中 文 信 息 学 报
第15卷 第6期 JOURNAL OF CHINESE INFORMATION PROCESSING Vol.15 No.6

二字词词义组合推理方法的研究*

郑家恒 钱揖丽 李 竞
(山西大学计算机科学系 太原 030006)

摘要:汉字是表义文字,具有丰富的语义内容,汉字是一个有限的封闭集,它的数目是有限的,而汉语的词是一个开放系统,它是无限的。本文以"字义基元化、词义组合化"为基本思想,从字义着手,研究二字词词义组合。首先以经过整理的《现代汉语规范字典》、《现代汉语词典》和《同义词词林》为资源,从中自动搜索、抽取出二字词词义组合,建立汉字字义、词义知识库,然后再采用《同义词词林》的语义体系,通过语义相关度等的计算确定它们的组合类型,为研究二字词词义的组合提供一定的参考价值。

关键字:词义;语义相关度;二字词词义组合;词汇学
中图分类号:TP391

Research of Reasoning Method of Two-character Words' Word-sense Combination

ZHENG Jia-heng QIAN Yi-li LI Jing
(The Department of Computer Science ,Shanxi University Taiyuan 030006)

Abstract: As an ideography of abundant semantic contents, the Chinese character is a closed set with limited number while the Chinese word is an open system which is unlimited. Following the idea of" character-sense elementalization and word-sense combinationalization", this paper researches the combination of word-sense with the character-sense as the starting point. Firstly, it establishes the database of character-sense and word-sense by searching automatically the combinations of two-character words' word-sense from three main dictionaries. Then it defines the combination types through the calculating of semantic relativity. The author hopes this paper can provide references for the research of the combination of two-character words' word-sense.

Keywords: word-sense; semantic relativity; combination of two-character words' word-sense; Lexicology

* **收稿日期**:2001-04-20;**修改稿收日期**:2001-09-03
基金项目:山西省自然科学基金(20001032)
作者**郑家恒**,女,1948年生,教授,主要研究领域为自然语言处理。**钱揖丽**,女,1977年生,硕士研究生,主要研究领域为自然语言处理。**李竞**,男,1975年生,硕士研究生,主要研究领域为自然语言处理

一、引言

经验告诉我们，在大多数情况下，即使我们遇到一个生词，我们也能对其进行语义分析以得出其含义。汉语的这个非常明显的特点是出于两个重要的事实：其一，汉语中的词是由语义信息含量很高的汉字所组成的；其二，汉字是有确定语义内容的表意文字。例如：

1）激光——其英文翻译 Laser 是 Light Amplification by Stimulated Radiation 的缩写。大多数人只知其拼写而不知其具体内容，单词本身并不能提供其具体意义的任何暗示。但是对于中文"激光"一词，我们立即知道这是一种光，而且与"刺激"、"激发"有联系。

2）母亲——其英文翻译为 Mother，不能把它分成 moth（飞蛾）和后缀-er 来分析。但在中文中，如果我们将"母"与母性、母亲；亲与亲爱、亲情联系起来，则母亲的意义就自然明了了。

这并不是说我们将词义分析归结为简单的语形分析，但至少可以说明汉字更容易与其它汉字或词组成新的语义单元。毕竟字义是有限的，而由汉字组成的词是无限的，所以我们以"字义基元化、词义组合化"为基本思想，从字义着手研究词义组合。

本文以《现代汉语规范字典》、《现代汉语词典》和《同义词词林》为知识源，运用计算机技术，首先搜索《现代汉语规范字典》和《现代汉语词典》，从中自动抽取得出二字词词义组合，建立汉字字义、词义知识库，然后再采用《同义词词林》的语义体系，通过计算语义相关度等，进行二字词词义组合的分析研究，确定其组合类型。

二、字义、词义知识库的建立

2.1 词典介绍

1）《现代汉语规范字典》

《现代汉语规范字典》，收录了《现代汉语常用字表》中全部 7000 个常用字和一部分能见到的生僻的字。按照"形同而音、义不同的字"，"形、义相同而读音不同，各有适用范围的字"和"形音相同而意义上没有联系的字"等分立字头，并标注词性。对义项不止一个的分条释义，且义项按词义的引申发展脉络排列。先列被引申义，后列引申义。字典中的例词、例句丰富生动，具有典型性，合乎规范，并贴近日常生活。

