: Mongolian; speech recognition; Time Delay Neural Network; Feed-forward Sequential Memory Network

0 引言

语音是人类最自然、便捷的交流方式，而语音识别技术，就是让机器能够“听懂”人类的语言并将语音信号转化为对应的文本或命令。基于高斯混合模型—隐马尔可夫模型(Gaussian Mixture Model-Hidden Markov Models, GMM-HMM)的语音识别框架在很长一段时间都是语音识别系统的主导框架，其核心就是用 GMM 对语音的观察概率进行建模，而用 HMM 对语音的转移概率进行建模[1]。近年来，深度神经网络(Deep Neural Network, DNN)[2]的研究和应用极大地推动了语音识别的发展，相比传统的基于 GMM-HMM 的语音识别系统，其最大的改变是采用 DNN 替换 GMM 对语音的观察概率进行建模来计算 HMM 状态的后验概率。根据文献[3]，基于 DNN-HMM 的声学模型采用固定长度的输入窗对语音的上下文特征进行建模，而语音是一种各帧之间具有很强相关性的复杂时变信号，所以这种方法不能充分利用语音的上下文时序信息。

相比 DNN，时延神经网络(Time Delay Neural Network, TDNN)[4]同样是一种前馈网络架构，它对每个隐藏层的输出都在时域进行扩展，即每个隐藏层接收到的输入不仅是前一层在当前时刻的输出，还有前一层在之前和之后的某些时刻的输出。在文献[5]中，通过选择正确的时间步长和对隐藏层输出进行降采样，TDNN 可以从输入上下文中的所有时间步长提取足够语音特征信息。因此，TDNN

收稿日期：2017-10-20　　定稿日期：2017-12-18

基金项目：国家自然科学基金(61563040, 61773224)；内蒙古自然科学基金(2016ZD06)

文章编号：1003-0077(2018)07-0044-08

蒙古文信息检索系统的设计与实现

温子潇，包飞龙，高光来，王勇和，苏向东

（内蒙古大学 计算机学院，内蒙古 呼和浩特 010021）

摘 要：该文针对传统蒙古文与西里尔蒙古文设计开发了一个功能完备的信息检索系统。在网页抓取方面，采用MD5算法对爬虫进行了改进，提升了爬虫的速度。在预处理阶段，对蒙古文文档进行了编码转换、词缀切分转换等操作。在检索方面，使用向量空间模型实现了对蒙古文文档的检索。在该文系统中加入了西里尔蒙古文到传统蒙古文转换和更新统计等模块，最终搭建了一个可以达到应用要求的蒙古文信息检索系统。

关键词：蒙古文；网络爬虫；信息检索系统

中图分类号：TP391　　文献标识码：A

Design and Implementation of Mongolian Information Retrieval System

WEN Zixiao, BAO Feilong, GAO Guanglai, WANG Yonghe, SU Xiangdong

(College of Computer Science, Inner Mongolia University, Hohhot, Inner Mongolia 010021, China)

Abstract: This paper presents a well-functioned information retrieval system for both traditional Mongolian and Cyrillic Mongolian. In the network crawling, MD5 algorithm is applied to improve the crawler performance. In the preprocessing, Mongolian documents are porcessed for code conversion, affix analysis and proofreading. The retrieval module is built upon the Vector Space Model. In addition, the Cyrillic Mongolian to the traditional Mongolian conversion module is developed to meet the application requirements.

Key words: Mongolian; Web crawler; information retrieval system

0 引言

随着科学技术的不断发展，互联网上的信息也在呈指数增长。目前，很多中英文信息检索系统层出不穷，但针对蒙古文的信息检索系统还不够完善，且相对较少。

蒙古文是蒙古族使用的语言文字，主要分布在中国的内蒙古自治区和蒙古国。中国与蒙古国使用的蒙古文字具有一定的差异。“语同文不同”，即指语言相同，但文字不同。蒙古国使用的蒙古文称为“西里尔蒙古文”（也称为新蒙文[1]），中国使用的蒙古文称为“传统蒙古文”（也称为旧蒙文或老蒙文）。随着信息的日益增长，蒙古文也急需一种信息检索系统，来满足人们的信息检索层次的需求[2]。

一些科研工作者对蒙古文信息检索系统进行了很多相关研究工作。金威[3]通过对传统蒙古文语法及构词进行详细分析后，解决了如何构建蒙古文索引词的问题。同时，搭建了一个较为完善的蒙古文信息检索平台。李业荣[4]根据传统蒙古文语言特点，利用信息检索技术实现了一个相对完善的蒙古文搜索引擎原型系统。刘娜[5]在基于传统蒙古文语义的基础上，利用信息检索模型，构建了蒙古文信息检索系统。以上研究工作均是基于传统蒙古文而言的，而基于西里尔蒙古文的信息检索系统研究成果还相对较少。上述研究人员不仅为蒙古文信息检索的发展起到了积极促进作用，还为本系统的构建提供了重要参考价值。

本文基于传统蒙古文和西里尔蒙古文，构建了一个性能优良的信息检索系统。该系统可以同时对传统蒙古文和西里尔蒙古文进行关键词检索。本文

收稿日期：2017-08-21　定稿日期：2017-09-08

基金项目：国家自然科学基金（61563040）；内蒙古自然科学基金重大项目（2016ZD06）；内蒙古自然科学基金（2017BS0601）

Combining Discrete Lexicon Probabilities with NMT for Low-Resource Mongolian-Chinese Translation

