基于知网的中文结构排歧工具——VXY

董强　郝长伶　董振东

Canada Keenage Inc.

Email:{dongqiang, support, dzd}@keenage.com

摘 要： 本文介绍了基于知网的中文结构排歧工具系列中的一种 — VXY。VXY 采取了一种独到的排歧技术，它对于语言难点采取“定点清除”的策略。它是用来解决那种被习惯地称之为“V+N+的+N”的类型的结构性歧义的。VXY 是一个自足的、可以现场考核检验的并可以真正付诸实用的系统，而不是仅仅某种方法论的表演或举例性的“游戏”。本文简要地介绍了 VXY 的组成部分，说明了它的意义计算的原理。本文并就如何更有效地利用知网进行结构和语义排歧，如何开辟不同于当前语言信息处理中的“三部曲”（语料标注、现成的计算、应试性的评测）的语言技术等问题进行讨论。

关键词： 语义，排歧工具，强支配，动词管辖，中文句法结构，知网，HowNet

HowNet-Based Disambiguator of Chinese Syntactic Structures

Qiang Dong　Changling Hao　Zhendong Dong

Canada Keenage Inc., Montreal, Canada

Email:{dongqiang, support, dzd}@keenage.com

Abstract: The paper introduces a HowNet-based disambiguator, named VXY. The disambiguator effectively tackles the ambiguity in syntactic structures, e.g. “削（V）苹果（X）的 皮（Y）”, which appear highly-frequently in Chinese. The ambiguity of the kind lies in which word is governed by V in the structure, either X or Y. The HowNet-based disambiguator VXY is not merely a demonstration of some stereotypic methodology or algorithm, but is a practical tool that can be tried and tested. A user can test the tool with any structures composed of any one of 98000 unique entries included in HowNet Chinese vocabulary. Hence, the paper presents a paradigm completely different from the state-of the-art human language technology. You are welcome to http://hownet.kookge.com and have a try on VXY by yourself.

Keywords: HowNet, semantics, disambiguator, strong government, Chinese syntactic structure

1. 引言

排除歧义是语言信息处理或计算语言学研究中的关键问题。中文里的 V-X-de-Y 歧义性结构在真实文本中是非常普遍存在的。我们曾经统计过一篇不到 4 页的短文，里面竟有 31 个这样的结构。我们将 31 个含有这样结构的句子，分别输入两个不同类型的机器翻译系统，可以清楚地看到这种歧义判别的正确与否会对自动翻译产生非常严重的影响。这个问题不论采取何种机译策略都是绕不过去的。读者自己不妨也可以自己试试。

我们研究和开发中文的排歧工具的目的有两个，一是用来检验知网的理论与方法的正确性以及数

据的规模和可靠性；二是检验我们对于中文处理的观点和方法，看看是否能够将知网有效地投入实用。就是基于这样考虑我们在不断改进、强化和提升知网知识库性能外，还利用知网先后开发了同时可适用中英两种语言的概念相关性计算器（Concept Relevance Calculator, CRC）、概念相似度计算器（Concept Similarity Measure, CSM）等来作为排歧工具开发的预备性资源。 近一年来我们完成了一个我们称之为基于知网的中文句法结构排歧工具（HowNet-based disambiguator of Chinese syntactic structures）的开发。它被简称为 VXY。

2. VXY

2.1 VXY 要解决的歧义

VXY 排歧工具所要排除的是中文里十分常见的句法结构歧义，即很多学者曾经讨论过的 V+NP1+的+NP2 （削苹果的皮/削苹果的刀）。中文的 V+NP1+的+NP2 结构歧义的关键点是：V 的管辖，即在这类结构里 V 管辖的是 NP1 呢，还是 NP2？与其他学者所讨论的有所不同，我们要解决的是要更加复杂和多样的歧义，即在他们所列出的 NP1 和 NP2 的位置上可以是其他词性的词语，如表 1 所示。 正因如此，我们更确切地命名我们的工具为：V-X-de-Y 排歧工具，简称为 VXY。

