互联网背景下的计算语言学研究

王海峰, 赵世奇
(百度, 北京 100085)

摘要：近年来，计算语言学（Computational Linguistics: CL）在学术界和工业界均得到了越来越多的关注，这主要得益于其在互联网领域越来越广泛的应用，如搜索引擎、在线翻译系统、社交网络等。计算语言学的很多技术在互联网应用中都能找到用武之地。这其中既包括词法、句法、语义等基础技术，也包括问答、翻译、文摘等应用技术。面对海量、高噪声的互联网数据及真实互联网应用需求，计算语言学技术也需要进行调整与改进。本文将主要讨论在互联网大背景下的计算语言学研究，包括新应用、新资源、新挑战，以及新方法等。

关键词：计算语言学;互联网应用;互联网资源

Computational Linguistics in the Internet Age

Haifeng WANG and Shiqi ZHAO
(Baidu, Beijing 100085)

Abstract: Computational Linguistics (CL) has attracted more and more interest in both academic and industry communities in recent years, since it plays an essential role in many Internet applications, including search engines, online translation systems, social networks, and so forth. Almost all CL techniques, ranging from morphological, syntactic, and semantic analysis of texts, to question answering, machine translation, summarization, and other complex techniques, can find their scopes in the Internet applications. However, conventional methodologies in CL research need to be adapted and improved, so as to deal with new challenges arise from the new application requirements as well as the large-scale and highly noisy web corpora. In this paper, we discuss the CL research in the background of Internet, including the new applications, resources, challenges, and methodologies.

Keywords: computational linguistics; internet applications; internet resources

引言

计算语言学（Computational Linguistics: CL）是语言学与计算机科学的交叉学科，是指使用计算机建立形式化的数学模型，来分析和处理自然语言，因而又叫自然语言处理（Natural Language Processing: NLP）。计算语言学的研究始于上世纪 50 年代，滥觞于机器翻译研究。随着时代的发展，计算语言学研究也先后历经了几个重要阶段，包括基于简单字词匹配的阶段、基于语言学知识与规则的阶段，以及基于大规模数据统计的阶段。进入新世纪以来，随着互联网的迅猛发展以及在人们生活中的快速普及，尤其是搜索引擎技术与产品的异军突起，计算语言学研究在实际应用中的重要性日益提升。互联网应用带给计算语言学研究的最主要变化是要处理海量的互联网真实数据，这一变化使得传统的计算语言学研究必须调整研究问题和研究思路，充分利用互联网上产生的新资源和新数据，克服互联网应用中遇到的新困难和新挑战。在本文接下来的内容里，我们将分别就上述各个方面进行详细的阐述和分析。

1 互联网背景下计算语言学的新应用、新趋势

1.1 计算语言学在互联网背景下的主要应用

随着搜索引擎、电子商务、网络游戏、社交网络的蓬勃发展，人们将信息获取、经商购物、休闲娱乐、社会交往等各种日常行为都搬到了网络上。随着这些领域的逐步深入发展，越来越多的互联网应用借助计算语言学技术来优化性能。

作为互联网上最重要的应用之一，搜索引擎中大量使用了计算语言学技术。计算语言学技术在这一领域的应用可以概括为两点，即“面向需求搜索信息”以及“利用信息生成知识”。其中，前者旨在从海量网络信息中搜索出满足用户需求的特定信息；后者则致力于对搜索到的信息进行深入加工，以生成更精炼的知识，从而进一步方便用户的信息和知识获取。

现阶段，以百度等为代表的搜索引擎对计算语言学技术的应用大量体现在查询（query）解析方面。如查询改写需要基于同义资源挖掘技术，即自动挖掘出同义词或同义短语，以供查询改写时对查询片段进行同义替换；又如，查询省略需要基于查询词重要性分析技术，即计算出查询中各个词的重要程度的不同，以确定对长查询进行省略时删减查询词的顺序；再有，检索结果排序则会基于查询词紧密度分析技术，即判断查询中各个词的紧密程度，以确定检索结果中哪些词是必须邻接出现的。可见，搜索引擎对于计算语言学技术有着很多特有的需求，有些并不属于计算语言学研究的传统范畴。此外，搜索引擎在面向不同语言的时候所依赖的计算语言学技术也有所不同。例如，对于中文而言，搜索引擎还需要基于自动分词技术对查询和网页进行切分，而在分词系统中又要重点解决新词发现等问题。