Li Jinting
Inner Mongolia University
Hohhot, China
justin_63@sina.com

Hou Hongxu
Inner Mongolia University
Hohhot, China
cshhx@imu.edu.cn

Wu Jing
Inner Mongolia University
Hohhot, China

Wang Hongbin
Inner Mongolia University
Hohhot, China

Fan Wenting
Inner Mongolia University
Hohhot, China

Ren Zhong
Inner Mongolia University
Hohhot, China

***Abstract*—Mongolian-Chinese neural machine translation (NMT) models often make mistakes in translating low-frequency words. We propose a method to alleviate this problem by improve NMT models with discrete translation lexicons that efficiently encode these low-frequency words. We describe a method to calculate the lexicon probability of generating the next word in the translation candidate by using the attention vector of the NMT model to select which source word lexical probabilities the model should focus on. The method use this probability as a bias to combine with the standard NMT probability. Experiments show an improvement of 4.02 BLEU score. We apply this method to large-scale corpus and improve the BLEU score. In addition, we also propose a novel approach to combine discrete probabilistic lexicons obtained from large-scale Mongolian - Chinese bilingual parallel corpus into NMT of small-scale corpus and enhance the performance of the system effectively.**

Keywords-Mongolian-Chinese Neural Machine Translation, Statistical Machine Translation, Discrete Lexicon Probabilities, Low-resource.

I. INTRODUCTION

Neural machine translation (NMT) [1-3], which directly models the translation process in an end-to-end way, has attracted intensive attention from the community. Although NMT has achieved state-of-the-art translation performance on rich-resource language pairs such as English-French and German-English [4-7], it still suffers from the unavailability of large-scale parallel corpus for translating low-resource languages. Due to the large parameter space, neural models usually learn poorly from low-count events, resulting in a poor choice for low-resource language pairs [8]. Zoph et al. (2016) indicate that NMT obtains much worse translation quality than a statistical machine translation (SMT) system on low-resource languages [9].

End-to-End Mongolian Text-to-Speech System

Jingdong Li, Hui Zhang, Rui Liu, Xueliang Zhang, Feilong Bao*

[1]College of Computer Science, Inner Mongolian University, Hohht 010021, China

jingdong.li@mail.imu.edu.cn, cszh@mail.imu.edu.cn

Abstract

Speech synthesis, or text-to-speech (TTS), generates a speech waveform of the given text. To build a satisfactory TTS system, a large natural speech corpus is requested. In the traditional approach, the corpus should be accompanied with precise annotations. However, the annotation is difficult and costly. Recently, end-to-end speech synthesis methods are proposed, which eliminated the requirement of annotation. The end-to-end methods make the development of TTS system less costly and easier. We used the state-of-the-art end-to-end Tacotron model in the Mongolian TTS task. With much more unannotated speech data (about 17 hours), the new system beats the old best Mongolian TTS system, which is trained with a small amount of annotated data (about 5 hours), with a big margin. The new mean opinion score (MOS) is 3.65 vs 2.08 which is the old one. The proposed system becomes the first Mongolian TTS system can be used in real applications.

Index Terms: Mongolian, speech synthesis, TTS, end-to-end

1. Introduction

Speech synthesis, or text-to-speech (TTS), generates a speech waveform of the given text [1]. TTS is one of the most important component of the voice user interface. With the widespread use of smart voice assistants (e.g. Siri from Apple), the research on speech synthesis draws more and more attentions [2]. Most traditional speech synthesis methods fall into two categories: unit-selection synthesis [3] and statistic parametric speech synthesis (SPSS) [4]. Unit-selection synthesis selectes appropriate sub-word units from large corpuses of natural speech, then concatenates the selected units to form the output. In contrast to selecting the actual speech instances, SPSS models the sub-words with parametric models, and generates speech from these parametric models. Unit-selection owns higher naturalness, while SPSS over the unit-selection synthesis in flexibility and controllability.

Both of the unit-selection synthesis and the SPSS request a large natural speech corpus. It appears that the larger the corpus the better the quality of the synthesized speech. Unfortunately, recording large corpus is very difficult and costly [5]. To build such a speech corpus, firstly, we need carefully select the sentences to make them phonetically and prosodically balanced. Secondly, we need carefully select a suitable speaker who should give a pleasant voice, with good voice quality and professional recording experience. Thirdly, we need to be equipped with professional recording devices and environment which can obtain high-quality noise-free recordings. Lastly and the most costly, we need to give a precise annotations on these recordings.

Data annotation for TTS is much difficult and costly than other applications. For example, speech recognition only needs to transcribe the recordings in word level. However, TTS system need to model a series expressive factors of speech, which includes intonation, stress, rhythm, and so forth. To fit the requirement of the TTS, we need to transcribe the recordings not only in word level, but also in sub-word (e.g. phoneme) level. The non-speech events like breathing or clicking also need to be picked out. The intonation, stress, rhythm, syllable and prosody need to be annotated, too. All of these annotations should be aligned to the time-line of the recording. The annotation boundaries also need careful fine turning. Because any misannotation often causes glitches in the synthesized speech, we need double check these annotations. All of these requirements make the data annotation difficult and costly. In our experience, the annotation of one hour's recording expends one man-month work, and involves at least two native speaking and professional annotators, where one for the first-phase annotation and another for revision. In average, one hour's TTS speech data cost over 1000 US dollars, in where the annotation cost the 95%.

The requirement of annotation much restricts the quality improvement of the TTS systems, especially for the languages whose resources are scarce. Mongolian is one of these languages. As far as our knowledge, the largest annotated Mongolian speech corpus, which can be used in TTS, only contains about 5 hours recordings. The state-of-the-art TTS system is built upon it only obtains 2.08 mean opinion score (MOS) [1], which means the perceived quality is "poor". In fact, this system cannot be used in any real applications. However, the requirement of a usable Mongolian TTS system is urgent. Mongolian is an influential language. There are about 6 million people who speak Mongolian language all over the word. Mongolian is one of the five major minority languages in China, and is one of the official languages in Inner Mongolia Autonomous Region of China. To improve the quality of the synthesized Mongolian speech, and develop a TTS system can be used in reality applications, we seek for some approaches which do not need the costly data annotation.

