汉语否定极项（NPI）自动抽取研究

王 栋　盛玉麒

山东大学 文学与新闻传播学院 济南 250100

E-mail: flasher@mail.sdu.edu.cn　yuqi-sheng@163.com

摘要：否定极项（NPI）指那些只能出现在否定句中的词语，如“丝毫”、“万万”、“绝”等。本文运用形式语义学的理论与语料库语言学方法，研究现代汉语文本中否定极项（NPI）的自动抽取。基本思路是：通过词表中词在语料库里否定句中的出现频度来判断是否是一个否定极项，通过对实验结果的分析，得出了基本结论和进一步研究的思路。

关键词：形式语义学、语料库、 否定极项、自动抽取

A Research on Extracting Chinese NPIs From Corpus

Wang Dong,　Sheng Yuqi

School of Literature and Journalism, Shandong University, Jinan, 250100

E-mail: flasher@mail.sdu.edu.cn　yuqi-sheng@163.com

Abstract: In this paper, we focus on giving a new method for a very significant problem in modern theoretical linguistics- automatically extracting NPIs from corpus. The approach is established on the combination of formal semantics and corpus linguistics. A solution is introduced by us: we can determine whether a word is an NPI or not based on its occurrence frequency in negative context. And then we can extract it if it is an NPI. After analyzing the data and result of this method, we give some suggestions on improving further research.

Key Words: formal semantics,　corpus linguistics,　NPI,　automatically extracting NPIs

1　引论

“否定极项”源自英文 Negative Polarity Items（以下简称 NPI），指那些仅能出现在否定句中的词、短语、或者固定结构。最早见于 Klima 关于英语中否定的研究（Klima 1964）。英语下列例子中的 any、at all、lift a finger、ever、give a red cent 等都是 NPI，比如：

1)　a. John didn't *say* anything.　（Klima 1964）
　　b. Nobody *lifted a finger* to me.　（Zwart 1998）

汉语中有一些词，比如：丝毫、万万、从来、压根、景气等，同样有类似的性质，比如：

2)　a. 你*万万*不可以轻敌。
　　b. 他*丝毫*不为眼前的情景所动。
　　c. 经济危机下，各个企业都不大*景气*。

由于其分布及语义上的特殊，早在 20 世纪 70 年代就引起了很多学者的关注，在句法、语义、语用等多各方面都有大量的相关研究（Klima 1964、Ladusaw 1970、Horn 1972、Krifka 1995、

Giannakidou 1998、Israel 2004 等），至今仍然是国际语言学界的一个研究热点。

以往关于否定极项的允准条件及语义，多数都是基于内省的方法，有许多创新的见解，例如，Horn 1972、Hirschberg、Julia Bell 1989、石毓智 1992、沈家煊 1999，以及 Israel 2004，Krifka 1994，Huang 1982，Li 1992，Lin 1996，伍雅清 2000 等。但是这些研究都是基于很少的几个 NPI，少有在穷尽式提取某种语言的 NPI 的基础上进行的，因此或多或少会受到语料缺失的影响。受形式语义学在国内发展现状的制约，关于汉语 NPI 的研究很少有人问津，相关研究成果很少，而自动抽取汉语 NPI 的研究更是一项空白。没有详实的语料做支撑，单纯依靠个别 NPI，就难以从宏观对 NPI 进行整体的相关研究。

本研究的基本思路是：穷尽式搜索语料库中的否定句，然后找出其中的 NPI。其中涉及到两个环节：一个是对否定句的判断；一个是对 NPI 的判断。相比之下，后一个判断难度更大，也是本研究的关键。

指导思想是：根据否定句中初选词的频度与该词在整个语料库中的频度之间的相关性，判断该词在否定句里出现的概率。按照概率大小降频排列初选词表，对该表中高频区的初选词进行人工判断，从中选出汉语的 NPI。

具体操作步骤是：

1）确定一个词表，作为所有可能成为 NPI 的词语集合；

2）确定一个语料库，作为否定句抽样的母本；

3）从词表中取出一个词 w，求出 w 在语料中出现的次数 N_w，再求出 w 在语料中的否定句中的出现次数 N_{neg}，那么用 N_w 除以 N_{neg}，即 N_w/N_{neg}，就得到了词 w 在否定句中出现的大致概率。

