

詞義預測研究：以語料庫驅動的研究方法

洪嘉馥¹、柯淑津²、黃居仁^{3,4}、安可思^{1,5}、李冠勳²

¹ 台灣大學語言學研究所研究所, 台灣

² 東吳大學信息管理學系, 台灣

³ 中央研究院語言學研究所, 台灣

⁴ 香港理工大學人文學院, 香港

⁵ 香港浸會大學語言中心, 香港

E-mail: jiafei@gate.sinica.edu.tw; ksj@cis.scu.edu.tw;

churenhuang@gmail.com; kathleenahrens@yahoo.com; franker.lee@msa.hinet.net

摘要: 在本研究里, 我們將要探究中文詞匯歧義的所有可能詞義, 主要是為了處理一些還沒有分析詞匯的詞義預測研究, 以及提出更多適合詞匯歧義的解決研究方法。我們打算運用語料庫驅動的語言探究方法在本研究中。我們將關注這些詞義的個別詞義特徵以預測一些還沒分析詞匯的詞義, 而我們所使用的語料庫和工具有: 中文十億字語料庫 (Chinese Gigaword Corpus) 和知網 (HowNet)。使用這些語料庫, 我們將可以藉由語意特徵與概念分群來確認我們在本研究中的目標詞的共現詞匯分群。而這些研究過程也將可證明語料庫為主的研究方法的明顯性。

关键词: 詞匯歧義、詞義預測、語料庫方法、共現搭配、概念分群

Corpus-driven Approaches to Sense Prediction

Jia-Fei Hong¹, Sue-Jin Ker², Chu-Ren Huang^{3,4}, Kathleen Ahrens^{1,5}, Guan-Xun Li²

¹ Graduate Institute of Linguistics, National Taiwan University, Taiwan

² Department of Computer Science and Information Management, Soochow University, Taiwan

³ Institute of Linguistics, Academia Sinica, Taiwan

⁴ Faculty of Humanities, The Hong Kong Polytechnic University, Hong Kong

⁵ Language Centre, Hong Kong Baptist University, Hong Kong

E-mail: jiafei@gate.sinica.edu.tw; ksj@cis.scu.edu.tw

churenhuang@gmail.com; kathleenahrens@yahoo.com; franker.lee@msa.hinet.net

Abstract: In this study, we would like to explore all possible senses of lexical ambiguity in Mandarin Chinese in order to deal with the undefined sense prediction study and to bring about more appropriate lexical ambiguity resolutions. We propose to use corpus-driven linguistic approaches for a sense prediction study. We will concentrate on individual semantic features to predict the senses of non-defined words by using corpora and tools, such as Chinese Gigaword Corpus and HowNet. Using these corpora, we will determine the collocation clusters of my target word through semantic features and concepts. This requirement will demonstrate the visibility of the corpus-based approaches.

Keywords: Lexical ambiguity, sense prediction, corpus-based approach, collocation, concept clustering.

1 引言

本文研究未分析詞匯的詞義預測。在信息工程研究中, 詞義預測可視為一種分群研究, 分群是一種非監督式學習法的方法, 在不需要人工介入的情況下將詞義相似的詞、句子分到同一群組, 分群不同於分類的地方在於分組一開始並不知道數據可以為幾個群組, 也沒有已經定義好的特徵來決定數據該被分到哪一群組。分群的方法可以從概念出發, 將具相同或相似詞義的詞匯分成同一群。目的是利用分群技術找出目標詞匯相似的特徵詞匯, 藉以將不同詞義的目標詞匯及所

属句子做分群；已分群的目标词汇将可以辅助人工词义标记的工作，人工只要将机器无法盼读而分群的词汇标上正确词义即可，以大幅降低人工工作量；已分群的例句将可以做为词义辨识的训练数据。

本研究将探讨分群方法的效果。我们是从概念分群出发。首先，自句子中撷取重要特征，再利用分群算法来计算各特征对于各群组的相似度。得到特征的相似度之后就以此做为分群的依据。

