汉语语义关联网的研究*

苑春法　许伟　黄昌宁
清华大学计算机系
智能技术与系统国家重点实验室

摘要：本文以《现代汉语辞海》和《同义词词林》为主要资源，在具体词的搭配实例的基础上通过语义类标注建立了语义关联网．语义关联网描述了语义类之间的所有可能的组合关系以及组合模式．它的建立将对汉语词组边界的辨识、句法结构歧义的排除以及汉语语义分析具有十分重要的意义．

A Study on the Chinese Semantic Association Net

Yuan Chunfa, Xu Wei and Huang Changning
Dept. of Computer Tsinghua University
State Key Laboratory of Intelligent Technology snd Systems

Abstract: This paper used the modern Chinese dictionary 《Cihai》 [8] and Chinese thesaurus 《Cilin》 [11] as the main resources to build the Chinese Semantic Association Net(CSAN). We tagged specific words in the collocation examples in 《Cihai》 dictionary with the semantic codes in the thesaurus 《Cilin》. CSAN describes all possible combinatorial relations and patterns between the semantic classes. It will be very useful for the recognition of the phrase boundaries in Chinese parsing, the disambiguation of the syntactic structures, and the semantic analysis of Chinese sentences.

1. 引言

在英语文本中句法分析借助词类(Part of Speech 简称 POS) 标记是十分自然的事情。因为在英语中，明确地规定了词类在句子中的句法功能。例如，动词在句子中一般作谓语，如果要作其它成份必须要有形态上的变化，如加 -ing 或加 .to等。尽管如此，这种模式的句法分析仍然是困难重重；主要是因为自然语言是一个复杂系统。在汉语中的情况还要严重得多，汉语的词类在句子中的句法功能是不确定的。动词不仅可以在句子中作谓语而且还可以在没有任何形态变化的情况下担当定语、宾语或主语。这样，在汉语中仅依赖 POS 标记对文本进行自动句法分析更是十分困难的事情。当然，也可以对句子中这些名词化（或名物化）了的动词加以明确的标记，但在尚未分析、理解的情况下很难界定哪些动词是名词化了的；或者可以以具体词之间搭配关系知识来进行句法分析[1]，但这样一来，由于知识的颗粒度太细，无疑使知识库无限制地膨胀。近年来一些学者越来越重视句法和语义交互分析在自然语言处理(NLP)中的应用[2][3][4]。

*国家自然科学基金资助项目

与传统的词法-句法-语义分析的自然语言处理方法不同，有的学者提出了语义驱动的自然语言处理(Semanticsally Driven NLP)的研究[5]。本文将根据汉语的特点，研究语义类之间存在的句法关联，从而建立一个汉语语义关联网，以求有助于摆脱当前汉语分析的困境。

2. 语义分类在汉语分析中的重要地位

张志公先生曾指出：“从语素到词，到词组，到句子就是一个组合过程，而组合的过程是‘一以贯之’的。各级的组合虽有小异不失大同”[6]。这个“一以贯之”的组合就是汉语的特点，这个“一”就是汉语组合的规律。三个语言层次上的组合遵循着一条规律，那么这个规律又是什么呢？又如何揭示呢？近年来清华大学基于汉语语素数据库研究了汉语复合词的构词规律[7]。人民中国出版社出版的《现代汉语辞海》（以下简称《辞海》）[8]则描述了由词组合成词组的组合规律。它们研究了什么样的语素（类）和什么样的语素（类）以什么样的方式（定中、动宾、壮中、补充和并列等）组合成什么样的词（类）。或什么样的词（类）和什么样的词（类）以什么样的方式（定中、动宾、壮中、补充和并列等）组合成什么样的词组。所有这些均是以语素或词的搭配实例为基础，在语素或词的层面上总结出来的规律。如，[7]指出在汉语复合名词中46.7%由“名+名”语素以定中方式构成，其次，“形+名”语素以定中方式构成的占20.6%，……。但在词的层面上无论如何也没有办法揭示哪些名词和哪些名词可以组合？而哪些名词和哪些名词又不可以组合？哪些名词和哪些名词可以以定中方式组合为名词词组？而哪些名词和哪些名词之间只可以以并列方式组合成为名词词组？这些问题只有通过对那些具体的搭配实例中的语素或词进行语义层面上的聚类或分类才能解决。国立新加坡大学赖金定（K.T.Lua） 博士多年来开展了由字组词的过程中汉字的语义组合规律以及组词的语义转化规律的研究，并取得了许多有益的成果[9]。本文则侧重于语义类之间的组合研究以寻求建立语义类之间的关联网。汉语是一种意合语言，在语义层面上阐述汉语的组合规律，对汉语的构词研究和词组组合规律的研究将具有重要意义。