The end-to-end learning is a solution to our question. It takes all of multiple stages required by the conventional processing, and replaces them usually with just a single neural network. With the development of deep learning, end-to-end model have achieved significant improvement in many tasks in recent years [6–10]. Specifically, several end-to-end speech synthesis models has been successfully applied to English and other languages [11–14]. The end-to-end TTS system can be trained to predict audio from the text directly, which minimize the costly annotation work. In this work, we used the state-of-the-art end-to-end Tacotron model [14] in the Mongolian TTS task. Because the Tacotron model can be trained with <text, audio> pairs only, we can build a larger corpus with our limited budget. As a result, a Mongolian corpus is built, which contains about 17 hours recordings with word-level transcriptions. Although it smaller than the actually used English TTS corpus, it much larger than the existing largest

[1]See our experiments in section 4 for details.

Exploring Different Granularity in Mongolian-Chinese Machine Translation Based on CNN

Wang Hongbin, Hou Hongxu, Wu Jing, Li Jinting, Fan Wenting
College of Computer Science, Inner Mongolia University
Hohhot, China
cshhx@imu.edu.cn

***Abstract*—In this paper, a translation model based on Convolutional Neural Network (CNN) architecture is introduced into the Mongolian-Chinese translation task. Mongolian language has rich morphology structure, so we use byte-pair encoding (BPE) to segment the Mongolian word. In addition, the Mongolian Correction approach is adopted to reduce coding errors occurred in Mongolian corpus. The statistics data show that BPE and Mongolian Correction are alleviate the data sparsity that results from very low-resource Mongolian-Chinese parallel corpus. On Mongolian-Chinese translation task, we achieve the best result 35.37 BLEU that exceeds the baseline system by 1.4 BLEU. In the experiments, effect of different translation granularity on the translation result is investigated. The experiment results show that sub-word unit is more suitable than word unit for Mongolian-Chinese translation.**

Keywords-component; CNN, Mongolian-Chinese, Machine Translation, Low-resource, BPE, Mongolian Correction.

I. INTRODUCTION

Neural machine translation (NMT) is an end-to-end approach to machine translation. NMT has been synonymous with recurrent neural network (RNN) based encoder-decoder architectures [1]. In machine translation, this architecture has been demonstrated to outperform traditional phrase-based models by large margins [2][3][4]. The NMT based on RNN has been applied in the Mongolian-Chinese machine translation [5]. Recent work has applied convolutional neural networks (CNN) to sequence modeling such as [6][7][8]. Architecture is partially convolutional have shown strong performance on larger tasks but its decoder is still recurrent [9]. A fully convolutional architecture for sequence to sequence modeling was proposed [10]. We apply a NMT based on CNN to Mongolian-Chinese machine translation.

Compared to RNN-based NMT model, CNN creates representations for fixed size contexts that depend on the kernel width, however, the effective context size of the network can easily be made larger by stacking several layers on top of each other. CNN does not depend on the computations of the previous time step and therefore allows parallelization over every element in a sequence. This contrasts with RNN which maintains a hidden state of the entire past that prevents parallel computation within a sequence.

Sub-word units show its advantage when out-of-vocabulary (OOV) words and rare in-vocabulary words are translated by NMT, and that reducing the vocabulary size of sub-word models can actually improve performance [11]. Byte-pair encoding (BPE) is proposed in [11] that is used to segment words. For very low-resource Mongolian-Chinese translation task, there are a lot of rare words and OOV words. However, the Mongolian language an agglutinative language which words are made by concatenating morphemes. A number of morphemes are shared by words including rare words and OOV words. We segment Mongolian words via BPE to get sub-word units. For Chinese, we use Chinese characters as sub-word unit.

Another reason that there are a lot of Mongolian rare words is the coding errors occurred in Mongolian corpus, so we adopt Mongolian Correction approach [12] to process the Mongolian corpus.

As shown in Table I, BPE and Mongolian Correction reduce the number of low frequency tokens and reduce the size of source vocabulary. We combine BPE with Mongolian Correction to get lower size of source vocabulary and number of low frequency tokens than BPE or Mongolian Correction.

II. CONVOLUTIONAL NEURAL NETWORKS MACHINE TRANSLATION

We follow the NMT architecture by [9][10], which we will summarize here. The NMT system is implemented as an encoder-decoder network with CNN.

Position embeddings are useful in CNN translation model since they give the model a sense of which portion of the sequence in the input or output it is currently dealing with. First, we embed input elements $\mathrm{x} = (x_1, \ldots, x_m)$ in distributional space as $\mathrm{w} = (w_1, \ldots, w_m)$, where w_j is a column in an embedding matrix. Then, we embed the absolute position of input elements as $\mathrm{p} = (p_1, \ldots, p_m)$. Both are combined to obtain input element representations $\mathrm{e} = (w_1 + p_1, \ldots, w_m + p_m)$.