随着“语义搜索”概念的兴起，计算语言学技术在搜索引擎中的应用显得越来越广泛和多样。所谓“语义搜索”，业界并没有一个明确的定义，因此不同的搜索引擎都有着自己的理解，在技术的使用上也有不同侧重。例如，百度提出“框计算”的理念，其核心功能之一便是可以更加准确的理解蕴含在用户查询背后的真实需求和意图，从而针对性地提供精确的搜索结果，而这背后则是以深入的查询理解技术作为支撑的。此外，WolframAlpha[1]采用自动问答技术，对于简单的问句式查询可以直接给出精确答案。SenseBot[2]基于自动文摘技术，为搜索引擎的每条结果生成文摘句，以便用户浏览。Cognition[3]运用词义消歧技术，对于有歧义的查询，将其搜索结果按照不同词义分开展现。人立方[4]基于信息抽取与关系抽取技术为每个人物类查询自动构建出其相关人物网络等。可以说，随着语义搜索领域的百花齐放，计算语言学的各项技术都获得了更大的尝试和应用空间。

除搜索引擎外，计算语言学技术在其他诸多互联网应用中也起着至关重要的作用。例如，以谷歌翻译[5]为代表的在线翻译系统使得统计机器翻译技术得以实际应用；信息定制系统的产生使得用户需求模型的自动建立与更新技术变得更加重要；伴随着电子商务系统而产生的大量在线评论信息对情感倾向性分析技术提出了更大的需求；社交网络的迅速普及使得“社会计算”进入研究者的视野，从而为实体关系挖掘、用户模型自动构建等研究方向提供了新的应用点。类似的需求在互联网应用的方方面面不断出现，预示着在互联网的大背景下计算语言学研究正进入一个活跃而繁盛的时期。

1.2 传统的计算语言学研究在互联网应用背景下产生的新变化

当经典的计算语言学研究遭遇当下的互联网应用，很多的问题定义、着重点，以及解决思路都发生了变化。本节将通过几个例子来分析一下传统的计算语言学研究在新的应用背景下究竟发生了哪些变化。我们要看的第一个例子是中文分词。分词可谓是中文计算语言学研发工作中最底层的技术模块之一，绝大多数上层应用都要基于分词。而互联网应用、尤其是搜索引擎的应用则对一个分词模块提出了以下三点需求：（1）快速高效，以能够及时处理数

以百亿计的网页；（2）及时更新，可以迅速挖掘出互联网上出现的新词、新概念、新专名等；（3）多种粒度，以满足不同应用中对于分词粒度的不同要求。鉴于以上三点需求，互联网应用中的分词模块在模型的选择上会更多的考虑运行速度和效率，而不是单一地追求准确率。同时，新词发现（即通常所说的未登录词识别问题）在一个分词模块中的重要性变得更大。为此，我们需要有一种方法，可以定期分析新近的网络语料并从中挖掘新词，再将挖掘到的新词补充到分词词表当中去。当然这也就意味着我们的分词模块要做到可以便捷地调整分词词表，并且将不同时期添加的新词进行合理的整合。而在分词粒度的把控方面，一个通用性强的分词模块会同时支持多种粒度，可供自如的切换和选择。