表 1 VXY 的各种类型举例

词性分布	举例	管辖类型	举例	管辖类型
VNN	削苹果的皮	type1	削苹果的刀	type2
VNN			骂邻居的孩子	type3
VNV	惨遭暴徒的蹂躏	type1	预报海啸的方法	type2
VVV	预防感染的发生	type1		
VNA	展现孩子的天真	type1		
VVN	看望生病的孩子	type1	搞营销的人员	type2
VAN	吃不卫生的东西	type1	爱美的小姑娘	type2
VVA	证明评测的必要	type1		

我们把 V 管辖的是 Y 的，定为 type1；V 管辖的是 X 的，则定为 type2；如果在判别中某一短语既能适用某条 type1 规则，又能适用某条 type2 规则，那么就被判定为 type3，也即它仍然存在歧义，如“骂邻居的孩子”，这样的歧义结构应该需要更大的语境来解决。

需要说明一点，V+NP1+的+NP2 的管辖关系，还可能包括 V 处于被管辖的关系，例如“失事飞机的残骸”、“进口商品的关税”、“遇害老人的亲属”等。然而我们不会利用 VXY 来解决这样的歧义性结构的判别。这类歧义我们会利用我们正在开发的其他判别工具加以解决。

2.2 VXY 的组成及其功能

VXY 自身主要由以下三个部分组成：

（a）判别器：它的主要功能是调用各种查询和匹配函数，进行词典访问、信息提取、规则匹配。用户填入的词语是它的输入；被判别的结果是它的输出。

（b）规则库：存有判别确定 V 对于 X 或者 Y 的管辖关系的规则。到目前为止，VXY 规则库的规则总数约 200 条。

（c）VXY 用户界面：界面显示 4 个部分：第一行列出 V、X、Y 测试短语输入框；第二部分显示判别器所选定的 V、X、Y 各自的 DEF，即义项的概念定义；第三部分列出判别中所选用的规则；第四行给出了判别的最终结果。这个界面是供用户测试的工具，也是供维护者调试和修改的工具。如图 1 所示。

特别要指出，实际上，整个知网也应视为 VXY 的组成部分。VXY 是完全基于知网的，它直接利用知网的全部资源，特别是知识词典。与知网的其他的意义计算工具一样，知网的更新会引起 VXY 内容上的改善或充实，但不会带来结构上的负面影响。

VXY 的功能是可以对于任意的 VXY 词语组合结构中的 V 对于 X 或者 Y 的管辖关系加以判别。唯有的条件是：（a）输入的组合在意义上应是合理的、真实的；（b）输入的各个词语是知网中所包含的。如前所述，判别的结果有三种：type1、type2 以及 type3 。