Both encoder and decoder networks share a simple block structure that computes intermediate states based on a fixed number of input elements. Each block contains a one dimensional convolution followed by a non-linearity. We denote the output of the lth block as $h^l = (h_1^l, \ldots, h_n^l)$ for the decoder network, and $z^l = (z_1^l, \ldots, z_m^l)$ for the encoder network. For a decoder network with a single block and kernel width k, each resulting state h_i^l contains information over k

Convolutional Neural Network for Machine-Printed Traditional Mongolian Font Recognition

Hongxi Wei[1,2(✉)], Ya Wen[1,2], Weiyuan Wang[1,2], and Guanglai Gao[1,2]

[1] School of Computer Science, Inner Mongolia University, Hohhot 010021, China
cswhx@imu.edu.cn
[2] Provincial Key Laboratory of Mongolian Information Processing Technology, Hohhot, China

Abstract. Although font recognition is a fundamental issue in the field of document analysis and recognition, it was usually ignored in the past. With the development of optical character recognition (OCR), font recognition becomes more and more important. This paper proposed a well-designed convolutional neural network (CNN) architecture for traditional Mongolian font recognition by means of a single word. To be specific, the whole word image is regarded as input of CNN. Hence, the word images should be normalized into the same size before being inputted into CNN. By comparison, an appropriate aspect ratio for the traditional Mongolian word images has been determined. Experimental results demonstrate that the proposed CNN architecture outperforms three classic CNN architectures, including LeNet-5, AlexNet and GoogLeNet. Therefore, the proposed CNN is much more suitable for the task of the traditional Mongolian font recognition in the way of a single word.

Keywords: Traditional Mongolian · Font recognition
Convolutional neural network · Word image · Aspect ratio

1 Introduction

A formal document generally contains multiple parts such as title, main body and so forth. These parts are usually edited in some certain fonts. The existence of multiple fonts makes the character recognition difficult, which results in decreasing the accuracy of optical character recognition (OCR) systems considerably. If the fonts are known before character recognition, an individual recognizer can be constructed for per font. Such the mono-font character recognition strategy can achieve higher accuracy. Moreover, reproduction of a digitized document requires the identification of the characters and the fonts used in the original document. Font recognition is very useful for determining the logical entities of a document including title, subtitle and paragraphs. Therefore, font recognition is able to not only improve the accuracy of OCR system, but also recover the layouts of a document exactly.

In the literatures, many approaches have been proposed for solving the problem of font recognition. Most of these approaches were applied to handling font recognition

L. Cheng et al. (Eds.): ICONIP 2018, LNCS 11305, pp. 265–274, 2018.
https://doi.org/10.1007/978-3-030-04221-9_24

2018 24th International Conference on Pattern Recognition (ICPR)
Beijing, China, August 20-24, 2018

Word Image Representation Based on Visual Embeddings and Spatial Constraints for Keyword Spotting on Historical Documents

Hongxi Wei, Hui Zhang, Guanglai Gao
School of Computer Science
Inner Mongolia University
Hohhot, China
cswhx@imu.edu.cn

Abstract—**This paper proposed a visual embeddings approach to capturing semantic relatedness between visual words. To be specific, visual words are extracted and collected from a word image collection under the Bag-of-Visual-Words framework. And then, a deep learning procedure is used for mapping visual words into embedding vectors in a semantic space. To integrate spatial constraints into the representation of word images, one word image is segmented into several sub-regions with equal size along rows and columns. After that, each sub-region can be represented as an average of embedding vectors, which is the centroid of the embedding vectors of all visual words within the same sub-region. By this way, one word image can be converted into a fixed-length vector by concatenating the corresponding average embedding vectors from its all sub-regions. Euclidean distance can be calculated to measure similarity between word images. Experimental results demonstrate that the proposed representation approach outperforms Bag-of-Visual-Words, visual language model, spatial pyramid matching, latent Dirichlet allocation, average visual word embeddings and recurrent neural network.**

Keywords—visual word; visual embeddings; spatial constraints; word image representation; query-by-example

I. INTRODUCTION

Until now, *optical character recognition* (OCR) is still a challenging task for historical documents due to degradation and low quality. When OCR is infeasible, keyword spotting is an alternative [1]. *Keyword spotting* can be defined as a task of image retrieval that relevant word images, similar to a given query word image, are obtained from a word image collection by image matching. Depending on the manner of providing query keywords, keyword spotting can be divided into two different approaches [2][3]: *query-by-example* (QBE) and *query-by-string* (QBS).

In the QBS approaches, query keyword is provided by text string [4-6]. However, the QBS approaches need to train a model to map from a text string to a word image on a large number of annotated word images. When there is no such annotated word images, the QBE approaches can be competent. The QBE approaches [7-10] require that an instance image of a query keyword is provided for being retrieved. In this study, we address a QBE based approach to accomplish the aim of keyword spotting. Therefore, it is assumed that historical document images have been segmented into corresponding word images. This paper mainly concentrates on how to represent word images and measure similarity between word images.

In the traditional keyword spotting, profile-based features are widely used to represent word images [11] and compared using dynamic time warping (DTW) algorithm [12][13]. Although the DTW algorithm works well, it is so computationally slow that cannot be suited for real-time matching. Recently, Bag-of-Visual-Words (BoVW) has been attracted much more attention and shown advantages in keyword spotting [14][15]. In the BoVW framework, each word image can be converted into a fixed-length histogram of visual words. Generally, cosine similarity between word images can be calculated on their histograms so that a ranked list of word images can be formed for a provided query keyword image. Hence, BoVW can be competent for real-time matching in the case of large-scale word image collections. However, visual words are independent each other in the BoVW framework, which results in not only discarding spatial orders of the neighboring visual words but also lacking semantic relatedness between visual words.

In this paper, we propose an approach to capture semantic relatedness between visual words. First of all, the corresponding visual words are extracted from a word image under the framework of BoVW. And then, a word image can be represented as a sequence of the labels of visual words along the writing direction. Given a collection of word images, all these sequences of the labels are concatenated together so as to form a training corpus. Next, a deep learning procedure is applied for the training corpus. By this means each visual word can be mapped to one vector in a semantic vector space. Consequently, the semantic relatedness between visual words can be measured by calculating Euclidean distance on their embedding vectors.