我们再来看一下机器翻译（Machine Translation: MT）的例子。机器翻译作为计算语言学的经典研究方向，已有半个多世纪的历史。从上世纪九十年代起，基于统计的方法逐渐在机器翻译研究中占据了统治地位。然而，由于大规模语料难以获得、运行速度相对较慢且模型规模过于庞大等原因，统计机器翻译技术一直难以应用到实际的翻译软件中去。而在线翻译系统的出现改变了这一局面。目前，谷歌、微软等公司都已推出了基于统计的在线翻译系统。而在线翻译系统的最大特色则是其超大规模的语料库，这其中既包括用以训练翻译模型的双语平行语料库，也包括用以训练语言模型的目标语端大规模单语语料库。以中文-英文这样的语言对为例，在互联网上可以比较容易的抓取到上亿对的双语平行语料。然而，在语料规模快速膨胀的同时也产生了相应的问题，即自动获取的语料中含有大量的噪声。因此，如何有效地过滤语料中的噪声、控制语料的质量成为对于在线翻译系统而言很关键的研究问题。除了语料，互联网上的在线翻译系统还需要考虑模型的选择、压缩、分布式存储及动态更新，翻译过程中的快速解码、领域自适应以及与搜索等不同应用的结合等。

本节的最后，我们再看一下自动问答（Question Answering: QA）技术的例子。自动问答和机器翻译一样，也是计算语言学的经典研究方向之一。初期的研究主要集中于如何深入的分析和理解输入问句，并且从一个给定的语料库（如 TREC 评测制定的语料库）中抽取问句的答案。之后，随着互联网的发展，研究者们开始更多的研究基于互联网的问答技术（Web based QA），即将整个互联网的资源看成是一个大的语料库，并从中抽取给定问题的答案。由于互联网信息的冗余性，基于互联网的自动问答系统可以在更大范围内抽取答案，并且更好地利用统计信息来对候选答案进行排序和选择。而近些年来，社区式问答（Community based QA）逐渐的成为了研究的热点。社区式问答最大的优势在于所有的问题都是用户提出的真实问题，所有的答案也都是用户人工编辑的答案，因此该资源的质量更高，准确性也更有保障。在社区式问答方面的研究目前主要集中在对社区式问答资源的检索以及问答资源质量自动评估等方面。

从以上的三个例子中我们可以看到，互联网应用一方面对传统的计算语言学研究提出了新的需求，使研究者必须按照实际需求重新定义研究问题、确定研究方法；另一方面，互联网资源也为研究者们带来了很多新的灵感和思路。

1.3 计算语言学研究在互联网应用背景下产生的新研究点

互联网应用不仅给传统的计算语言学研究带来了变化，还催生了很多个新的研究方向。例如，情感倾向性分析（Sentiment Analysis）便是在互联网上主观型评论信息越来越多的背景下应运而生的研究方向。尤其是在电子商务应用中，情感倾向性分析技术可以自动识别前人评论信息所蕴含的褒贬性，进而帮助人们判断一件商品的质量和口碑。事实上，在最近几年，几乎每一种互联网应用的产生都带动了一股研究热潮。比如，以Wikipedia[6]和百度百科[7]等为代表的在线百科全书的出现带动了基于相关资源的信息抽取和知识挖掘的研究，而博客、Twitter、微博等的流行也使得面向这些资源的检索、挖掘，以及用户兴趣建模、社交网络构建等方面的研究广受关注。可以说，互联网应用极大的促进了新的研究问题和研究方向的产生。

2 互联网背景下可供利用的新数据、新资源

互联网应用除了为计算语言学研究提供了更多的研究问题和内容之外，还同时提供了更丰富多样且规模巨大的数据和语料资源。这些数据资源对于很多研究方向的开展带来了更多可供选择的思路和方法。如何利用好互联网带来的新数据和新资源也成为了当下研究界很感兴趣的问题。以下我们将就互联网提供的几大类语料资源的特点和适用领域分别加以介绍。