4）将每个词按照否定句中出现概率降序排列，作为 NPI 备选词表；

5）对 NPI 备选词表进行人工判断，基本上就可以得出汉语的 NPI。

其中较为重要的环节有：语料库的选取、词表的选取和否定句的判断。

2 语料库的建立与词表的选取

2.1 语料库的选取

1．语料库选材与加工

理想的语料库是经过加工标注词性的熟语料库，因为具有否定成分和句法属性等标注，可以减少程序判断的复杂性，提供效率和信度。但是由于目前的自动分词和词性标注结果往往需要人工校对。工作量巨大，不是个人力量短期能完成的，多人集体校对也会因为标准掌握的差异影响质量。因此，本研究使用的是未进行分词和标注的生语料库。

2．语料类型

在语料选取类型方面，我们选取的语料可以大致分为书面语语料和口语语料。口语语料库以相声小品脚本为主，共约 300 余个文件，500 余万字；书面语语料库以《人民日报》语料和网络小说语料为主，共约 1000 个文件，4000 余万字，所有语料均未经过自动分词标注。

3．语料库文件格式

为了便于程序运行，所有文本一律采用“.txt”纯文本格式，GB2312 内码。

2.2 词表的选取

词表选择原则

1. 通用性

确保词表中的词语是现代汉语中普遍使用的，不收罕用词、专业术语等词语。

2. 规范性

词表中的词语应该符合现代汉语词汇规范，不收古语词、方言词语、行业语词以及歇后语、俗语、习语等。

3. 均衡性

主要指词表中的词语在词类分布上应该具有一定的均衡性。

4. 数量适中

词表是除语料库外的另一个重要资源，其选取直接影响到统计结果的精确度。如果词表过小，就不能覆盖所有的词，会漏掉统计某些词；如果词表过大，会影响程序的效率。

基于各种因素的考虑，我们从现代汉语八百词、HSK 甲乙丙级词等词表中选取了约 25000 词，组成该研究所用词表。词表以二字词为主，另外还包括单字词、多字词甚至词组等。

3 否定句的抽取

3.1 否定句的判断

否定句指含有否定副词的句子。现代汉语的否定副词有“不”、“没”、“非”、“否”、“莫”、“别”等。为了验证本研究思路的可行性，只选择“不”和“没”两个典型的否定副词作为判断项。如果需要，今后可以将所有否定副词都加入进来。

一个比较特殊的情况是“双重否定”句的处理。对此，经过抽样分析发现，即使在双重否定句中，也可以根据单个否定副词进行 NPI 的筛查依据。例如：

他[不是[*丝毫*没有生气]]。

其中的“丝毫”存在于最靠近否定副词的结构中，可见，NPI 实际上是与否定副词直接相关的成分，双重或多重否定并不会影响对它的判断。实际上，关于 NPI 与否定范围的关系，也是国外理论语言学界集中讨论的问题之一，基于最小否定范围内的 NPI 判断参见 Progovac（1994）等。

3.2 否定句的过滤

1．过滤含有“不”和“没”的词语

程序一旦发现句子中含有“不”、“没”，就会认定该句是否定句。而含有“不”和“没”的词并不是否定极项，如：“不仅”、“没头没脑”、“不慌不忙”等。

根据最小范围判断的方法，进行匹配，共有 600 余个含有“不”、“没”的词语，所以这些词在否定句或者否定范围内的出现的概率虽然是 100%，但它们并不是我们需要找的 NPI。

2 过滤低频词

在概率判断中，低频词没有信度价值，无法提供足够的数据作为判断筛选的依据。为此确定以 20 次为限，把出现次数不足 20 次的词都过滤掉，以确保统计分析的信度的可靠性。