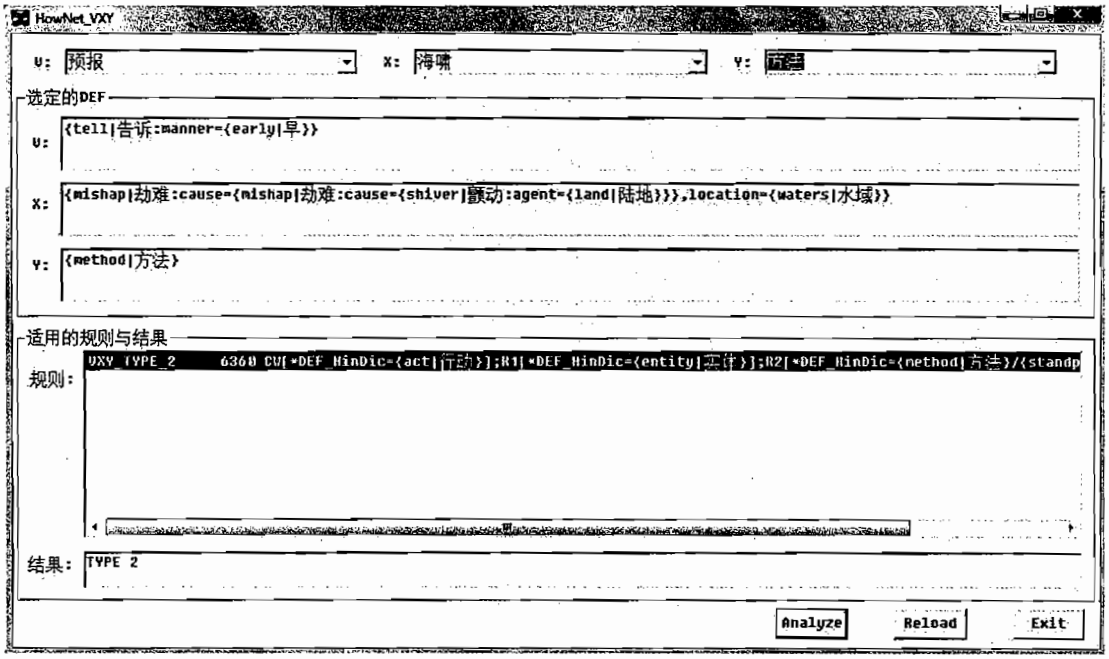

图 1

2.3 VXY 规则的表示

每一条 VXY 规则包括如下几个部分：（a）规则名，（b）序号，（c）规则内容。规则内容由两部分组成：（a）条件部分，（b）动作部分。条件部分是 VXY 三元组：其中 CW 代表当前输入的 V 词语；R1 代表 X，即 V 右邻的词语；R2 代表 Y，实际上是“的”后邻的词语。

CW，R1 和 R2 后面置于方括号[ ]中的是它们各自的语义内容，它们由知网的意义计算函数及其赋值所构成的，三者之间由“;”分割，表示“和”。其中动作部分由“$”引导，@表示“调用”动作类型。每条规则均以句号结束。例如：

VXY_TYPE_2 6360 CW[*DEF_HinDic={act|行动}];R1[*DEF_HinDic={entity|实体}];
R2[*DEF_HinDic={method|方法}/{standpoint|立场}]$@type(2).

2.4 VXY 的意义计算函数

如前所述，VXY 规则采用了知网的意义计算函数。这些函数是知网通用的，应用于各个基于知网的意义计算工具，而并非是 VXY 所专用的。VXY 现今采用如下函数：DEF_HinDic，DEF_inDic，DEF_HostInDic，DEF_WinDic，它们的操作符分别是：=，-，>。

请看下面一条规则，它是在判别“挫伤孩子的自尊心”时，匹配成功的那条规则：

VXY_TYPE_1　2490 CW[*DEF_HinDic={act|行动}];R1[*DEF_HinDic={human|人}];
R2[*DEF_HinDic={mental|精神}]$@type(1).

根据知网，CW 词语“挫伤”有如下 2 个义项，它们的 DEF 分别是：{wounded|受伤}和{discourage|泼冷水}。其中后者的类义原的上位在知网的分类体系（taxonomy）中表示为：

event|事件
　　==> **act|行动**
　　==> ActSpecific|实动
　　　　==> AlterSpecific|实变
　　　　==> AlterState|变状态
　　　　　　==> AlterMental|变精神
　　　　　　==> AlterEmotion|变情感
　　　　　　　　==> excite|感动
　　　　　　　　==> discourage|泼冷水

这样，CW 满足了规则的条件：**{act|行动}**。

R1 词语“孩子”有 3 个义项，其类义原的上位在知网的分类体系（taxonomy）中都是：

entity|实体
　　==> thing|万物
　　==> physical|物质
　　　　==> animate|生物
　　　　==> AnimalHuman|动物
　　　　　　==> **human|人**