2.1 海量网页语料库

互联网的规模有多大？恐怕很难有人能给出一个精确的数字，不过早在 2008 年，谷歌就曾宣称检测到了超过一万亿的各种语言独立页面。而百度抓取到的中文页面也达到了千亿量级。海量网页数据的利用可以大致分成两大类。一类是基于这样一个超大规模的语料库进行特定语言现象和分布的统计。例如，谷歌于 2006 年发布了 1T 五元组（5-gram）数据[8]，该数据即是从大规模网页语料中抽取出来，可以用于语言模型的训练。当然，还可以利用互联网语料库统计词语共现（co-occurrence）、词语搭配（collocation）等数据。互联网语料库在计算语言学研究中的另一大应用是信息抽取与数据挖掘。例如，Banko[9]等基于无指导的方法利用网页语料库进行关系抽取；Ravichandran和Hovy[10]使用一些人工制定的种子实例从互联网上学习信息抽取模板；Szpektor等[11]使用类似的方法从互联网语料中挖掘语义蕴含（entailment）模板；Bhagat和Ravichandran[12]基于“分布假设（Distributional Hypothesis）”从互联网语料库中抽取复述短语。另外，也有人尝试从大规模网页库中抽取双语句对资源[13]等。计算语言学研究中对于海量网页语料库的利用远不止以上所列内容。研究者所能应用的除了网页内容之外，还包括网页结构信息与网页间的链接关系。而海量网页资源的优点除了规模巨大之外，还体现在该资源在获取方面不受限制，任何研究者都可以从互联网上获取到相当规模的语料资源。海量网页语料的主要缺点是其中含有大量噪声，可能会对某些应用产生负面影响。

2.2 搜索行为数据

搜索行为数据，又叫用户日志（query log），是用户在使用搜索引擎时生成的数据。搜索行为数据包括用户查询（query）、用户点击（click），以及查询序列（session）等信息。对于像谷歌、百度这样规模的搜索引擎，每天记录下来的用户行为数据数以亿计，其中蕴含着大量可供研究与利用的信息和知识。例如，从用户查询集合中可以学习到查询语言的分布和特点，还可以挖掘出网络用户的兴趣分布与网络热点；从用户点击数据中可以学习到用户查询与网页资源的匹配情况，从而改善信息检索模型，提升搜索引擎的性能；从查询序列数据中则可以学习到用户对查询的改写与重构，以及用户搜索兴趣的自然转移情况等。信息检索领域对于用户搜索行为数据的研究非常广泛，主要包括查询分类[14]、查询推荐[15]、查询扩展[16]等有关用户查询的分析与处理，以及利用搜索行为数据改善搜索结果排序的相关性[17]等内容。

与此同时，搜索行为数据在计算语言学的研究中也有着非常重要的用途。例如，有研究者将搜索行为数据用于命名实体（Named Entity: NE）识别与新词发现[18,19]。这其中的基本想法是，若一个未知词或实体w_1与一个已知类型为T的词或实体w_2常出现在相似的查询中，则w_1的类型也应为T。与之类似，有人基于搜索行为数据进行语义关系抽取以及本体（Ontology）构建的工作[20,21]。还有人利用用户查询与对应的点击文档的标题抽取复述资源[22]。除规模巨大之外，用户搜索行为数据的另一大优势是时效性强，可以及时的反映用户的新需求、新兴趣等。然而，用户日志中也含有大量噪声，尤其是用户查询语序和语法随意，

内容省略，且含有大量的拼写错误等，这些都给基于搜索行为数据的应用设置了障碍。另外，出于隐私保护等方面的考虑，搜索行为数据在获取上十分受限，往往只有搜索引擎公司可以获取和使用，这也使得基于该资源的研究受到了一定的局限。

2.3 用户生成内容数据

用户生成内容，即 User Generated Content（UGC），是近些年来计算语言学研究的新热点。UGC 数据又可具体细分为三类，即社区/论坛数据，知识分享网络，以及博客/微博数据。

社区/论坛数据：社区/论坛数据是几类UGC数据中出现最早的一类。尽管此类数据中噪声问题十分严重，但还是有研究者利用该数据资源从事计算语言学方面的研究工作。其中最典型的例子是从社区/论坛的发帖、回帖数据中抽取问答知识，以提供给自动问答系统使用[23]。