2 文献探讨

当一个词汇同时拥有两个以上的词义，称做是词汇歧义，例如：*bank*。当一个歧义词，事先不知道有多少个词义，以及有哪些词义，这样的歧义辨别，是最困难的歧异辨别。对于概念分群，学者也曾经以语意相似度 (semantic similarity) 来衡量一个新的、尚未分析的词汇与已知、已分析的旧词汇之间的关连程度。语意相似度的衡量大致上可分为两种方式，分别是基于语料库和基于语意距离两种方式。基于语料库的方法则是找出词汇所提供的信息，例如词汇共现频率等，透过较复杂的统计模式来计算相似度的方法 (Resnik, 1999; Jiang 和 Conrath, 1997; Lin, 1998; Li 等人, 2003)。基于语意距离的方法可利用同义辞典所提供的字词树状结构来计算两个字词间的距离，进而推导出相似度。例如有一些研究利用同义词词林 (梅家驹 等人, 1993)、Princeton WordNet (Miller 等人, 1993) 所提供的词汇间的关系与阶层树，使研究者可以利用这些特性来计算语意相似度，知网 (HowNet, Dong 和 Dong, 2000) 则进一步提供义原间的关系，并用以描述概念。

Dai 等人 (2008) 利用知网的特性，从计算义原间相似度延伸到概念间相似度，再导出词汇间相似度。在概念间相似度方面，从每个词汇的定义中找出能够代表大部分词义的「主要义原」与用来描述主要义原的「修饰义原」，得到一个词的概念网络图用以计算概念间相似度，将两概念所属的主要义原、修饰义原做义原相似度计算并加入义原修饰语给予适当权重值，得出概念间相似度，最后取两个词汇中最相似一组概念的相似度得到词汇间相似度。

Jing 等人 (2008) 同样利用义原上下位关系所建立的阶层树来进行义原间的相似度计算，并利用改良式的向量空间模式 (Vector Space Model) 建立概念间相似度，是最有效率的方法之一，并进一步推导至词汇间相似度，也证明可以把知网提供的资源和向量空间模式做结合，增加效率。

3 假设与研究问题

本文中有两个假设。首先，词汇歧义是一些词汇的语言现象，他们同时拥有多样的意义或词义。所以，在中文词汇网络系统里 (Chinese Wordnet (CWN), Huang 等人, 2003)，我们可以藉由这些词汇的词义来决定他们的语意关系。第二，共现搭配的现象呈现词汇结合的情形，不同的结合分布与趋势可作为词义分群的证据。事实上，当词汇共现搭配特征、句法特征、语意特征和词类标记等，在中文原始文档上，必须先准备好的。

基于以上总总因素与假设，本研究提出两个主要研究议题：(1)、我们如何预测词汇歧义的词义并表现出不同的词义在不同的语境或领域？(2)、我们如何使用语料库当作数据来支持我们的词义预测研究？我们将透过中文词汇网络、十亿词语料库 (Gigaword Corpus) 和知网，以探讨词汇歧义的词义语意、词汇特征和词汇共现搭配特征来检测这些研究议题。

4 重要特征收集

在本研究当中，我们将探讨的研究方法，是从概念出发的分群方法。我们使用具丰富语料的中文十亿词语料库第二版 (Lexical Data Consortium, 2005) 做为我们的实验语料库。另外，我们也使用知网将特征词汇转换成概念，利用概念间相似度进而计算出词汇间相似度。

特征词汇撷取阶段，我们选取特征词汇做为句子的特征，以特征来代表句子，透过对特征词汇分群，完成对句子分群。在分群算法中，概念分群算法是将特征词汇进一步撷取出概念群，将概念视为特征词汇的代表，透过对特征概念进行分群，达到对句子分群的工作。

为了收集大量资料以进行我们的词义预测研究，我们选定十亿词语料库中的台湾中央社新闻的内容(繁体中文的语料库)，十亿词语料库的内容(从1991年到2004年)，包含有十四亿的中文字，其中，有八亿字是来自于台湾中央社新闻的内容，有五亿字是来自于中国新华社新闻的内容，和将近3000万自来自于新加坡早报的内容。

在本文中，我们将选取歧义特性的动词做为目标词汇，然后再收集其相关共现搭配组合词汇。我们以三种不同的方式来收集：(1)、动词后紧邻的名词；(2)、动词后的名词组的主要名词；(3)、动词后的第一个标点符号前的名词。这段区间里文字中受词(名词)往往可以做为判断目标词汇词义的重要依据。在收集到的例子中，我们可以归纳几类如下：