3. 汉语语义关联网的建立

由前所述，研究汉语中哪些语义类之间可以组合以及以什么方式组合，在汉语的计算语言学中具有十分重要的地位。那么我们如何建立这些语义类之间的组合关系呢？本文采用的方法是在大量具体的搭配实例中对具体词进行语义分类。

3.1 《同义词词林》和《辞海》简介

《同义词词林》[10]是 80 年代出版的一部对汉语词汇按语义全面分类的词典。虽然把它应用于汉语的计算语言学尚有许多不尽人意的地方，但目前仍不失为一个对汉语进行词义分类的主要参考体系。 《同义词词林》（以下简称《词林》）中大类有12个（A类为人，B类为物，C类为时间与空间，D类为抽象事物，E类为特征，F类为动作，G类为心

理活动，H类为活动，I类为现象与状态，J类为关联，K类为助语，L类为敬语），中类有94个，小类有1428个。

《辞海》是一部词语搭配词典，共收入词条 7781 条，搭配实例约77万个。在每个词条下分若干个义项(总计 13292个义项)，在每个义项下有前搭配词、后搭配词。每个词条的每个义项下标有词类，搭配词也标有词类，搭配关系分为定中、状中、动宾、并列、补充和其它等。例如:

#完善｜1.〈形〉完美。

〔在前〕①主谓｜连语:（~+动）~难得｜（~+形）~好｜②补充｜连语:（~+动）~不了｜（~+形）~得多｜~得早｜~得喜人……。

〔在后〕①主谓｜连语:（名+~）办法~｜法规~｜体系~｜制度~｜组织~；……。

2.〈动〉使完善。

〔在前〕①动宾｜连语:（~+名）~措施｜~制度｜~计划｜~构画｜~体制｜……。

〔在后〕①主谓｜连语:（名+~）法制~了｜制度~了｜措施~了｜……。

3.2 语义类间搭配关系的建立

《辞海》中有70余万个词间搭配实例。以具体词间的搭配实例为基础来建立语义类间的搭配，我们可以采取聚类或分类的办法。但从后续汉语分析的可操作性考虑，采用《词林》的分类体系便于文本的语义类标注，因此这里采用了分类的办法。从《辞海》中70余万个词间搭配实例中可以归纳出若干搭配关系，$r_{d1},r_{d2},...r_{di}...r_{dn}$。设这些搭配关系的集合为$R_d, r_{di} \in R_d(i=1,2,.....n)$。则定义，

$$r_{di}=(w_m,w_c,r_{sk},num,w_t) \quad \text{（搭配词在后）} \quad (3\text{-}1)$$

或 $$r_{di}=(w_c,w_m,r_{sk},num,w_t) \quad \text{（搭配词在前）} \quad (3\text{-}2)$$

其中w_m为某个词条的某个义项，w_c为与w_m搭配的某个词。w_t为w_m与w_c相互搭配时出现在它们中间的特征词，如“的”、“和”、“与”……等。r_{sk}为第k种搭配关系，$r_{sk} \in R_s(k=1,2,.....6)$. R_s={定中、状中、动宾、并列、补充、其它}，num为出现关系r_{di}的次数。

同样我们设$r_{sd1},r_{sd2},...r_{sdl}...,r_{sdm}$为语义类之间的搭配关系，设其集合为$R_{sd}$，$r_{sdl} \in R_{sd}(l=1,2,.....L)$则有，

$$r_{sdl}=(s_m,s_c,r_{sk},num,w_t) \quad \text{（搭配词在后）} \quad (3\text{-}3)$$

或 $$r_{sdl}=(s_c,s_m,r_{sk},num,w_t) \quad \text{（搭配词在前）} \quad (3\text{-}4)$$

其中$s_c \in S$，$s_m \in S$，S为语义类的集合。$r_{sdl} \in R_{sd}$，R_{sd}为语义类间搭配关系的集合。设$w_m \in s_m, w_c \in s_c$，即w_m的语义类为s_m，w_c的语义类为s_c。则

$$r_{sdl} \rightarrow r_{di} \quad (3\text{-}5)$$

反之，$r_{di} \rightarrow r_{sdl}$ 不成立。 (3-6)

