社交文本规范化研究综述

罗程多[1] 吴晓蕊[2] 薛 凯[2] 杨 飞[2] 王保录[2]

([1]北京邮电大学 网络与交换技术国家重点实验室 北京 100876
[2]空间物理重点实验室 北京 100076)

摘要:社交媒体中的海量文本已经成为自然语言处理和数据挖掘领域的重点研究对象。然而社交文本中存在的不规范特征是对文本进行处理和挖掘的重要障碍。消除这一障碍的方法之一是将不规范的文本转化成规范的形式,即社交文本规范化。本文将对社交文本规范化的基本方法和研究现状进行综合介绍,同时也对社交文本规范化未来的研究方向进行了讨论。

关键词:社交文本,文本规范化,微博文本

An Overview of Social Text Normalization

LUO Chengduo[1], WU Xiaorui[2], XUE Kai[2], YANG Fei[2], WANG Baolu[2]

([1]Beijing University of Posts and Telecommunications, State Key Laboratory of Networking and Switching Technology, Beijing,100876, China,
[2]Science and Technology on Space Physics Laboratory, Beijing, 100076, China)

Abstract: With the popularity of social media, the huge volume of the social media text becomes the important research object of natural language processing and data mining communities. However, the informal features of social text are significant obstacles to process and mine the text. One way to overcome the obstacles is to transform the informal text into formal one. This process is called social text normalization. This paper reviews the basic methods and recent advances of social text normalization research. And it discusses the future research direction of social text normalization.

Keywords: social text, text normalization, micro – blog text

随着社交媒体的普及,各个社交平台上每天都会产生海量的文本数据。这些文本数据所蕴含的信息对政治、商业、灾害应对等诸多领域都有着极高的价值。例如当下的时事热点、某些产品的评价、大选的民意、实时灾情报告等等。要从文本中提取相应的信息,首先要进行分词、词性标注、句法分析等一系列自然语言处理操作。但是社交文本的不规范性严重影响了自然语言处理操作的有效性[1-3]。

在互联网时代之前,文本的生成和发表往往都经过严格的编辑和修订,这保证了文本在词法和句法上实现统一的规范性。而如今互联网用户产生的文本脱离了这种规范性,以英文为例,大量单词和词组以缩写、简写和谐音变形等形式出现在文本中,如 cu (see you)、2gether (together)、hw (homework/how)等;此外,口语式表达使文本中出现了大量主语省略、宾语省略等情况,还有诸如不规范的大小写、符号省略等情况。这些特征构成了社交文本中的噪声,对基于传统规范文本的自然语言处理工具产生了严重的干扰。解决上述问题的方法有两种:一种是恢复文本的规范化,将文本中不规范的词法和句法形式转换成规范的形式;另

本文于 2017 – 06 – 12 收到。

一种则是对现有的自然语言处理工具进行调整，令其适应社交文本中的噪声。本文主要介绍第一种方法，即文本规范化。

文本规范化最早由 Sprout 等人提出[4]，目的是为了解决文本到语音(text - to - speech，TTS)系统无法将文本中的非标准词转化成语音的问题。根据 Sprout 的定义，文本规范化是将非标准词转化成符合上下文的相应标准形式的过程。由于非标准词是社交文本不规范性的主要部分，目前针对社交文本的大多数研究延续了上述的定义，提出了多种针对非标准词的规范化方法。仅有少量研究针对不规范大小写、符号省略、主语或宾语省略等问题进行了初步探索。

然而正如 Eisenstein 所指出的[5]，Sprout 的定义并不能明确划分文本规范化的目标和范围，同时规范本身在语言学上就不是一个确定的概念，例如 flvr 应该被转化成 flavor 还是 flavour 是由相应的规范来决定的。为了明确文本规范化的定义，Han 定义了词汇规范化(lexical normalization)[6]，即将非标准词与其在词典中对应的标准形式关联起来的映射过程。同时附加了两个限制条件：①只有词典外(out - of - vocabulary，OOV)的单词被考虑视为非标准词加以规范化；②只考虑单个单词到单个单词的规范化。这个定义明确了规范就是所选择的词典。两个限制条件将非标准词的复杂性进一步简化，排除了出现频率较低的几类非标准词。例如，cu (see you)、l o v e (love)这样的一对多或多对一映射；wit (with)这样的非标准词本身是词典内(in - vocabulary，IV)单词等。