于是 R1 也满足了规则的条件：**{human|人}**。最后 R2，其词语“自尊心”的类义原的上位在知网的分类体系（taxonomy）中是：

entity|实体
　　==> thing|万物
　　==> **mental|精神**
　　　　==> emotion|情感

也满足了规则的条件：**{mental|精神}**。因此歧义结构“挫伤孩子的自尊心”被判定为 type1，即 V（“挫伤”）管辖的是 NP2（“自尊心”）。

应注意，规则的意义计算同时兼有词语义项选择的功用。如前面 CW 本来是有两个义项，为何没有选中{wounded|受伤}这一义项呢？那是因为该义项类义原的上位在知网的分类体系（taxonomy）中是：

event|事件

==> static|静态
==> state|状态
==> StatePhysical|物理状态
==> change|变
==> BeBad|衰变
==> SufferFrom|罹患
==> ill|病态
==> wounded|受伤

这样就没有一个上位可以满足规则的条件的，而被摒弃了。

3. 讨论

第一，歧义是语言信息处理的关键问题。歧义有种种不同的类型及其不同的特点，解决歧义也就应该有不同的策略和方法。本文所介绍的中文排歧工具是专门对付中文里普遍存在的一种结构性或管辖性歧义的。这类歧义的一个重要特点在于表面上似乎是因为词类分布产生的歧义如V+N(V)+的+N(A/V)，但实际上它是高度语义依赖的，是由内在的三元的语义约束决定的，是V对于X或Y的强支配性决定的。请比较下面一组例子

新颖的基于语料库的统计分析方法

容易判断自己工作的质量

该组的“V+N+的+N”的前一词语词性均为adj（“新颖的”、“容易”），但其中的第一例为type2，而第二例却为type1。这种词性组合的结构歧义的排除主要是要依靠、或也只能是依靠词语本身的语义。只有当其自身的语义没有可能解决时例如“咬死猎人的狗”，才需求助于更大的语境。至于词汇意义的歧义，虽然也是高语义依赖的，但它们在性质上是完全另外一种类型。因此我们将采用另外的策略和方法。知网已经为此准备好了三种资源：除前已提及的概念相关性计算器（Concept Relevance Calculator, CRC）、概念相似度计算器（Concept Similarity Measure, CSM）外，更重要的是知网的知识词典中为多义词语给出的实例。试以词语“材料”为例，它的三个义项在知网的知识词典中分别列出如下实例：

（1）DEF={InfoElement|信息元素}，（英语=data）

实例：收集~，鲜活的~，熟悉~，调查~，整理~，给~分类编目，手头的~，掌握~，考研~，又发给我们一堆学习~，参考~，第一手~，上报的~中有记载的，一本~

（2）{Quality|质量:host={human|人}}，（英语=makings）

实例：唱歌的~，跳舞的~，不是干这的~，上大学的~

（3）DEF={material|材料}，（英语=material）

实例：建筑~，装修~，买~，家装~，航天飞机外壳是用什么~做的，房屋~，纳米~，~科学

我们相信将上述三种资源有机地结合使用，是实用性地解决词汇多义的有效途径。

第二，进一步讨论关于VXY工具的规则。首先是规则的依据。知网的“事件语义角色框架”和“语义角色与典型演员”是规则的基本依据。知网的这两个文件描述了语义角色与典型演员的强支配关系，例如：

“娶” 对于其 possession 角色：“人，女性”的强支配性

“开办”对于其 PatientProduct 角色：“机构”的强支配性

“医治”对于其 content 角色：“疾病”的强支配性

“买” 对于其 cost 角色：“钱”的强支配性

其次是规则对于词语的义项的选择性。当 VXY 三个词语的任何一个有多个义项时，规则有能力进行自动的选择，这是很重要的机制。再者是规则的自动的上下位查询的机制。

第三，我们应该采取怎样的策略和方法来解决歧义问题呢？现在我们看到有两类做法，一类是本质上应属于语言学本体研究的，或者属于我们称之为无计算的“计算语言学”（computational linguistics without computation）的方法；另一类是眼下尚流行的“三部曲”（语料标注、现成的算法、应试性的评测）方法。这两种方法都不是我们所赞成的。我们主张的是：对于汉语的语言难点应采取“定点清除”的策略，不同类型的歧义应采用不同的方法去解决， 我们正在努力开发不同类型的排歧插件，供用户选择、嵌入用户自己的语言信息处理系统，如文本挖掘、机器翻译系统等。 换句话说，我们要的是可以经得起任意考核的排歧系统，而不是只能演示或评测几十个多义词的“玩具”。