知识分享网络：知识分享网络有两种常见类型，一种是wikipedia、百度百科等在线百科类资源；另一种则是百度知道[24]等社区式问答资源。近几年，计算语言学界基于上述两种资源的研究很多。其中，基于在线百科资源的研究工作又可以细分为四类：（1）利用百科类资源内容准确、噪声较少的特点，进行关系抽取等研究[25,26]；（2）利用百科类页面的特定格式、句式模板、半结构化数据等进行信息抽取[27]、自动文摘[28]等研究；（3）利用词条内链关系等信息进行词义消歧[29]、词关系挖掘[30]、跨语言NE识别[31]等研究；以及（4）利用用户的编辑历史信息，进行文本简化[32]、句子压缩[33]等研究。另一方面，有关社区式问答的研究主要可分为两类，一类是研究社区式问答资源的检索与推荐[34,35]，另一类则是研究社区式问答资源的数据质量自动评估[36,37]。

博客/微博数据：现如今，博客和微博越来越成为普通网民展示和表达自我以及相互沟通交流的平台。以中国的微博为例，最大的两个微博平台新浪微博和腾讯微博的注册用户数量在 2011 年上半年均已超过一亿。博客和微博数据的最大特点在于其内容的个性化、主观性，以及时效性。这些特点使其在计算语言学的很多方向上被加以研究和应用，例如用户个性化兴趣模型的构建[38]、博客和微博内容的个性化推荐[39,40]、主观性内容的情感倾向性分析[41,42]、热点事件及舆情的检测与跟踪[43,44]等等。

从上述内容中我们可以发现，互联网可以提供的数据资源可以惠及计算语言学研究的大多数方向。因此所有研究人员在改善算法和模型的同时都要认真考虑如何从新数据与新资源中受益。

3 互联网应用背景下出现的新问题、新挑战

正所谓“世上没有免费的午餐”。互联网在提供给我们丰富多样的资源和数据的同时，也同时提出了更多的问题和挑战，具体体现在以下几个方面：

对存储/运算能力提出更高要求：海量网页、用户日志等数据需要超大的存储空间，同时处理和统计这些数据需要很强的运算能力，尤其是并行计算的能力。此外，互联网数据“时效性”的特点也意味着很多运算结果需要频繁更新（比如热点新闻的挖掘与跟踪），这也要求必须能在很短的时间周期内完成对大规模数据的运算。

需要应对快速涌现的新的语言现象：这里，“新的语言现象”主要指的是新词（如“给力”、“雷人”）、新概念（如“80 后”、“啃老族”）、新专名（如“筷子兄弟”、“旭日阳刚”），新用法（如“粉丝”、“围脖”），以及大量的网络语言，甚至“火星文”等。UGC 数据的膨胀催化了新的语言现象的出现，但同时也给自然语言处理技术设置了更大的障碍。只有准确地对新词进行切分、对新概念/新专名进行挖掘、对新用法进行统计、对网络语言进行改写和规范化，才能够满足信息抽取、机器翻译、自动问答等诸多上层应用需求。

数据噪声的过滤与纠错：传统的媒体信息通常是由专业人员编辑生成的，但互联网信息中含有大量的UGC数据，这些数据中含有很多错误内容，最主要的体现为错别字、表达随意、句法不规则等。这些错误对计算语言学的一系列底层技术，包括分词、词性标注、句法分析等提出了难题。因此一方面需要考虑如何进行纠错，另一方面也要考虑如何对噪声数据进行过滤。其中，前人对于拼写改错或信息检索中的查询改错的研究较多[45]，但对于数据噪声过滤尚且缺乏系统的、有通用性的研究成果。

数据内容可信性的甄别与分析：互联网数据的另一大特点是信息的质量良莠不齐。具体地，在网页数据中含有为数不少的不实新闻、虚假广告、“软文”等内容；在用户日志中含有大量的用户对不确定信息的搜索；在UGC数据中更是含有非常多的主观性内容。因此在利用互联网数据的时候应首先考虑数据内容的真实性、可信度、主观性等方面，否则便容易受到错误或不实信息的误导。对此，已经有研究者从事“可信计算”方面的研究，即利用信息的来源、发布者、内容等多方面特征来判别和度量信息可信度[46,47]，但这方面的研究工作还很不足。