1　非标准词

在经过校对的传统文本中，非标准词主要是指首字母缩写和各种不同含义的数字(包括年份、金额、邮编等等)[4]。而在社交文本中，非标准词有如下类型：

(1)单词简写：将单词用词干或者少数字母表示，例如(seriously，srsly)；

(2)近音拼写：将单词的部分音节用发音相同的字母或数字表示，例如(you，u)、(forever，4ever)；

(3)首字母缩写：除了专有名词的首字母缩写外，还有很多词组的缩写，例如(to be honest，tbh)、(happy birthday，hb)；

(4)语气强调：刻意重复单词中的字母以表达强烈的情绪，例如(so good，soooo goood)；

(5)俚语、俗语：社交文本中混杂着来自不同地区、文化背景的各种俚语和俗语，例如推特上常见的牙买加英语、非裔英语方言等。这类单词有一部分存在对应的标准单词，另一部分则相反。因此是否将这类单词视为非标准词，目前仍有争议；

(6)拼写错误。

2　文本规范化方法

2.1　问题模型

文本规范化问题可以被形式化定义为给定一个原始单词序列 s，找到一个目标单词序列 t，使得条件概率 $P(t|s)$ 最大。我们可以用噪声信道模型(noisy channel model)来表示这个问题：

$$\arg\max_t P(t \mid s) = \arg\max_t \frac{P(s,t)}{P(s)} = \arg\max_t \frac{P(s \mid t)P(t)}{P(s)} \tag{1}$$

在公式(1)中，由于 $P(s)$ 是固定不变的，因此 $P(t|s)$ 与 $P(s|t)$ 和 $P(t)$ 的乘积成正比。先验概率 $P(t)$ 往往通过对目标领域的文本进行语言建模得到。$P(s|t)$ 则表示由标准文本序列 t 产生不规范文本序列 s 的过程。

以噪声信道模型为基础，目前衍生出了三种基于不同前提假设的规范化方法，分别是拼写修正(spell checking)方法、序列标注(sequential labeling)方法和机器翻译方法。

2.2　拼写修正方法

假设单词变成非标准词的过程是相互独立的，则文本规范化问题可以简化成单词拼写修正问题。传统

的单词拼写修正方法主要包括基于词形相似性[7,8]和基于分布相似性两类[9]。基于词形相似性的方法中最具有代表性的是通过计算单词的编辑距离(edit distance)来表示单词的相似性。单词的分布相似性则指的是不同单词出现在相似上下文中的概率[10]。最后再选择相似性最高的单词对文本中的非标准词进行替换。

而在社交文本中,非标准词的词形很可能与标准形式大相径庭,例如单词的刻意拉长(good, goooooood)、单词的简写(before, b4)等。文献[11]提出了社交文本中的几种非标准词的变化形式,如近音拼写(epic, epik)、单词裁剪(walking, walkin)等。文献[12]提出了4种针对推特文本的单词相似性模型。

由于社交文本的数量极大,在海量处理的场景下,实时的相似性计算往往满足不了效率上的要求。因此从拼写修正方法衍生出了基于词典的文本规范化方法[13]。即预先从语料中提取出常见的非标准词和标准词的词对,保存成词典的形式,然后再通过查找词典来对文本进行规范化处理。

2.3 序列标注方法

在考虑到文本中单词与单词的上下文关系时,文本规范化可以被看作一个序列标注问题进行求解。首先针对文本中的每个单词生成候选的若干个规范化单词,然后采用维特比算法基于语言模型进行求解,得到联合概率最大的单词序列作为规范化结果。通常所采用的序列模型有隐马尔科夫模型(Hidden Markov Model)[14]和条件随机场(Conditional Random Field)[15]。

序列标注模型是自然语言处理领域最常用的模型之一,但是在文本规范化的场景下略有不同。以词性标注为例,单词的词性标签是一个已知的有限集,而在文本规范化中,一个单词的候选规范单词集合往往是未知的。因此如何确定一个单词的候选规范单词集合是一个重要的问题。文献[16]提出了三种候选词生成方法的组合,分别是单词转换模型、视觉启动模型(Visual Priming)和拼写修正。单词转换模型是指对单词中的字母进行插入、删除、替换等操作并产生变形单词的过程。视觉启动模型是基于认知科学中启动的概念,在计算单词相似性的基础上采用了单词频率进行加权,然后基于计算结果得到单词的候选词集合。最后将三种方法生成的候选词以一定的优先级进行重排序,得到最终的候选词集合。