Meanwhile, each word image will be divided into a quantity of sub-regions with equal size along rows and columns. Such the spatial constraints are integrated into word

978-1-5386-3787-6/18/$31.00 ©2018 IEEE

Multiple Deep CNN for Image Annotation

Wei Wu and Deshuai Sun
Computer Science Department, Inner Mongolia University, Hohhot, China
Email: cswuwei@imu.edu.cn; 1169180464@qq.com;

ABSTRACT

Achieving better performance has always been an important research target in the field of automatic image annotation. This paper draws on the current popular deep learning model for the field of automatic image annotation. We propose a multiple convolutional neural networks (CNN) combination model for image annotation, which achieves satisfactory performance. First of all, we use three classical convolutional neural networks, and subsequently we examine the annotation accuracy for each CNN model. Then we take full advantage of the powerful feature representation capabilities of deep CNN, thus the last two layers of the deep CNN are extracted for each model and merged to form a new combined feature. Finally, we form our combination models by concatenating these features from each CNN model, and utilize these concatenated features to linear SVM classifier for image annotation. Experimental results on ImageCLEF2012 image annotation dataset illustrate that our combination method outperforms the traditional classifiers and the individual CNN models.

Keywords: Image annotation, deep learning combination model, CNN, SVM.

1. INTRODUCTION

Automatic image annotation or classification has always been a research hotspot in the field of computer vision, and has attracted wide attention in both research and industry. The goal of automatic image annotation is to automatically identify semantic visual concepts from the images, including natural scenes, objects, events (sports, travel), and even expressions (happy, unpleasant, etc.). Due to large intra-class variations and inter-class similarities, the study of automatic image annotation is very challenging. In recent years, many research groups have been engaged in this work, and there are several well-known large image data sets, such as ImageCLEF [1], MSCOCO [2] and ImageNet [3], which confirm the challenges in this field.

Early image automatic annotations are concerned with the overall category of images, which is closer to image classification. And some works show promising results on these traditional image annotation tasks, such as probabilistic latent semantic analysis (PLSA model) [4], cross media relevant model (CMRM model) [5], and sparse kernel learning for continuous relevance model (SKL-CRM model) [6]. But all of these methods rely on various global and local features of the image, which limits the accuracy of the annotation when the selected image features are not appropriate.

At present, the latest research progress in the field of image annotation is to perform large-scale, semantically conceptual annotation for natural scene images. The annotation semantics are not only limited to specific single physical objects, such as "cats", "dogs", and "cars", etc., but also includes various abstract semantic concepts, such as "animals", "children", "traffic", "sports", and "play", etc., which increases the difficulty of image annotation [7-9].

In recent years, deep learning technology has achieved great success in the field of computer vision. In 2012, Hinton et al. applied Deep Learning to the field of image recognition [10], and achieved amazing results on the large-scale image database ImageNet, the recognition error rate was reduced to 15%. Since then, deep learning technology has opened up a wave of academia and industry. And many new CNN models are presented for image classification, scene recognition, and object detection etc., all these models are performed well on the ImageNet dataset. The most famous of these models are VGGNet [11], GoogleNet [12] and ResNet [13], etc.

In this paper, we mainly use three CNN models including LeNet5 [14], AlexNet [10] and VGGNet [11] combined together for image feature extraction, which is used to construct traditional SVM classifier for image annotation. ImageCLEF2012 dataset is experimented for validating our method, and experimental results show that the proposed method achieves higher accuracy than the traditional classifiers and the individual CNN models.

TRAINING SUPERVISED SPEECH SEPARATION SYSTEM TO IMPROVE STOI AND PESQ DIRECTLY

Hui Zhang, Xueliang Zhang, Guanglai Gao*

College of Computer Science, Inner Mongolia University, China

alzhu.san@163.com, {cszxl, csggl}@imu.edu.cn

ABSTRACT

Supervised speech separation methods train learning machine to cast the noisy speech to the target clean speech. Most of them use mean-square error (MSE) as loss function. However, MSE is not the perfect choice because it doesn't match the human auditory perception. Short-time objective intelligibility (STOI) and perceptual evaluation of speech quality (PESQ) are closely related to the human auditory perception and widely used in speech separation research as evaluation criteria. Therefore, STOI and PESQ may be better choices for the loss function. However, they are nondifferentiable functions which cannot be optimized by the conventional gradient descent algorithm. In this work, a gradient approximation method is used to calculate the gradients of the STOI and PESQ. Then the calculated gradients are used in the gradient descent algorithm to optimize the STOI and PESQ directly. Experimental results show the speech separation performance can be improved by the proposed method.

Index Terms— Monaural speech separation, Short-time objective intelligibility (STOI), perceptual evaluation of speech quality (PESQ), gradient approximation

1. INTRODUCTION

Monaural speech separation separates target speech from additive noise signal by using only one microphone. It has been widely studied to improve the performance of various signal processing systems, including hearing prosthesis, mobile telecommunication, and robust automatic speech and speaker recognition [1]. For a few decades, monaural speech separation systems have achieved considerable performance improvements, especially after formalizing it as a supervised learning problem and using deep learning algorithms.

Early studies for monaural speech separation, e.g. spectral subtraction, are mostly based on the mean-square error (MSE) criterion [2] which can improve the perceptual speech quality. However, these approaches typically assume that background noise is stationary, i.e. its spectral properties do not change over time, or are stationary than the speech at least. Therefore, they have difficulties in tracking non-stationary noises, which limits its application in real-world environments.