也就是说，具体词间的搭配关系r_{di}与其相应的语义类之间的搭配关系r_{sdl}之间不存在简单的一对一映射关系。即r_{di}成立，而r_{sdl}不一定成立，但利用由r_{di}抽取出来的r_{sdl}对汉语分析（不是汉语生成）仍不乏指导意义。

3.3 对搭配实例中的词进行语义类标注

主词条w_m是指《辞海》中的某个词条的某个义项。对它进行语义类标注包括如下步骤:

（1）如w_m在《词林》中只有唯一的语义类代码，直接标上即可，否则

（2）根据w_m的词类标记来确定其属于那种语义类代码，如果仍不能确定，则

（3）根据w_m在当前义项下的《辞海》释义文本来确定它属于哪个语义类代码，如果仍不能确定，则

（4）通过人机交互方式加以标注。

搭配词w_c的语义类标注有类似的步骤。关于搭配实例中词的语义类标注的具体算法将另文作介绍。

3.4 语义关联网的建立

在《辞海》中有77万个搭配实例，其中有10万左右是属于由语素组词的实例。由于本文只研究由词间搭配关系来建立语义关联网，所以这10万实例这里不予考虑。在剩下的60余万个搭配实例中仍有一些是新词或即使在人机交互中一时难以确定其语义类代码。基于这些搭配实例建立的语义关联网共有关系178507个；综合为语义中类后的语义关联网共有关系16070个；综合为语义大类后的语义关联网共有关系897个。

例如，

语义类s_m	语义类s_c	r_{sk}	num	w_t
Ga01	Dc01	动宾	2	*
Ga01	Dc01	定中	4	的
Ga01	Dc02	动宾	4	*
Ga01	Dc02	定中	2	的
Ga01	Dc03	定中	12	的
Ga01	Dc04	动宾	1	*
Ga01	Dc04	定中	80	的
Ga01	Dd05	动宾	1	*
……				

注: “*” 表示两类词之间没有特征词

4. 语义关联网中的定中与并列关系分析

在名词词组中出现最多的句法关系是定中。在语义类平面上研究定中和并列关系的组合规律对于名词词组的分析有重要意义。

4.1 语义关联网中的定中关系

在语义类层面上构成定中关系的分布情况如表1 、表2 所示。表1 揭示了修饰词为某一语义类时，哪些语义类的词可以作为中心词构成定中关系，以及它们分别所占的比

例。表2 揭示了中心词为某一语义类时，哪些语义类的词可以作为修饰词构成定中关系及其分别所占比例。从两表中可以看出，D类（抽象事物类）作为中心词的百分比最高（约42%），其次是H类（活动类），再其次是E类（特征类）；在修饰词中，E类做修饰词的百分比最高，其次是H类，再其次是D类。总之 D、 E 、H 是组成定中关系的搭配对中最活跃的三种语义类。经过统计，语义关联网中共有148129个定中关系搭配对，其中以D类做中心词的有62175个约占42%，而在这62175个搭配对中，以E类为修饰词的有17770个，以H类为修饰词的有16461个。

表1: 某类修饰词与各类中心词构成定中关系的分布

修饰词		中心词											
类	比例	A	B	C	D	E	F	G	H	I	J	K	L
A	.06	.03	.03	.01	***.35***	.14	.02	.09	***.27***	.05	.02	.00	.00
B	.05	.01	.09	.03	***.24***	***.27***	.02	.01	.19	.11	.03	.00	.00
C	.06	.04	.07	.03	***.33***	.15	.01	.05	***.23***	.07	.02	.00	.00
D	.17	.02	.02	.01	***.29***	***.28***	.00	.04	.02	.08	.05	.00	.00
E	.25	.07	.14	.04	***.48***	.04	.01	.02	***.15***	.03	.01	.00	.00
F	.02	.05	***.26***	.09	***.43***	.06	.00	.02	.03	.05	.01	.00	.00
G	.06	.10	.04	.03	***.50***	.09	.01	.03	.12	.05	.03	.00	.00
H	.22	.07	.08	.06	***.51***	.10	.00	.02	.07	.06	.01	.00	.00
I	.07	.04	.13	.07	***.41***	.07	.01	.03	.15	.07	.02	.00	.00
J	.04	.04	.06	.04	***.47***	.09	.01	.07	.14	.06	.02	.00	.00
K	.00	.00	.00	.00	.00	.50	.00	.00	.50	.00	.00	.00	.00
L	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00