2.4 机器翻译方法

在前两类方法中,文本规范化都基于单词的一对一映射。这个前提在机器翻译方法中得到了扩展。文本规范化可以被看作是将不规范文本翻译成规范化文本的单语言机器翻译过程。借助于机器翻译方法中的词对齐(Word alignment)概念,可以对非标准词-标准词关系中的一对多、多对一和多对多映射进行建模。文献[17,18]各自提出了基于词组的机器翻译方法。文献[19]提出了基于字符的机器翻译方法。这种更细粒度的对齐方式可以更好地捕捉到单词变形的共性,例如训练数据中存在(fight,fite)这样的变形,经过训练后可以将"nite"还原成"night",即使训练数据中没有出现(night,nite)词对。

3 研究现状的分析与讨论

2011年多篇文献[1-3]报道了自然语言处理工具在处理社交文本时所出现的性能显著下降,从这时起,社交文本规范化作为解决这一问题的手段之一获得了广泛关注和研究。然而由于社交文本的复杂性和规范定义的不确定性,目前大部分社交文本规范化的研究都只针对非标准词。从广义上看,文本的规范化还应该包括分词、断句、句法结构等方面。在目前的研究中,文献[20]进行了社交文本中单词大小写修正的尝试;文献[21]进行了社交文本断句和助动词(is、are、am)补充的初步尝试。但是相对于非标准词规范化的研究,这些方向仍有待进一步深入。

在非标准词方面,文献[22]提出了非标准词识别问题,即如何在社交文本中区分一个单词是否为非标准词。首先,基于给定的词典我们可以将单词划分为词典内(in-vocabulary)单词和词典外(out-of-vocabulary)单词,然后单词被进一步划分为四类,即词典内标准词、词典内非标准词(例如wit,with)、词典外标准词(例如未收录在词典中的专有名词,Obama)、词典外非标准词。有效地从文本中识别出词典内非标准词和词典外非标准词是一个极具挑战的问题。此外,由于社交文本中用户来自世界各地,文本中往往混合了多

种语言，这使得非标准词的识别更加困难。对于拼写修正这一类上下文无关的方法而言，非标准词的识别能够有效提升规范化的准确率。

对于序列标注和机器翻译这两类考虑了上下文的方法，最大的问题在于训练数据的获取。由于大部分方法都采用有监督的训练，训练数据需要耗费大量人力进行手动标注。然而社交文本每天都会产生大量新的词汇和单词变形[5]，这对文本规范化方法的时效性提出了非常苛刻的要求。因此，非监督的方法将是社交文本规范化未来的研究方向。文献[23]提出了一种非监督的对数线性模型来对语料的整体进行学习，进而生成每个单词的候选词集合。

在语种方面，除了主流的英语社交文本的研究，还存在大量针对不同语言的社交文本规范化研究，包括了法语[24]、西班牙语[25]、德语[26]、日语[27]等等。在中文方面，文献[28]基于拼音对网络聊天文本中的谐音词进行了建模；文献[29]总结了中文社交文本中的三种不规范形式，分别是谐音（如“河蟹”，“和谐”）、简写（如“桌游”，“桌面游戏”）和释义（paraphrase，如“给力”，“很棒”），并针对这三种形式分别建模，最后采用基于规则和统计特征的综合方法进行规范化。

4　结束语

随着自然语言处理领域逐渐将研究兴趣从传统文本转向以微博为代表的社交媒体文本，社交文本规范化也受到了越来越多的关注。针对社交文本中的不规范特征进行研究，不仅可以提高自然语言处理工具的健壮性，还可以为社交媒体中语言演变的研究提供帮助。本文对社交文本规范化的基本方法和研究现状进行了介绍，同时也对未来的研究方向进行了分析。

参　考　文　献

[1] Gimpel K, Schneider N, O'Connor B, et al. Part - of - speech tagging for twitter: Annotation, features, and experiments[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2. Association for Computational Linguistics, 2011. 42 - 47

[2] Ritter A, Clark S, Etzioni O. Named entity recognition in tweets: an experimental study[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2011. 1524 - 1534

[3] Foster J, Cetinoglu O, Wagner J, et al. From news to comment: Resources and benchmarks for parsing the language of web 2.0[J]. 2011.