3．过滤非词语成分

还有一些不是词的汉字组合也会出现在其中，比如“显山”、“举妄动”、“红皂白”等等，这是由文本或词表的错误造成的，同样要去掉。

4 结果分析

4.1 抽取出的NPI样表

自动抽取的备选NPI共有100个左右，经过人工干预审定后按降频排列，本文列出了具有代表性的前50个词语（详见附录）。与其他语言相比，我们得出的汉语NPI数量居于中游水平，比如Hoeksema （1997）收集了760个荷兰语中的NPI，Beata & Jan-Philipp（2008）收集了84个德语NPI、58个罗马尼亚语NPI，Von Bergen (1993) 收集了141个英语NPI。

4.2 抽得的NPI样表分析

1．表示极小量的NPI

“丝毫、分毫、吹灰之力、吭声、理睬、压根儿、丝毫、动弹、一丝一毫、半点、介意、声张、吱声、分毫”等词语，虽然词性不同，有副词也有动词，但在语义上确实基本相同的，都表示一种极小的量，或者一种极端情况。在否定模式中，对这类NPI的否定，能表达全面或充分否定。这是利用否定的不对称性，通过对极小量的否定达到完全否定或者接近完全否定，涉及部分语用因素（石毓智 1992， 沈家煊 1999）。这类NPI占得比例很大，这与其他语言中NPI的情况很类似，比如英语中也存在大量的此类NPI：drink a drop（喝一杯）、hurt a fly（伤害任何东西）、lift a finger（举手之劳）等，这类NPI在英语中占37.1%（Von Bergen 1993）。

2．表示极大量的NPI

“再也、久久、万万、从来、再也、无时无刻、全然、源源、久留”等，在语义上恰恰与上一类相反，表示某种极大的量。与否定副词搭配使用时，表示的是一种程度很高或者全称量化的否定，如：

“久久不能忘怀”、“万万不可粗心大意”、“我从来不玩网络游戏”等等。

该类NPI在英语、德语、罗马尼亚语（Timm 2005）中也存在，但是所占比例非常小，比如英语中此类NPI只有5%（Von Bergen 1993），而汉语中该类NPI所占的比例要比上述语言中大得多，相比较下，这一特征是汉语独有的。

5 余论

本研究提出基于语料库语言学的方法自动抽取现代汉语否定极项（NPI）的解决方案，发现了汉语中表示极小量和极大量的NPI词语，以及汉语特有的表示较大程度的NPI等，与目前已有的对NPI的分布与语义研究基本吻合。

本研究成果和解决方案的思路对于汉语句法语义理论研究、自动翻译、机器学习、对外汉语教学、语言类型学研究等多个领域的基础及应用研究，都具有很高的应用价值。

我们采用的统计方法存在统计精度不高的问题，某些本不该是NPI的词反而排在了结果列表的前列，属于统计模型、选取语料等导致的噪音问题。今后应选取经过分词标注词性的“熟”语料库，特别要标注肯定否定、句法依存关系等属性，这不但有助于NPI的自动抽取，而且将有助于相关问题的深入研究。

欣闻全国第十届计算语言学学术会议胜利召开，谨以此文表示祝贺，并就教于各位专家。恳请批评指正。

作者谨识 2009-5-10

参 考 文 献

[1]. 沈家煊，1999，《不对称和标记论》[M]，南昌：江西教育出版社。

[2]. 石毓智，1992，《肯定和否定的对称与不对称》，学生书局，台湾。

[3]. 伍雅清，2000，《单位词是极端 WH 词项的允准语》，《现代外语》 第 4 期。

[4]. Beata Trawinski and Jan-Philipp Soehn: A Multilingual Database of Polarity Items. Poster presentation at LREC 2008 in Marrakech, Morocco.

[5]. Giannakidou, A. (1998). Polarity Sensitivity as Nonveridical Dependency. John Benjamins, Amsterdam.

[6]. Hoeksema, J. (1997). Corpus study of negative polarity items. Html version of a paper which appeared in the IV-V Jornades de corpus linguistics 1996-1997, Universitat Pompeu Fabre, Barcelona.

[7]. Horn, L.: 1972, On the Semantic Properties of Logical Operators in English, PhD thesis, UCLA.

[8]. Huang, James C.-T.1982. Logical Relations in Chinese and the Theory of Grammar. Ph. D dissertation, MIT.