4. 今后的工作

我们现已将 VXY 上传至 http://hownet.kookge.com，我们将通过读者和用户的测试反馈来改进和完善它。我们真心地愿意看到有人采用其他的方法（如词性标注下的“三部曲”）或其他的资源（如 Chinese WordNet 等）来做出类似的排歧工具并进行开放性的考核，我们相信这样的比较才会更有意义。

如今，我们已开始开发新的中文排歧工具，如 VN、NV 工具，用以解决诸如“医治疾病”/ “走私集团”，“太空行走”/ “群众抱怨”等管辖关系歧义。同时我们通过我们正在研发的基于知网的英中机器翻译系统，开发英语的排歧工具。最后，我们愿意与其他机构合作共同开发更多的实用的排歧工具。

参考文献

[1] Zhendong Dong, Qian Dong, HowNet and the Computation of Meaning, World Scientific, 2006

[2] 冯志伟 自然语言的计算机处理，上海:上海外语教育出版社，1996

[3] 冯志伟，论歧义结构的潜在性，中文信息学报，1995 年，第 2 期

[4] 苑春法，黄锦辉等，基于语义知识的汉语句法结构排歧，中文信息学报，1999，13（1）

[5] 张克亮，基于 HNC 理论的句法结构歧义消解，中文信息学报，2004，第 6 期 pp 43-52

附录

1. 关于建议设立“汉语句典”课题的刍议	V 设立课题的刍议
2. 难以打开局面的看法，的确反映了很多人的忧虑	A 打开局面的看法 + 反映很多人的忧虑
3. 较为新颖的基于语料库的统计分析方法以外	A 基于语料库的方法
4. 那是值得研究的问题	V1 值得研究的问题
5. 少数人期待有关自然语言的“日心说”的出现	N 期待日心说的出现 （1）

6. 目前研究自然语言处理的方法好比托勒密的理论	N 研究语言处理的方法
7. 坚持这种扭曲的理论的结果是...。	坚持理论的结果 V
8. 儿童学习自然语言的过程	N 学习自然语言的过程
9. 要是我们把观察和思考问题的角度变换一下	P 思考问题的角度 V
10. 即采用适合计算机的特点的方法	V 适合计算机的特点 （1）
11. 可以用一套形式语法系统来描述是这种方法的基石	V 是方法的基石 （1）
12. 也是处理这种语言的切入点	V1 处理语言的切入点
13. 支撑自然语言大厦的主要支柱可能不是	支撑大厦的支柱 V
14. 我们仔细观察小孩子学说话的过程	N 学说话的过程
15. 一个一个地掌握各种句模的用法	Ad 掌握句模的用法（1）
16. 从而提高他们的说话和理解能力	Ad 提高他们的能力 （1）
17. 尤其是在研究别人的言语	Ad 研究别人的言语 （1）
18. 才打破了不能开口的局面	Ad 打破开口的局面（1）
19. 我们也有教外国人的《汉语 400 句》了	V 教外国人的汉语 400 句
20. 以上的说法并不是完全否认“语法”的作用	Ad 否认“语法”的作用（1）
21. 旧句模的消亡过程受到全社会成员的参与	N 受到成员的参与（1）
22. 《汉语 400 句》就是一个《1 级汉语句典》的雏形	N 是句典的雏形（1）
23. 类似于人类自己掌握自然语言的过程	N 掌握自然语言的过程
24. 容易考核工作的实际进展	A 考核工作的进展（1）
25. 容易判断自己工作的质量	A 判断工作的质量（1）
26. 我提出上述建立《句典》的建议	N 建立句典的建议
27. 任一语句是否属于本句型的算法	N 属于句型的算法（1）
28. 而这个课题所要解决的是面向计算机的句典	V1 面向计算机的句典
29. 根本无法纳入我们心目中的《句典》中	Aux 纳入心目中的句典（1）
30. 以上是个人的浅见	A 是个人的浅见（1）