[4] Sproat R, Black A W, Chen S, et al. Normalization of non - standard words[J]. Computer speech & language, 2001, 15(3): 287 - 333

[5] Eisenstein J. What to do about bad language on the internet[C]//HLT - NAACL. 2013. 359 - 369

[6] Han B, Baldwin T. Lexical normalisation of short text messages: Makn sens a# twitter[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1. Association for Computational Linguistics, 2011. 368 - 378

[7] Brill E, Moore R C. An improved error model for noisy channel spelling correction[C]//Proceedings of the 38th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2000. 286 - 293

[8] Toutanova K, Moore R C. Pronunciation modeling for improved spelling correction[C]//Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2002. 144 - 151

[9] Li M, Zhang Y, Zhu M, et al. Exploring distributional similarity based models for query spelling correction[C]//Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2006. 1025 - 1032

[10] Lin D. Automatic retrieval and clustering of similar words[C]//Proceedings of the 17th international conference on Computational linguistics - Volume 2. Association for Computational Linguistics, 1998. 768 - 774

[11] Cook P, Stevenson S. An unsupervised model for text message normalization[C]//Proceedings of the workshop on computational approaches to linguistic creativity. Association for Computational Linguistics, 2009. 71 - 78

[12] Xue Z, Yin D, Davison B D. Normalizing microtext[C]//Workshops at the Twenty – Fifth AAAI Conference on Artificial Intelligence. 2011.

[13] Han B, Cook P, Baldwin T. Automatically constructing a normalisation dictionary for microblogs[C]//Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning. Association for Computational Linguistics, 2012. 421 – 432

[14] Choudhury M, Saraf R, Jain V, et al. Investigation and modeling of the structure of texting language[J]. International journal on document analysis and recognition, 2007, 10(3): 157 – 174

[15] Liu F, Weng F, Wang B, et al. Insertion, deletion, or substitution?: normalizing text messages without pre – categorization nor supervision[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers – Volume 2. Association for Computational Linguistics, 2011. 71 – 76

[16] Liu F, Weng F, Jiang X. A broad – coverage normalization system for social media language[C]//Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers – Volume 1. Association for Computational Linguistics, 2012. 1035 – 1044

[17] Aw A T, Zhang M, Xiao J, et al. A phrase – based statistical model for SMS text normalization[C]//Proceedings of the COLING/ACL on Main conference poster sessions. Association for Computational Linguistics, 2006. 33 – 40

[18] Kaufmann M, Kalita J. Syntactic normalization of twitter messages[C]//International conference on natural language processing, Kharagpur, India. 2010.

[19] Pennell D, Liu Y. A Character – Level Machine Translation Approach for Normalization of SMS Abbreviations[C]//IJCNLP. 2011. 974 – 982

[20] Nebhi K, Bontcheva K, Gorrell G. Restoring capitalization in# tweets[C]//Proceedings of the 24th International Conference on World Wide Web. ACM, 2015. 1111 – 1115

[21] Wang P, Ng H T. A Beam – Search Decoder for Normalization of Social Media Text with Application to Machine Translation [C]//HLT – NAACL. 2013. 471 – 481

[22] Han B. Improving the utility of social media with Natural Language Processing[D]. The University of Melbourne, 2014.

[23] Yang Y, Eisenstein J. A Log – Linear Model for Unsupervised Text Normalization[C]//EMNLP. 2013. 61 – 72

[24] Beaufort R, Roekhaut S, Cougnon L A, et al. A hybrid rule/model – based finite – state framework for normalizing SMS messages[C]//Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2010. 770 – 779

[25] Cerón – Guzmán J A, León – Guzmán E. Lexical normalization of Spanish tweets[C]//Proceedings of the 25th International Conference Companion on World Wide Web. International World Wide Web Conferences Steering Committee, 2016: 605 – 610

[26] Sidarenka U, Scheffler T, Stede M. Rule – based normalization of German Twitter messages[C]//Proc. of the GSCL Workshop Verarbeitung und Annotation von Sprachdaten aus Genres internetbasierter Kommunikation. 2013.

[27] Ikeda T, Shindo H, Matsumoto Y. Japanese Text Normalization with Encoder – Decoder Model[J]. WNUT 2016, 2016: 129.

[28] Xia Y, Wong K F, Li W. A phonetic – based approach to Chinese chat text normalization[C]//Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2006. 993 – 1000

[29] Wang A, Kan M Y, Andrade D, et al. Chinese Informal Word Normalization: an Experimental Study[C]//IJCNLP. 2013.

作者简介

罗程多,男,1987 年 8 月,博士,主要研究方向:自然语言处理,机器学习。