表 1: 动词后紧邻的名词

相关联句子	相关共现搭配词汇
民众除了多食用蔬菜，多吃鱼也有益健康。	鱼{Na}

表 2: 动词后的名词组的主要名词

相关联句子	相关共现搭配词汇
一些空军指挥官认为，伊拉克总统沙丹·胡笙玩拖延战术游戏，并希望在地面战斗中有惊人之举。	游戏{Na}

表 3: 动词后的第一个标点符号前的名词

相关联句子	相关共现搭配词汇
现代人多会一种语言,就等于为自己多开一扇文化的窗子。	窗子{Na}

根据以上的三种方法，在表 1 中的句子可以发现，对于动词词义来讲，最具影响性的名词是出现在目标动词后的第一个名词，如：吃「鱼」。但这样的状况若是出现在一个名词词组，则最重要的名词词汇就不是目标动词后的第一个名词，而是对于动词词义最具影响性的名词，是常出现在目标动词后的第一个名词组中的最后一个名词，例如表 2 中，「玩拖延战术游戏」，重点应该是玩「游戏」。不过，在表 3 的句子中可发现，对于动词词义影响最大的名词常出现在目标动词后的最后一个名词，如「开一扇文化的窗子」的开「窗子」。

5 概念分群演算法

由于词形分群算法仅考虑特征词汇的词形，虽然能将一些词形相同、词义相近的词汇分到同一群，但同时也可能将词形相同，但词义并不相近的词汇分到同一群，造成分群上的错误如：山药、药。因此，我们提出一种将特征词汇透过知网转为概念，再透过分析每个词汇的义原组合来进行相似度的计算方式，做为分群的依据。于每个特征词汇我们透过知网撷取出其概念的组成义原。由于多个词汇可能会对对应上同一组概念，这些词汇某种程度上可以被视为同义，如「西瓜」、「柿子」、「苹果」、「葡萄」…等的概念都是水果，都可以视为同义，应该要被分到同一群，也就是说，可以将概念做为特征并计算概念相似度。

对于计算概念相似度，我们使用了修改过的 dice 系数(Dice, 1945)计算公式来做概念相似度的计算，如公式 1。再透过公式 2 来产生最后的平均相似度，以决定分群群组。

公式 1: 概念相似度计算

$$dice_def(m,n) = \frac{2|m \cap n|}{|m| + |n|}$$

再者，我们将 2,915 种的特征词汇对应到知网，其中，有 1,682 种词汇可成功对应。在 1,682 种词汇里，其中，1,286 种特征词汇是属于单义词，共 12,019 个句子，396 个特征词汇是属于多义词，共 7,127 个句子；也就是说，在 1,682 种词汇对应到知网的句子，共有 19,146 笔。在 1,286 种属于单义词的特征词汇，对应到知网，共可分为 666 种不同的概念。在 396 个属于多义词的特征词汇中，经过与知网的对应，取其较常使用的概念，则可得到 195 种不同的概念。换句话说，对于目标动词「吃」来讲，在概念分群的分析中，对应知网的词汇，我们将单义词和多义词合并且去除重复之后，总共可分到 744 种概念。

在概念分群的分析方法上，我们以 744 种概念为依据，所对应到目标词「吃」的句子，共有 19,146 笔，同时也是可以分到我们所设定的群组中，其概念分群的分析，举例如表 8：
表 6：概念分群群组成员

群组	句子	特征词汇	分析结果
「事情+吃」群	1 到了吃饭时间却因此吃不下饭。	1 饭	1 错误
	2 吃晚饭是件辛苦的事，因为不断有人要求与马英九合照。	2 晚饭	2 正确

以表 6 来看，在「事情+吃」群中，例句 1 是概念分群错误的句子，原因是特征词汇「饭」有两个以上的概念，在分群的时候，因为相似度比重的关系，所以错将两个没那么相近的概念分在同一群。