表2: 某类中心词与各类修饰词构成定中关系的分布

修饰词												中心词	
A	B	C	D	E	F	G	H	I	J	K	L	类	比例
.03	.01	.05	.08	***.32***	.02	.10	***.31***	.05	.03	.00	.00	A	.05
.03	.06	.05	.04	***.40***	.06	.02	***.20***	.11	.03	.00	.00	B	.08
.02	.04	.05	.05	***.24***	.04	.05	***.35***	.12	.04	.00	.00	C	.04
.05	.03	.05	.12	***.29***	.02	.07	***.26***	.07	.04	.00	.00	D	.42
.07	.11	.07	***.37***	.08	.01	.04	***.18***	.04	.03	.00	.00	E	.13
.14	***.17***	.10	.09	***.27***	.01	.08	.04	.07	.04	.00	.00	F	.01
.17	.02	.09	***.19***	***.19***	.01	.05	.15	.06	.08	.00	.00	G	.03
.12	.07	.09	***.22***	***.25***	.00	.05	.10	.07	.04	.00	.00	H	.15
.05	.10	.07	***.24***	.13	.02	.05	***.23***	.08	.04	.00	.00	I	.06
.05	.07	.06	***.39***	***.14***	.01	.07	.11	.07	.03	.00	.00	J	.02
.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	K	.00
.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	L	.00

4.2 语义关联网中的并列关系

在综合为大类的语义关联网中共有并列搭配对3777个（在词类层面上这些可能是“名词+名词”，“形容词+形容词”及“动词+动词”），而其中有相同语义类的组合关系或两语义类之间有特征词如，“和”、“与”……等的搭配对有3661个占总数的96.9%。而无规律可循的仅占3.1%。一般情况下，相同语义类的组合关系为并列关系。但这里也有一个特殊情况，在D类与D类的组合中大部分却是定中关系。据统计D类与D类的搭配对共有7950个，是定中关系的有7261个，是并列关系的有689个。而在689个搭配对中，没有特征词如，“和”、“与”……等的搭配对只有59个。这种情况有待进一步深入研究。

5. 语义关联网研究中存在的问题及展望

由于《辞海》中的搭配实例对是以举例方式罗列的，所以这个资源并不是完备而无遗漏的，因此以《辞海》中的搭配实例而不是以大规模语料库为资源建立的汉语语义关联网也是不完备的。另外，目前建立的汉语语义关联网首先是在细类（清华大学计算机系在小类的基础上又划分为若干细类）层面上构造的，然后综合为小类，综合为中类，再综合为大类，最后形成几个相互独立的语义关联网。这种综合在有些情况下是不成立的。研究哪些关系可以综合而哪些关系又不可以综合乃是今后研究中要解决的问题。语义关联网对汉语分析以及分析中岐义的排除有着重要的意义，但这些应用还依赖于对汉语文本语义类代码自动标注的研究，以及汉语分析的具体算法。我们将对这些问题陆续开展研究，并希望和同行开展广泛的交流。

参考文献

[1]Yuan Chunfa, Huang Changning and Pan Shimei, Knowledge Acquisition and Chinese Parsing Based Corpus , *In Proceedings of the International Conference on Computational Linguistics,* Nantes, 1992.

[2] Mellish,Chris 1985. *Computer Interpretation of Natural Language Descriptions.*Ellis Horwood.

[3] Hirst, G. 1987. *Sematic Interpretation and the Resolution of Ambiguity.* Cambridge University Press, Cambridge, England.

[4] Haddock, Nicholas 1987. Incremental Interpretation and Combinatory Categorical Grammar. *In Proceedings of the 10th International Joint Confrence on Artificial Intelligence,* Milan,Italy.

[5] Schank. Roger C. 1975. *Conceptual Information Processing.* Elsevier, Netherlands.

[6]张志公，谈汉语语素，《语言教学与研究》1981 年第四期。

[7]苑春法、黄昌宁等《现代汉语中二字复合词的构词格式研究》，《计算语言学进展与应用》清华大学出版社，1995年。

[8] 倪文杰、张卫国等《现代汉语辞海》人民中国出版社，1994。

[9] K. T. Lua, A Study of Chinese Word Semantics and its Prediction, In Journal of Computer Processing of Chinese and Oriental Languages, Vol.7 December, 1993.

[10]梅家驹等《同义词词林》上海辞书出版社，1983。