In order to enhance the noisy speech in various noisy environments, more powerful models are involved. Deep neural networks (DNNs) and long short term memory networks (LSTMs) can model the complicated relationship between the input variables and the output targets. They were successfully introduced to the speech separation area as supervised speech separation, and obtained considerable performance improvements. In these approaches, a learning machine (DNN or LSTM) is trained to cast the acoustic features of the noisy speech to a time-frequency mask, or the spectrum of the clean speech, where these two categories methods can be generally referred as the masking-based and the mapping-based methods. Many works devoted to the supervised speech separation, which covered the most aspects of the supervised learning: Wang concluded the related works on features [3] and training targets [4], and many works studied the learning machine and its training methods [5–9]. But very few studies investigated the loss function, and most of the learning-based method employ MSE. For example, the masking-based method minimizes the MSE between the estimation and the ideal mask target, and the mapping-based method minimizes the MSE between the estimation and the target clean spectrum.

Although a lot of works show the effectiveness of the MSE. In fact, the MSE is not a perfect loss function to evaluate the estimation, because it is not closely related to the human auditory perception. The MSE has two weaknesses: it treats the estimation elements independently and equally. a) the MSE will lead to over-smooth speech trajectories and may result in muffled sound quality and decreased intelligibility [6]. Because the MSE measures are derived from each time-frequency (T-F) unit separately rather than from whole spectral trajectory. b) it treats every estimation elements with equal importance, in fact, they are not. For speech intelligibility, the distinguishable phones are more important, and for speech quality, the isolated points are more harmful which may lead to musical noise. The MSE is usually defined in the linear frequency scale, but the human auditory perception follows the Mel-frequency scale. Therefore improving the human auditory perception quality

Citation Count Prediction Based on Academic Network Features

XinPing Zhu
School of Computer Science
Inner Mongolia University
Hohhot, China
zhuxinping1993@163.com

ZhiJie Ban
School of Computer Science
Inner Mongolia University
Hohhot, China
banzhijie@imu.edu.cn

***Abstract*—Citation count is an important factor to measure the influence of academic publications. Identifying future citation count in advance can help scientists to find references and research area. There are many academic network features which are related to citation count. However, these features have not been completely explored in the existing studies. In this paper, we propose a citation count prediction model based on academic network features. Firstly, some important features are introduced and analyzed in detail. Then, we verify the importance of each feature and use a neural network model to select a set of optimal features. Finally, we present several machine learning methods and one multiple linear regression strategy to predict a paper's future citation. Experimental results on real datasets demonstrate that our model significantly outperforms the baseline method.**

Keywords—citation count prediction; academic social network; feature selection;

I. INTRODUCTION

Usually, researchers tend to focus on the current influential papers. There are many indicators to measure the influence of a paper, such as paper's citation count, author's *h*-index, journal's impact factor and so on. In these indicators, citation count is the simplest measurements. The definition of citation count is the number of times that a paper is cited by other publications. In this paper, we use the citation count to represent the influence of papers. Highly cited papers indicate the recognition among peers.

Due to the rapid development of scientific research, the volume of publications increases exponentially every year. For researchers, they are not likely to read each literature in the data set of publications. The result may cause them to miss important references. Meanwhile, there are only a handful of papers that promote the development of the field. Figure 1 counts the total number of papers published every year in Computer Science [1]. The number of papers in 2009 was almost three times than that of 10 years ago. Identifying potentially influential papers in advance can help researchers to choose references and research area. Therefore, effectively predict the citation count of papers in the future is of great significance.

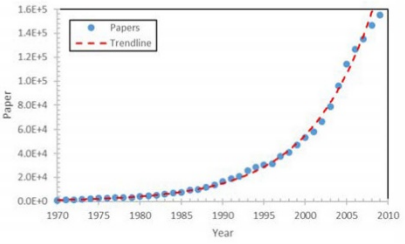

Fig. 1. The total number of papers published each year.

The 2003 KDD Cup introduces a citation prediction task that estimates the change in the number of citations of papers between two different periods of time [28]. After that, the researchers studied the similar tasks from different aspects. Yan et al. consider several regression models and propose several features to predict the number of citations [1]. The effect of his experiment is not ideal in the prediction of shorter time and he doesn't use any features constructed from the citation network. Using traditional regression analysis to estimate the citation count is exceedingly difficult [31]. Instead, Yuxiao Dong et al. use linear regression to predict the change of *h*-index [32]. Unlike previous studies, we consider some features related to the academic social network and utilize the neural network model to select a set of optimal features and then utilize machine learning methods to predict the exact citation count for each individual paper.

Combined with previous work, our model includes the following steps:

- Some academic network features are introduced and analyzed, such as author, paper, venue and network's related features. Then, we utilize the neural network model to verify the importance of each feature and select a set of optimal features.
- We present different methods (e.g. deep neural network, support vector machine, and multiple linear regression) to predict a paper's future citation count.

1550-445X/18/$31.00 ©2018 IEEE
DOI 10.1109/AINA.2018.00084

贝叶斯网络在信息检索中的应用

郑伟[1,2]　侯宏旭[1]　武静[1]

（1.内蒙古大学，计算机学院，呼和浩特010021；2.河北北方学院，理学院，河北张家口 0750002）

摘要:[目的/意义]贝叶斯网络是描述随机变量之间依赖关系的图形模式，被广泛应用于不确定性问题的智能求解。[方法/过程]文章介绍了信息检索领域中基于贝叶斯网络技术的三种检索模型-推理网络模型、信任度网络模型、贝叶斯网络检索模型,详细地分析了其工作原理、论述了国内外研究者使用贝叶斯网络在信息检索领域的研究现状，探讨了每种模型的优势与不足。[结果/结论]指出了贝叶斯网络技术在信息检索领域的发展趋势.