2）《现代汉语词典》（以下简称《现汉》）

《现汉》是常用的一部汉语工具词典，包含 44389 个词条，52276 个义项。对于任一词条，《现汉》列出了若干释义，其中一条释义对应一个义项。文中二字词的释义就来源于《现汉》。

3）《同义词词林》（以下简称《词林》）

《词林》是一部汉语义类词典，其中包含 52719 个词条，词被分为 12 个大类，94 个中类和 1428 个小类。《词林》中的每个大类、中类、小类都分别有 1 个字母、2 个字母和 4 个字母的语义代码，大类的代码为 1 位，是大写字母 A 到 L，中类的代码为 2 位，其中的第一个字母是它所属的大类的语义代码，小类的代码为 4 位，其中的前两个字母是它所属的中类的语义代码。这样，给定任一小类，我们即可根据其语义代码推知它所属的大类和中类。例如：Aa01 是词林中一个小类的语义代码，则 A 和 Aa 分别是它所属的大类和中类的语义代码。

一个词可以是歧义的，因此它可能属于词林中的几个小类，我们将所有这些小类的 4 位的语义代码都视为是这个词的语义代码，那么一个词就可能有几个语义代码。例如："大"属于词林中的 11 个小类，其语义代码分别为：Ab02 Ah04 Ah05 Dj05 Dn04 Ea03 Ec05 Ed26 Ed38 Jb04

Ka01,那么我们认为“大”这个词就有以上 11 个语义代码。

2.2 知识库的建立

对三部字典的电子版本进行加工整理、以文本形式建立知识库。重点对《现代汉语规范字典》进行了处理,由原始的方正排版文件转换为自定义的文本格式库文件。

《现代汉语规范字典》知识库样例:

＃zx 阿＃xh01＃dyā＃cx＃sy 词的前缀。附着在姓、名、排行或某些亲属名称的前面,常具有亲昵的意味,多用于方言＃jl 阿王|阿毛|阿大|阿婆|阿姨|阿哥 阿妹＃

＃zx 阿＃xh02＃dyā＃cx＃sy|阿昌族|āchāngzú|名|我国少数民族之一,分布在云南。＃

＃zx 阿＃xh03＃dyā＃cx＃sy 音译用字,用于“阿訇”、“阿门”、“阿拉伯”、“阿斯匹林”等。＃

＃zx 阿＃xh04＃dyē＃cx 名＃sy 弯曲的地方＃jl 山阿＃

＃zx 阿＃xh05＃dyē＃cx 动＃sy 曲从;逢迎;偏袒＃jl 刚直不阿|阿谀|阿附＃ys4＃

＃zx 阿＃xh06＃dyē＃cx 名＃sy 姓。＃

＃zx 阿＃xh07＃dyē＃cx 名＃sy 指山东东阿＃jl 阿胶＃

＃zx 阿＃xh08＃dyē＃cx＃sy 音译用字,用于“阿弥陀佛”。＃

＃zx 谀＃xh01＃dyyú＃cx 动＃sy 奉承献媚＃jl 阿谀|谄谀|谀辞＃

其中＃zx、＃xh、＃dy、＃cx、＃sy、＃jl、＃ys 分别表示字形、义项号、读音、词性、释义、举例和引申自。

《现汉》知识库样例:

阿斗:三国蜀汉后主刘禅的小名。阿斗为人庸碌,后来多比喻无能的人。

阿飞:指身着奇装异服、举动轻狂的青少年流氓。

阿附:逢迎附和。

《词林》知识库样例:

阿 = Hi46

阿爸 = Ah04

阿伯 = Ah05

三、二字词词义组合库

3.1 二字词词义组合现象分析

对于词的语义结构前人已有诸多研究,如 Wang(Patrick S P Wang, 1987)认为词的语义结构可分为象形、指事、会意、形声、转注和假借。符淮青(《词义的分析和描写》,语言出版社)认为语素组合成合成词时,常见的关系有一致关系、种类关系、关联关系、借代关系、比喻关系、部分语素义模糊、部分语素义消失和词的全部语素义消失。刘开瑛教授(刘开瑛,2000)将二字结构合成词义的组合方式分为合义、加义、同义、偏义、转义和反义,且进行抽样统计的结果为:合义类型占 26%,加义类型占 50.2%,合义类型和加义类型共占 76.2%。这一结果说明绝大多数词是合义型和加义型,给以“语素义”定“合成词义”的研究提供了良好的基础。

参考以上的词义组合类型研究,考虑到便于计算机处理,我们将二字词词义组合类型分为以下四大类:

合　义:词义与组成词的每个字义有着直接的关系。表示为:X.X→X。

如:悲哀:伤心。=悲:哀痛;伤心+哀:悲痛;伤心

偏义A:词义偏重于第一个字义。表示为:X.Y→X。

如:瞄准:射击时为使子弹、炮弹打中一定目标,调整枪口、炮口的方位和高低。

=瞄:目光集中在一个目标上;注视+准:正确无误

偏义B:词义偏重于第二个字义。表示为:X.Y→Y。

如:沉迷:深深地迷惑。=沉:程度深+迷:醉心于某事物

转　义:词义由其本义转化而来,因而与组成词的每个字义没有直接的关系。

表示为:X.Y→Z。

如:亏损:支出超过收入;亏折。=亏:损失;损耗+损:减少;丧失

其中X,Y,Z表示字义或由其组成的词的词义。

3.2　二字词词义组合库的建立

按照"字义基元化、词义组合化"的思想,一个词的词义是由组成该词的字的字义(语素义)组合而成的。在已有的字典资源中,对于某一字的某一义项,通常都列出了可用该义项解释的若干个例词,因此我们可以从例词入手,自动搜索得出二字词词义组合,表示某词的词义是由组成它的字的哪个具体义项组合而成的。

如:阿谀。《规范字典》中"阿"、"谀"的释义如前所示。可以看出,"阿"的第五个义项(＃zx阿＃xh05＃dyē＃cx动＃sy曲从;逢迎;偏袒＃jl刚直不阿|阿谀|阿附＃ys4＃)和"谀"的第一个义项(＃zx谀＃xh01＃dyyú＃cx动＃sy奉承献媚＃jl阿谀|谄谀|谀辞＃)中均有例词"阿谀",由此可以得出一条字义组合词义现象("阿谀"一词的释义由《现汉》中给出):

阿谀:迎合别人的意思,说好听的话。=阿:曲从;逢迎;偏袒+谀:奉承献媚

说明"阿谀"的词义(迎合别人的意思,说好听的话)是由"阿"的第五个义项(曲从;逢迎;偏袒)和"谀"的第一个义项(奉承献媚)组合而成的。

我们按此算法建立的二字词词义组合库(单义词)中共有2204条记录。部分示例如下:

傲慢:轻视别人,对人没有礼貌。=傲:自高自大,看不起别人+慢:对人没有礼貌

爱惜:因重视而不糟蹋。=爱:怜惜;爱护+惜:爱护;十分疼爱

沉迷:深深地迷惑。=沉:程度深+迷:醉心于某事物

故旧:旧友。=故:指老朋友;旧交+旧:特指老朋友、老交情

四、词义组合推理方法

4.1　语义相关度

1)定义《词林》中两个小类之间的语义相关度。假设C1和C2是任意两个语义类,其语义代码分别为x1x2x3x4和y1y2y3y4,定义它们之间的语义相关度为:

a)REL(C1,C2)=10.0　if x1x2x3x4=y1y2y3y4;

b)REL(C1,C2)=6.6　if x1x2=y1y2 and x3x4<>y3y4;

c)REL(C1,C2)=3.3　if x1=y1 and x2<>y2;

d)REL(C1,C2)=2.0　if x1<>y1 and C1,C2同属于A、B、C、D类或同属于F、G、H、I、J类;

e)REL(C1,C2)=0.0　除以上情况外。

a)表示 C1 和 C2 是相同的,则它们的语义相关度是 10.0;b)表示 C1 和 C2 小类不同,但属于同一中类,则它们的语义相关度是 6.6;c)表示 C1 和 C2 仅属于同一大类,则它们的语义相关度是 3.3;d)表示 C1 和 C2 大类不同,但它们同属于 A、B、C、D 大类(即名词类)或同属于 F、G、H、I、J 大类(即动词类),则它们的语义相关度是 2.0;e)表示除以上情况外,它们的语义相关度为 0.0。