我们在所设定的群组中，随机抽取了 10 个群组，以概念分群为主，所做的人工验证结果，其平均正确率为 89.81%，这 10 个群组的结果，大致如表 7 所示：

表 7：以概念分群的随机 10 组分析结果及准确率

	群组										總句數
	1	2	3	4	5	6	7	8	9	10	
各组句子数	5	4	22	5	6	3	54	3094	24	13	3230
正确句子数	0	3	0	0	4	1	29	2864	0	0	2901
正确率											89.81%

从词形分群与概念分群的方法来做词义预测的研究，由上得知，词形分群演算的正确率是 77.26%、概念分群算法的正确率是 89.81%，两者比较，我们很清楚地知道，对于词义预测的研究，加上了概念的分析元素，是可以达到较高的准确度。

7 结论与未来发展

对于词义预测的研究，本文主要采以概念分群两种算法来进行。在词形分群算法中，虽然有效将特征词汇词形相近的句子分到同一群，对于句子的重要特征词汇在还没有任何的定义时，仍可对于那些句子分群。至于在概念分群算法中，就能有效将句子的特征词形看起来不相似，但概念相近的句子分到同一群，这是基于义原距离的概念分群算法，则更进一步提升群组的质量，更接近人类直观的想法。

未来我们将进一步整合从词形分群与概念分群为出发点的分群算法，我们也将提供更好的特征抽取方式，期望能抽取对目标动词而言，更具代表性的特征词汇，尽可能协助人工词义预测与降低人工词义标记工作量的进行。

8 致谢

本文作者感谢评审委员对于本篇论文提供之宝贵建议，使本文内容更加严谨。本研究

承蒙行政院国科会部分经费补助(计划编号: NSC 97-2221-E-031 -003), 以及中央研究院 94-深耕计划-22, 仅表达感谢之意。

参 考 文 献

- [1] Dai, Liu-Ling, Bin Liu, Yuning Xia and Shi-Kun Wu. 2008 “Measuring Semantic Similarity between Words Using HowNet”, International Conference on Computer Science and Information Technology, pp.601-605.
- [2] Dice, Lee R. 1945. Measures of the Amount of Ecologic Association Between Species. *Journal of Ecology*, 26: 297-302.
- [3] Dong, Zhen-Dong, Dong, Qiang. 2000. HowNet Knowledge Database, <http://www.keenage.com>.
- [4] Huang, Chu-Ren, Elanna I. J. Tseng, Dylan B. S. Tsai, and Brian Murphy. 2003. “Cross-lingual Portability of Semantic Relations: Bootstrapping Chinese WordNet with English WordNet Relations.” *Languages and Linguistics*. 4.3: 509–532.
- [5] Jiang, Jay J. and David W. Conrath. 1997. “Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy”, In *Proceedings of International Conference Research on Computational Linguistics (ROCLING X)*, Taiwan.
- [6] Jing, Peng, Dong-Qing Yang, Shi-Wei Tang, Teng- Jiao Wang and Jun Gao. 2008. “A new similarity computing method based on concept similarity in Chinese text processing”, *Science in China Series F: Information Sciences*, Vol. 51, pp. 1215-1230, no. 9.
- [7] Lexical Data Consortium. 2005. Chinese Gigaword Corpus 2.5.: <http://www ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2005T14>
- [8] Li, Yu-Hua, Zuhair A. Bandar and David McLean. 2003. “An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources”, *IEEE Transactions on Knowledge and Data Engineering*, vol. 15, pp. 871-882.
- [9] Lin, Dekang. 1998. “Automatic Retrieval and Clustering of Similar Words”, *The 36th Annual Meeting of the Association for Computational Linguistics*, pp. 768-774.
- [10] Miller, George A., R. Beckwith, Christiane Fellbaum, D. Gross, and K. Miller. 1993. “Introduction to WordNet: An On-line Lexical Database,” In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence*. Chambéry, France. 28, August–3, September.
- [11] Resnik, Philip. 1999. “Semantic Similarity in a Taxonomy: an Information-based Measure and Its Application to Problems of Ambiguity in Natural Language”, *Artificial Intelligence Research*, Vol. 11, pp. 95-130.
- [12] 梅家驹, 竺一鸣, 高蕴琦, 殷鸿翔. 1984. 《同义词词林》。香港: 商务印书馆香港分馆, 上海: 上海辞书出版社。