关键词：　贝叶斯网络；信息检索；术语；文档

中图分类号：G350 **文献标识码**：A

Application of Bayesian Network for Information Retrieval

ZHENG Wei[1,2]　HOUHong-xu[1]　WU Jing[1]

（1. College of Computer　Science, Inner Mongolia　University, Hohhot 010021, China；2.College of Science, Hebei North University, Zhangjiakou 075000, China)

Abstract:[Purpose/significance]Bayesian network is graphical model to describe dependencies between random variables and is widely used to solve uncertainty problems in intelligent. [Method/process] Article introduces three kinds of retrieved model, named reasoning network model, belief network model and Bayesian network retrieval model, and analyses their working principle detailed, and discusses research status of using Bayesian network in information retrieved field, and discusses the advantage and shortcoming of each model. [Result/conclusion] The future research trend has pointed out about information retrieved field using Bayesian network.

Keywords: Bayesian network ， Information retrieval，term，Document

1　贝叶斯网络概述

贝叶斯网络(Bayesian Network, BN)是一种有向无环图，可表示为 $G=(V,E)$，其中 $V(x_1,x_2,\cdots,x_n)$，x_i表示要解决问题的随机变量节

收稿日期：2017-07-14；**修返日期**：

基金项目：国家自然科学基金项目(编号：61362028)，张家口市科学技术研究与发展计划项目(编号：1611056)

作者简介：郑伟(1978－)，男（蒙古族），内蒙古呼伦贝尔市人,副教授，博士生，主要研究领域为信息检索、文本分类、数据挖掘；侯宏旭 1972-），男，内蒙古呼和浩特人，教授，博士生导师，主要研究方向为自然语言处理，信息检索；武静（1989-），女，内蒙古呼和浩特人，博生生，主要研究方向为自然语言处理，机器翻译。

文章编号：

基于统计和神经网络的蒙汉机器翻译研究

任众[1] 侯宏旭[1] 武静[1] 王洪彬[1] 李金廷[1] 樊文婷[1] 申志鹏[1]
（1.内蒙古大学 计算机学院，内蒙古自治区 呼和浩特，010021）

摘要： 本文对基于传统统计模型的蒙汉机器翻译模型和基于神经网络机器翻译模型进行了研究，其中神经网络翻译模型分别为基于 CNN，RNN 的翻译模型，并通过将所有翻译模型结果进行句子级融合得到一个融合模型。针对蒙汉翻译面临资源稀少、蒙古文形态复杂等困难，我们提出多种翻译技术，对各个模型进行改进，并对蒙古文进行形态分析与处理。在翻译效果最好的 CNN 模型上，我们采用字和短语融合训练方法；基于 RNN 的翻译模型除用上述方法外，还采用 Giza++指导对齐技术调整 RNN 注意力机制；SMT 我们采用了实验室提出的重对齐技术。本文对实验结果进行了对比和分析，这三种技术方法对相应系统翻译效果有显著提升。此外，蒙古文形态分析与处理对缓解数据稀疏，提升译文质量也有重要作用。

关键词： 汉蒙机器翻译；蒙古文形态分析；融合训练方法

中图分类号： TP391 **文献标识码：** A

Research on Mongolian-Chinese MT Based on Statistical and Neural Network

Ren Zhong[1], Hongxu Hou[1], Jing Wu[1], Hongbin Wang[1], Jinting Li[1], Wenting Fan[1], Zhipeng Shen[1]
（1.Inner Mongolia University, Hohhot, 010021,China）

Abstract： In this paper, the Mongolian-Chinese machine translation model and the neural network-based machine translation model based on the statistical model are studied. The neural network translation models are respectively based on the CNN and RNN translation models. A fusion model are obtained by sentence-level fusion of all the translation model. The lack of resource is the main obstacle Mongolian-Chinese faces. Besides, the Mongolian morphology is complex. To tackle these, we proposed multiple methods to improve the three translation models. For the best performance CNN model, we use character and phrase joint-training method; we also use this method as well as a Giza++ guided alignment to RNN model; we also use a realignment method to the SMT model. This report evaluates these methods with more contrast experiments. We also analyze and process the Mongolian morphology to alleviate the data sparsity. The proposed methods and the Mongolian morphology process improved the Mongolian-Chinese translation performance significantly.

Key words: Mongolian-Chinese machine translation; Mongolian morphology process; Joint-training method

0 引言

蒙汉机器翻译属于稀少资源及少数民族语言翻译领域任务，对于促进语言、文字和文化交流，以及民族团结进步具有重要意义。然而此类翻译任务普遍面临双语对齐语料不足，资源稀少，蒙古文形态复杂，翻译研究时间短，成果少等困难。

本文中共有四个系统分别为：CNN(Convolutional Neural Network)，RNN(Recurrent Neural Network)，SMT (Statistical machine translation)系统和以上三个系统的句子级融合系统。其中 CNN 取得最好的翻译效果（BLEU5-SBP=0.7024），其后依次是融合系统、RNN 系统以及 SMT 系统。蒙汉翻译任务主要面临的困难是资源稀少和蒙古文形态复杂，针对这

收稿收稿日期： ；**定稿日期：**
基金项目： 国家自然项目基金(61362028)

文章编号：1003-0077(2018)12-0000-00

基于主题网络的伪主题分析

闫 蓉[1,2]，高光来[1,2]