2)为了区别频率高的语义代码和频率低的语义代码的不同影响,我们定义一个语义代码 c 与一语义代码集 S 间的加权语义相关度:

$$REC(c,S)=\sum_{x\in S}K_x\cdot REL(c,x)$$

其中:

$$k_x=\frac{freq\ (x)}{\sum_{y\in S} freq\ (y)}$$

freq(x)表示语义代码 x 在语义代码集 S 中出现的次数

另外,为了进一步考虑释义文本之间在句法层次上的差异,引进以下两条规则:

- 如果某个释义文本的末尾是"的"字结构,则加大"的"字后的名词性语义标记的权值;
- 如果某个释义文本中含有单字词"使",则加大"使"字后的动词性语义标记的权值。

4.2 词义组合类型分析算法

对于字义组合词义分析结果的每一条记录(形如:AB:……=A:……+B:……)

1)首先对其进行分词,然后查《词林》得出各词的语义标记("的"、"使"为特殊词,标记分别为[de00][shi0])。

2)根据释义文本的语义标记集对二字词及组成其的单字词进行义项选择:

a. 从释义文本的语义标记集中选择与被标词有关的语义标记,组成释义标记集 S。

b. 计算被标记词的每一语义标记 c 与 S 的加权语义相关度 REC(c,S),选择其中具有最大值的标记。分别算出词 AB 的语义标记 tag_{AB},字 A 的语义标记 tag_A 和字 B 的语义标记 tag_B。

3)分别计算 $r_{AB,A}=REL(tag_{AB},tag_A)$;$r_{AB,B}=REL(tag_{AB},tag_B)$。然后依以下公式标出词 AB 的组合类型:

合　义: if $r_{AB,A}>=2.0$ and $r_{AB,B}>=2.0$;　偏义 B: if $r_{AB,A}<2.0$ and $r_{AB,B}>=2.0$;

偏义 A: if $r_{AB,A}>=2.0$ and $r_{AB,B}<2.0$;　转　义: if $r_{AB,A}<2.0$ and $r_{AB,B}<2.0$.

五、实验结果分析

我们对所有 2204 条词义组合记录作测试,统计结果如下:

类　型	自动标注记录数
合　义	1322
偏义 A	348
偏义 B	435
转　义	99
合　计	2204

对全部 2204 条进行人工标注,结果有 1807 条与自动标注结果相同,所以:

$$自动标注正确率=\frac{自动标注结果与人工标注结果相同的记录数}{标注记录总数}=\frac{1807}{2204}=82\%$$

例:把守

把守|:|守卫|。|=|把|:|守卫|;|看守|+|守|:|保护|,|使|不|受|损害|;|防卫|	
把守 Hb04	结果:把守 Hb04
守卫 Hb04	
把 Bb03 Bc04 Bh12 Bo25 Dn05 Dn08 Dn09 Ed6 Fa04 Fa06 Hc10 Id21 Kb03	结果:把 Hc10
守卫|看守 Hb04 Ae04 Hm12	
守 Hb04 Hi39 Id21	结果:守 Hb04
保护|使|不|受|损害|防卫 Hi37 Shi0 Ka18 Je13 Je14 Je10 Hb04	

$r_{AB,A}$=REL(Hb04,Hc10)=3.3	结果:合义
$r_{AB,B}$=REL(Hb04,Hb04)=10.0	

例:白昼

白昼|:|白天|。|=|白|:|明亮|+|昼|:|白天|,|从|日出|到|日落|的|一|段|时间|	
白昼 Ca28	结果:白昼 Ca28
白天 Ca28	
白 Bp07 Dk06 Ec04 Ed47 Ee37 Fc04	结果:白 Ec04
明亮 Eb18 Gb08	
昼 Ca28	结果:昼 Ca28
白天|,|从|日出|到|日落|的|一|段|时间 Ca28 Ed28 Ed58 Hi39 Hj19 Hj36 Ka10 Kb02 Ed11 Hf08 Hj20 Hj63 Ie16 Kb02 De00 Dn04 Eb020 Jb01 Dn08 Ca03 Ca23	