4 互联网背景下解决问题的新思路、新方法

为应对互联网应用提出的新问题和新挑战，计算语言学界的研究者们应当与时俱进地转变研究思路，提出因应新趋势变化的新的研究方法。笔者认为，互联网应用背景下的计算语言学研究应注重以下几方面：

贴近真实需求：认真分析实际应用对每一项研究的真实需求，从需求出发设计研究题目和内容。根据互联网背景下需求的变化来调整传统研究问题的定义和解决方法。在设计研究方法时要充分考虑方法在实际应用中的可行性与易行性，尽可能兼顾方法的性能与效率。与此同时，要善于利用真实的数据、资源、工具来解决问题，尤其是要善于对各种可利用资源进行整合与互补。

聚焦前沿研究：对互联网应用的新趋势、新发展保持敏感，善于捕捉新需求、抽象新问题、发现新现象、总结新规律。既要尝试将经典方法运用于新问题，又要针对新问题思考全新的解决方法。比如，情感倾向性分析、面向微博和博客的研究，以及“众包（crowdsourcing）[48]”等便是随着互联网应用的发展而被抽象出来的新研究问题。另一方面，又要杜绝盲目跟风、人云亦云、浅尝辄止。不能一窝蜂的拥到“流行”问题或者“时髦”方法上，而是要将对新问题的研究建立在理性分析的基础上。

平衡数据算法：在传统的计算语言学研究中，通常是“小数据、大算法”的特点，即训练和测试的数据规模有限，研究者致力于不断的改善算法性能，到一定程度后会出现过拟合、算法过于复杂、脱离实际条件等问题。而在互联网背景下的计算语言学研究中，通常是“大数据、小算法”的特点，即数据规模足够大，从而使得研究者必须考虑方法的可行性和实现效率，提出更简单快捷的方法。研究者应对数据和算法的规模做一个合理的权衡与折中，而不是一味地追求算法的高深和复杂。

搭建面向应用的实验平台：研究问题和内容的转变还同时要伴随着实验方法和条件的转变。面向互联网应用的众多不同需求，若想让科研成果真正付诸实用，还需要搭建面向应用的实验平台，这其中既包括真实的实验设置、实验数据等，也包括更实用和更全面的评测指标。在这方面，科研院所与互联网企业的研究者应当通力合作，取长补短。一方面，互联网企业要积极向学术界反馈用户实际应用需求，同时在力所能及的范围内公开资源和数据；另一方面，学术界要将实际应用需求抽象为研究问题，并且从企业公开的真实数据中分析出更深层的现象，进一步提升至理论高度。

5 结语

时至今日，互联网技术和产业仍在蓬勃发展，互联网应用背景下的计算语言学研究更是方兴未艾。本文总结了计算语言学研究在互联网背景下的新应用，梳理了互联网能够带给研究界的新资源，分析了互联网应用的新形势下研究者们面临的新挑战，并且提出了面向互联网应用做研究的新方法。由于互联网包罗万象、瞬息万变，我们的总结和分析难免管中窥豹，有失全面。我们坚信，计算语言学界的广大研究者在与互联网应用打交道的过程中一定都积累了很多宝贵的心得体会，这必将对我们整个研究领域产生深远影响，也必将使得我们的研究一方面紧跟学术前沿，一方面面向应用需求，正所谓“仰望星空、脚踏实地！”

参考文献

[1] WolframAlpha：http://www.wolframalpha.com/

[2] SenseBot：http://www.sensebot.net/

[3] Cognition：http://www.cognition.com/

[4] 人立方：http://renlifang.msra.cn/

[5] 谷歌翻译：http://translate.google.com/

[6] Wikipedia：http://www.wikipedia.org/

[7] 百度百科：http://baike.baidu.com/

[8] Google 1T 5-gram语料：http://www.ldc.upenn.edu/Catalog/docs/LDC2006T13/readme.txt

[9] Michele Banko, Michael J Cafarella, Stephen Soderland, Matt Broadhead and Oren Etzioni. 2007. Open Information Extraction from the Web. In Proceedings of IJCAI, pages, 2670-2676.