[9]. Israel 2004. the Pragmatics of polarity. The Handbook of Pragmatics. Horn, L. and G. Ward, Blancwell. pp. 701-723

[10]. Klima, E. (1964). Negation in English. In J. A. Fodor and J. Katz (Eds.), The Structure of Language, pp. 246–323. Prentice Hall, Englewood Cliffs, New Jersey.

[11]. Krenn, B. (1999). The Usual Suspects. Data-Oriented Models for Identification and Representation of Lexical Collocations, Volume 7 of Saarbr¨ucken Dissertations in Computational Linguistics and Language Technology. Saarbr¨ucken: DFKI and Universit¨at des Saarlandes.

[12]. Krifka, Manfred (1995), The semantics and pragmatics of polarity items in assertion. Linguistic Analysis 15：209-257.

[13]. K¨urschner, W. (1983). Studien zur Negation im Deutschen. Gunter Narr, T¨ubingen.

[14]. Ladusaw, W. (1980). Polarity Sensitivity as Inherent Scope relations. Garland Press, New York.

[15]. Ladusaw, W. (1996). Negation and polarity items. In S. Lappin (Ed.), The Handbook of Contemporary Semantic Theory, pp. 321–341. Blackwell Publishers.

[16]. Lemnitzer, L. (1997). Akquisition komplexer Lexeme aus Textkorpora. T¨ubingen: Niemeyer.

[17]. Lin, Jo-wang. 1996. Polarity Licensing and Wh-phrase Quantification in Chinese, Ph.D dissertation, University of Massachuesetts, Amherst.

[18]. Linebarger, M. C. (1980). The Grammar of Negative Polarity. Ph. D. thesis, MIT. cited after the reproduction by the Indiana University Linguistics Club, Indiana, 1981.

[19]. Progovac, Ljiljana. 1994. Negative and Positive Polarity. Cambridge： Cambridge University Press.

[20]. Rayson, P. and R. Garside (2000). Comparing corpora using frequency profiling. In Proceedings of the Workshop on Comparing Corpora, ACL, 1–8 October 2000, Hong Kong, pp. 1–6.

[21]. Timm Lichte. (2005). Corpus-based Acquisition of Complex Negative Polarity Items Proceedings of the Tenth ESSLLI Student Session Judit Gervain (editor) Chapter 14.

[22]. Von Bergen, A. and K. Von Bergen (1993) *Negative Polarität im Englischen.* Gunter Narr Verlag, Tübingen.

附 录

词表	总出现次数	否定中出现的次数	在否定中的出现频率	词表	出现总次数	否定中出现的次数	在否定中的出现频率
对劲	559	544	0.973166	动弹	899	697	0.775306
吹灰之力	169	164	0.970414	好气	887	686	0.773393
亚于	258	249	0.965116	理会	1847	1422	0.769897

吭声	328	312	0.95122	死活	540	413	0.764815
理睬	521	487	0.934741	一丝一毫	172	131	0.761628
服气	568	522	0.919014	半点	1339	1013	0.756535
再也	5279	4822	0.913431	好意	3627	2742	0.755997
压根儿	219	200	0.913242	未尝	348	262	0.752874
好意思	2530	2289	0.904743	尽如人意	43	32	0.744186
从来	7559	6796	0.899061	断流	78	58	0.74359
丝毫	4096	3559	0.868896	少于	172	127	0.738372
景气	123	106	0.861789	多久	2885	2129	0.737955
确定性	40	34	0.85	源源	432	311	0.719907
久留	174	147	0.844828	万万	863	612	0.709154
搭理	199	168	0.844221	忍心	740	516	0.697297
示弱	316	264	0.835443	强求	179	124	0.692737
介意	679	565	0.832106	介意	2243	1544	0.688364
无时无刻	131	108	0.824427	声张	195	134	0.687179
相干	686	562	0.819242	罢休	399	274	0.686717
在乎	2349	1890	0.804598	怠慢	474	322	0.679325
何尝	491	390	0.794297	乱动	188	127	0.675532
根本	11401	8915	0.781949	全然	667	446	0.668666
相容	128	100	0.78125	吱声	199	132	0.663317
凡响	159	124	0.779874	分毫	389	256	0.658098
截然	622	483	0.776527	磨灭	142	91	0.640845