（1. 内蒙古大学 计算机学院，内蒙古 呼和浩特 010021；
2. 内蒙古自治区蒙古文信息处理技术重点实验室，内蒙古 呼和浩特 010021）

摘 要：传统无监督的主题建模方法利用相互独立的主题变量抽象描述文本语义，忽略了各主题内部隐含的结构和联系，粗粒化的文本主题分析加剧了“强制主题”问题对文本建模的影响。本文通过研究主题网络社区内部结构，结合主题内部语义耦合关系与网络拓扑结构，提出伪主题分析方法来识别和解释主题，实现从网络结构角度描述文本语义特征，弥补统计主题分析方法对文本语义结构刻画的不足。

关键词：伪主题分析；主题网络；文本理解

中图分类号：TP391　　**文献标识码**：A

Pseudo Topic Analysis Based on Topical Networks

YAN Rong[1,2], GAO Guanglai[1,2]

(1. College of Computer Science, Inner Mongolia University, Hohhot, Inner Mongolia 010021, China;
2. Inner Mongolia Key Laboratory of Mongolian Information Processing Technology,
Hohhot, Inner Mongolia 010021, China)

Abstract: Traditional unsupervised topic models usually represented the document semantic by using a set of topics where no relationships between the topics, which would result in the imperfection to the text topic modeling in a coarse-grained manner, and intensify the effect of the 'forced topic' problem for the text topic modeling because ignoring the internal structure and relationships within each topic. Based on the study of the community inner structure, and combined with the internal coupling relationships and network topology, this paper proposed a novel pseudo topic analysis approach. It achieved identify and explain the topic, and accomplished the description the textual semantic features from the network structure point of view so as to remedy the deficiency of the textual semantic structure of the statistical topic modeling methods.

Key words: pseudo topic analysis; topical network; text understanding

1 引言

概率主题模型，如LDA(Latent Dirichlet Allocation)[1]和PLSA(Probabilistic Latent Semantic Analysis)[2]为用户在海量信息中筛选和挖掘有效信息发挥了重要的作用[3]。已经有很多工作致力于构建新的主题模型和改进算法来捕获主题结构[4-6]和实现主题模型的可视化[7-9]。但是，该类文本主题分析技术多数是利用统计方法实现文本主题获取，通常考虑词频较大的词项对于文本内容的贡献，核心假设是利用文本集中包含特定数目的潜在主题变量，来构建文本语义描述空间。这些数目的潜在主题变量在表达文本集固有抽象的同时，也利用多个不同主题变量抽象地表示文本的不同语义，实现了文本间的区别。但是这种方法由于受到其概率主题建模机理的限制，文本主题分析结果并不理想，原因有三点，分别是：第一，利用统计方法获取这些潜在主题变量的同时，假设各潜在主题变量之间是相互独立的，尽管各潜在主题变量之间有结构，但是潜在

收稿日期：2017-00-00　**定稿日期**：2017-00-00

基金项目：国家自然科学基金项目(61662053)；内蒙古自然科学基金项目(2018MS06025)；内蒙古大学高层次人才项目(21500-5175128)

文章编号：1003-0077（2017）00-0000-00

融合先验信息的蒙汉神经网络机器翻译模型

樊文婷[1]，侯宏旭[2]，王洪彬[3]，武静[4]，李金廷[5]

（1.内蒙古大学 计算机学院，内蒙古 呼和浩特，010021）

摘要：重现神经网络机器翻译模型在蒙古文到汉文的翻译任务上取得了很好的效果，然而，神经网络翻译模型仅利用双语语料获得词向量，有限的双语语料规模限制了词向量的表示。该文将先验信息融合到神经网络机器翻译中，首先将大规模单语语料训练得到的词向量作为翻译模型的初始词向量，同时在词向量中加入词性特征，从而缓解单词的语法歧义问题。此外，神经网络翻译模型为了控制训练规模和训练时间，通常会限制目标词典大小，这导致大量未登录词的出现。该文利用加入词性特征的词向量计算单词之间的相似度，将未登录词用目标词典中与之最相近的单词替换，以缓解未登录词问题。最终实验结果在蒙古文到汉文的翻译任务上将译文的 BLEU 值提高了 2.68 个点。

关键词：重现神经网络；未登录词；词向量；词性标注

中图分类号：TP391 **文献标识码**：A

Mongolia-Chinese Neural machine translation with priori information

FAN Wen-ting[1],HOU Hong-xu[1],WANG Hong-bin[1],WU Jing[1], LI Jin-ting[1]

(1. College of Computer Science, Inner Mongolia University, Hohhot 010021, China)

Abstract：Neural machine translation (NMT) has become an extremely prominent model in Mongolian-Chinese translation task. We implement neural machine translation model with priori information. On the one hand, we train word representations using large - scale monolingual corpus to act as the initial word vectors. On the other hand, we add part-of-speech feature for word vector to solve the problem of grammatical ambiguity. NMT usually limits the target vocabulary size. To solve the out of vocabulary problem, we use word embedding to calculate the similarity of words, then replace the out-of-vocabulary words by the most similar words who are covered by the target vocabulary to improve the utilization of the vocabulary. In the task of Mongolian-Chinese machine translation, experimental results show that BLEU increased 2.68 points.

Keywords: Recurrent neural network; Out-of-vocabulary; Word embedding; Part-of-speech

1 引言

神经网络机器翻译模型被提出之后成为了机器翻译的一个研究热点[1][2]。神经网络机器翻译模型是基于词、短语和句子的连续表示，连续的词向量更准确的表示词的形态、语义和语法信息，刻画近义词之间的关系[3][4]。蒙古文的构词形式

收稿日期：2017-03-16；**定稿日期**：2017-04-26

基金项目：国家自然科学基金项目（61362028）；