[10] Deepak Ravichandran and Eduard Hovy. 2002. Learning Surface Text Patterns for a Question Answering System. In Proceedings of ACL, pages, 41-47.

[11] Idan Szpektor, Hristo Tanev, Ido Dagan, and Bonaventura Coppola. 2004. Scaling Web-based Acquisition of Entailment Relations. In Proceedings of EMNLP, pages, 41-48.

[12] Rahul Bhagat and Deepak Ravichandran. 2008. Large Scale Acquisition of Paraphrases for Learning Surface Patterns. In Proceedings of ACL, pages, 674–682.

[13] Long Jiang, Shiquan Yang, Ming Zhou, Xiaohua Liu, and Qingsheng Zhu. 2009. Mining Bilingual Data from the Web with Adaptively Learnt Patterns. In Proceedings of ACL, pages, 870-878.

[14] Xiao Li, Ye-Yi Wang, Alex Acero. Learning Query Intent from Regularized Click Graphs. In Proceedings of SIGIR, pages, 339-346.

[15] Huang C, et al. Relevant Term Suggestion in Interactive Web Search based on Contextual Information in Query Session Logs. Journal of the American Society for Information Science and Technology, 2003. 54(7):638-649.

[16] Hang Cui, Ji-Rong Wen, Jian-Yun Nie, Wei-Ying Ma. 2002. Probabilistic Query Expansion using Query Logs. In Proceedings of WWW, pages, 325-332.

[17] Thorsten Joachims. 2002. Optimizing Search Engines using Clickthrough Data. In Proceedings of KDD, pages, 133−142.

[18] Jiafeng Guo, Gu Xu, Xueqi Cheng, Hang Li. 2009. Named Entity Recognition in Query. In Proceedings of SIGIR, pages, 267-274.

[19] Junwu Du, Zhimin Zhang, Jun Yan, Yan Cui, Zheng Chen. 2010. Using Search Session Context for Named Entity Recognition in Query. In Proceedings of SIGIR, pages, 765-766.
[20] Satoshi Sekine and Hisami Suzuki. 2007. Acquiring Ontological Knowledge from Query Logs. In Proceedings of WWW, pages, 1223-1224.
[21] Marius Pasca and Benjamin Van Durme. 2008. Weakly-supervised Acquisition of Open-domain Classes and Class Attributes from Web Documents and Query Logs. In Proceedings of ACL, pages, 19-27.
[22] Shiqi Zhao, Haifeng Wang, and Ting Liu. 2010. Paraphrasing with Search Engine Query Logs. In Proceedings of COLING, pages, 1317-1325.
[23] Jizhou Huang, Ming Zhou, Dan Yang. 2007. Extracting Chatbot Knowledge from Online Discussion Forums. In Proceedings of IJCAI, pages, 423-428.
[24] 百度知道：http://zhidao.baidu.com
[25] Yulan Yan, Naoaki Okazaki, Yutaka Matsuo, Zhenglu Yang, and Mitsuru Ishizuka. 2009. Unsupervised Relation Extraction by Mining Wikipedia Texts Using Information from the Web. In Proceedings of ACL, pages, 1021–1029.
[26] Dat P.T. Nguyen, Yutaka Matsuo, and Mitsuru Ishizuka. 2007. Subtree Mining for Relation Extraction from Wikipedia. In Proceedings of NAACL, pages, 125-128.
[27] Fei Wu and Daniel S. Weld. 2010. Open Information Extraction Using Wikipedia. In Proceedings of ACL, pages, 118-127.
[28] Shiren Ye, Tat-Seng Chua, and Jie Lu. 2009. Summarizing Definition from Wikipedia. In Proceedings of ACL, pages, 199-207.
[29] Rada Mihalcea. 2007. Using Wikipedia for Automatic Word Sense Disambiguation. In Proceedings of NAACL, pages, 196–203.
[30] Eyal Shnarch, Libby Barak, and Ido Dagan. 2009. Extracting Lexical Reference Rules from Wikipedia. In Proceedings of ACL, pages, 450-458.
[31] Alexander E. Richman and Patrick Schone. 2008. Mining Wiki Resources for Multilingual Named Entity Recognition. In Proceedings of ACL, pages, 1-9.
[32] Mark Yatskar, Bo Pang, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2010. For the Sake of Simplicity: Unsupervised Extraction of Lexical Simplifications from Wikipedia. In Proceedings of NAACL, pages 365–368.
[33] Elif Yamangil and Rani Nelken. 2008. Mining Wikipedia Revision Histories for Improving Sentence Compression. In Proceedings of ACL, pages, 137-140.
[34] Xiaobing Xue, Jiwoon Jeon, W. Bruce Croft. 2008. Retrieval Models for Question and Answer Archives. In Proceedings of SIGIR, pages 475-482.
[35] Yunbo Cao, Huizhong Duan, Chin-Yew Lin, Yong Yu, and Hsiao-Wuen Hon. 2008. Recommending Questions Using the MDL-based Tree Cut Model. In Proceedings of WWW, pages 81-90.
[36] Pawel Jurczyk and Eugene Agichtein. 2007. Hits on Question Answer Portals: Exploration of Link Analysis for Author Ranking. In Proceedings of SIGIR, pages 845-846.
[37] Young-In Song, Chin-Yew Lin, Yunbo Cao and Hae-Chang Rim. 2008. Question Utility: A Novel Static Ranking of Question Search. In Proceedings of AAAI, pages 1231-1236.
[38] Dongwoo Kim, Yohan Jo, Il-Chul Moon, and Alice Oh. 2010. Analysis of Twitter Lists as a Potential Source for Discovering Latent Characteristics of Users. Workshop on Microblogging at CHI.