$r_{AB,A}$=REL(Ca28, Ec04)=0.0	结果:偏义 B
$r_{AB,B}$=REL(Ca28, Ca28)=10.0	

例:亏损

亏损|:|支出|超过|收入|;|亏折|。|=|亏|:|损失|;|损耗|+|损|:|减少|;|丧失|	
亏损 Dj08 He14	结果:亏损 Dj08
支出|超过|收入|亏折 Dj08 He10 Jb04 Je09 Dj08 Dj08 He14 If24	
亏 He14 He15 If24 Ka26	结果:亏 If24
损失|损耗 If24 Dj08	
损 Je10	结果:损 Je10
减少|丧失 Ih05 Jd08	

$r_{AB,A}$=REL(Dj08, If24)=0.0	结果:转义
$r_{AB,B}$=REL(Dj08, Je10)=0.0	

错误及分析:

经对错误结果分析,我们总结出以下两条错误原因:

1.由于所用资源的信息不全引起的错误:

例:安闲:安静清闲。=安:舒适;快乐+闲:无事可做;空闲

机器计算结果为偏义B,但人工标注为合义,这是由于在《词林》中"快乐"为单义词,只有一个标记Ga01,所以在选择"安"的标记时 REC(Ga07,{Ef07 Ga06 Ga01|)> REC(Ef08,{Ef07 Ga06 Ga01}),导致错选为Ga07,其实应为Ef08。

2.算法本身存在缺陷引起的错误:

例:门生:学生。=门:特指老师或师傅的门庭+生:学生;学习的人

计算结果为合义,但很明显应为偏义。这是由于算法赋予词性相同的词间一定的相关度,说明算法本身不能解决对这类词的分析。 (下转26页)

六、结论

本文将多种选词策略结合起来，提出了一种以实例比较为主，辅以语义模式匹配的选词模型。利用该模型，汉英翻译中的生成系统可以从分析所得的中间语言出发，在单词意义可能的多个翻译结果中选择较符合英语习惯的结果。这种方法已经用在我们开发的汉英机翻译实验系统中。目前，系统中有正反语义模式共计 8642 条，转换成中间语言表示的正反实例共计 22012 条。通过调整词语实例集的例子及词语的语义模式可达到一个比较好的选词效果。同时我们采用的语义知识资源《知网》是专为自然语言处理而设计的，无论是词条数还是语义定义及分类体系都十分适用于自然语言处理。

参 考 文 献

[1] 杨晓峰，李堂秋．洪青阳．基于实例的汉语句法结构分析歧义消解．中文信息学报，2001，15(3)

[2] 董振东，董强．知网，http://www.keenage.com

[3] 熊文新．中间语言机器翻译的有关问题．语言文字应用，1998.3

[4] 王海峰等．汉英双向机器翻译系统 BT863 的研究与实现．情报学报，1997.10

[5] Nirenburg. Sergei and Nirenburg. Irene, A framework for lexical selection in NLG, Proceedings of the 12th International Conference on Computational Linguistics, 1988

(上接 6 页)

从以上实验可以看出大部分词义组合类型均为合义，即词义与组成词的字义有着直接的关系，同时说明通过语义相关度的计算可以表达出词义的组合类型，为研究二字词词义的组合提供一定的参考价值。

参 考 文 献

[1] Patrick S P Wang, The Intelligent Chinese Characters, Proceedings of 1987 International Conference on Chinese and Oriental Language Computing, 85-88

[2] 符淮青．《词义的分析和描写》．北京：语文出版社

[3] 刘开瑛．现代汉语词汇语义网中词义采掘技术研究．多语言处理国际会议，2000.8

[4] 李行健等．《现代汉语规范字典》．北京：语文出版社．1998.1

[5] 吕叔湘等．《现代汉语词典》．北京：商务印书馆．1996

[6] 梅家驹等．《同义词词林》．上海：上海辞书出版社．1983