[39] Jia Wang, Qing Li, Yuanzhu Peter Chen, and Zhangxi Lin. 2010. Recommendation in Internet Forums and Blogs. In Proceedings of ACL, pages, 257–265.

[40] Jilin Chen, Rowan Nairn, Les Nelson, Michael Bernstein, and Ed H. Chi. 2010. Short and Tweet: Experiments on Recommending Content from Information Streams. In Proceedings of CHI.

[41] Feifan Liu, Dong Wang, Bin Li, Yang Liu. 2010. Improving Blog Polarity Classification via Topic Analysis and Adaptive Methods. In Proceedings of NAACL, pages, 309-312.

[42] Bernard J. Jansen, Mimi Zhang, Kate Sobel, and Abdur Chowdury. 2009. Twitter Power: Tweets as Electronic Word of Mouth. Journal of ASIS&T, 60(9), 1-20.

[43] Thomas Heverin and Lisl Zach. 2010. Microblogging for Crisis Communication: Examination of Twitter Use in Response to a 2009 Violent Crisis in Seattle-Tacoma, Washington Area. In Proceedings of the 2010 ISCRAM Conference.

[44] Amanda Lee Hughes and Leysia Palen. 2009. Twitter Adoption and Use in Mass Convergence and Emergency Events. In Proceedings of the 2009 ISCRAM Conference.

[45] Xu Sun, Jianfeng Gao, Daniel Micol, Chris Quirk. 2010. Learning Phrase-Based Spelling Error Models from Clickthrough Data. In Proceedings of ACL, pages, 266-274.

[46] Miriam J. Metzger. 2007. Making Sense of Credibility on the Web: Models for Evaluating Online Information and Recommendations for Future Research, Journal of the American Society of Information Science and Technology, 58(13), 2078-2091.

[47] Ramanathan Guha, Ravi Kumar, Prabhakar Raghavan, and Andrew Tomkins. 2004. Propagation of Trust and Distrust. In Proceedings of WWW, pages, 403-412.

[48] Jeff Howe. 2006. The Rise of Crowdsourcing. Wired Magazine, 14.06.