汉英机器翻译系统中的一种词义排歧方法

杨晓峰　　李堂秋　　洪青阳
厦门大学计算机系，厦门，361005

摘要：本文论述了一种基于中间语言的机器翻译的词汇排歧方法。该方法从大规模未标注文本中获取词语义原的同现集合，并以此为参考成生义原的语义限制规则，对机器翻译的分析阶段产生的中间语言中进行词义排歧。文章首先提出了这种方法的总体思路，并对其语义知识资源—《知网》作了简要的介绍。然后详细地描述了排歧的算法。最后的实验结果证明，这种方法是对词义排歧是有效的。

关键词：词义排歧、 机器翻译、 义原、 中间语言、相似度、 知网

Word Sense Disambiguation Method in Chinese-English Machine Translation System

Yang Xiaofeng　　Li Tangqiu　Hong Qingyang
Department of Computer Science, Xiamen University, Xiamen ,361005

Abstract: This paper presents a description of a word sense disambiguation method applied in machine translation. This method generates semantic restricted rules for each sense-atom according to the co-occurrence sets obtained from a large corpus, and then makes the word sense disambiguation for the intern-lingual. In this paper we first put forward the general idea about this method and give a brief introduce to its semantic knowledge resource — *the Hownet Dictionary*. Then the main algorithm for the method is proposed with detail. The experiment result given in the end proves our method to be effective.

Keywords: Word Sense Disambiguation, Sense Atom, Machine Translation, Similarity, *HowNet*

词义的排歧，也称为词义标注，是自然语言处理研究领域的一个难题。词义排歧主要有基于知识和基于语料库两种方法。前者如 Wilks 提出的应用规则来选择限制从而进行词义消歧。这种方法需要由专家组织手工进行编制规则库，主观性太强并且规则库存在着一致性、扩充性、及完备性等问题[1]。基于语料库的方法一般是利用语料库进行词语或义项的同现统计。语料库要事先进行词义的人工标注，因此该方法也难以大规模地使用。还有的研究者使用未标记的语料库进行无指导的词义排歧，取得比较高的正确率，但是排歧结果的优劣受到多义词的语法特性的影响，对于复杂的句式结果的排歧效果就不如简单句式理想[2]。

本文研究目的是为基于中间语言的机器翻译系统提供词汇级译文选择。在本文中提出了一种将基于规则与基于统计相结合的词义排歧方法。该方法以《知网》为主要的语义源，利用无指导的学习从语料库中抽取词义的排歧信息，并以此为参考按指定的转换模板构造出词义的限制规则。我们将语义限制规则运用于机器翻译分析阶段产生的中间结果上，中间结果中得到的词语上下文的语法语义信息可以为词义排歧提供更准确的判断依据，从而能有效地提高排歧的正确率。该方法在解决词义消歧的同时也可以对中间结果进行语义结构调整。

1．知网简介

《知网》是一个以汉语和英语的词语所代表的概念为描述对象，以揭示概念与概念之间以及概念所具有的属性之间的关系为基本内容的常识知识库[3]。与《同义词词林》不同，在《词林》中，是通过对词语进行语义分类来表示其意义的，而在《知网》中没有类别的概念，每个词语的意义来是由若干个义原及其关系来进行定义的。义原是知网中最基本的、不易于再分割的意义的最小单位，知网通过对约六千个汉字进行考察和分析来抽取了 800 多个义原，并总结了如部分、主体、客体、从属、时空、材料等若干种义原间的语义关系。由这些义原并上语义关系可以定义《知网》的所有词语义项。

例如“面”这个词语包含有如下两个义项：

W_C=面
G_C=N
W_E=noodles
G_E=N
DEF= food|食品

W_C=面
G_C=N
W_E= face
G_E=N
DEF= part|部件,%AnimalHuman|动物,skin|皮

第一个义项的“面”作“面条”解，它的义项定义是 fool|食物；第二个义项的“面”作“脸面”解，它的义项定义说明在这个义项中面是动物的部件，是“皮”。

在我们的算法中语义相似度是建立在义原级上，而不象传统的是建立在义项级上的。《知网》提供了义原分类树，分类树把各个义原及它们之间的联系以树的形式组织在一起，父子结点的义原具有上下位的关系，即父结点是比子结点更抽象的概念描述。图 1 描述的《知网》中动作类义原的分类树：

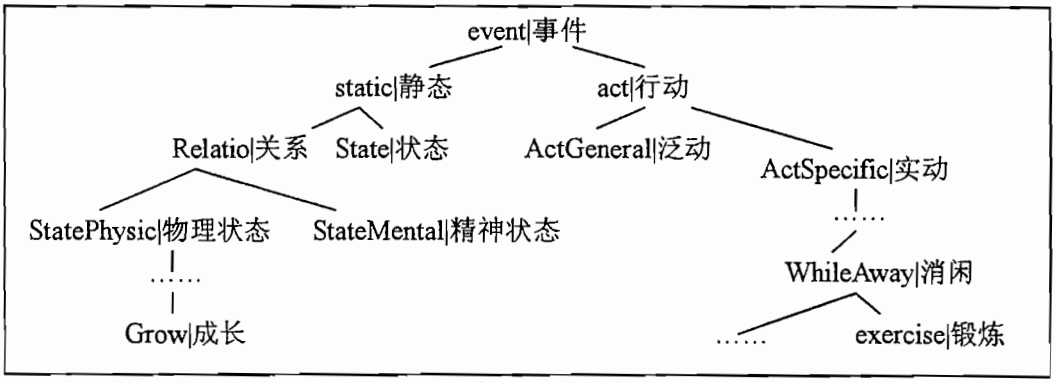

图 1《知网》中的动作类义原分类树

2．排歧算法描述

机器翻译分析阶段所产生的中间结果里包含有句子的语法及不完全的语义结构信息，如果能在词典中为词语的各义项标记其在实际句子中各语义格应出现的特征义原，我们在进行词义排歧时就可以在当前词语的特定上下文中选择一种搭配的意义组合，使之最满足词典中的语义特征和限制。在《知网》中，词语的义项是由义原构成的，我们可以考虑将限制规则建立在义原一级，而不是义项一级上。如果我们能为动作类义原及属性类义原定义适当的语义规则，就可以很容易地通过义项的义原的构成情况自动进行义项的语义限制比较计算。

2．1 从语料库中获取义原的上下文同现义原集

语言学家认为，词语的意义只能在上下文中才能得以辨识。如果一个词语的某一词义在语料库中出现多次，我们在其出现的上下文中可以发现某些词语出现的频率很高，而这些词语与该词义之间有着比较密切的搭配的关系。词义由义原构成，同样地，与一个特定义原在上下文中共同出现次数越多的义原也与该义原有越密切的语义关联，而这些的搭配义原就可以为我们制定这个义原的语义限制规则提供很好的参考依据。动词、形容词的词义对句子的语义影响最大，而动词、形容词的主要构成是动作类义原和属性类义原，因此我们希望得到这两类义原的同现集合。义原的同现集合定义为：

$S(a)=\{(b\ Prob(a,b)) \mid b\in ATOMSET, Prob(a,b)>THRESHOLD\}$ **(1)**

Prob(a,b)为 a,b 义原的同现概率，ATOMSSET 是《知网》中所有义原的集合。THRESHOLD 是预定义的一个阈值，集合中义原的同现概率都应大于这个阈值。对于动作类义原，其前后出现的词语充当的语法成分是有很大区别的，因此我们有必要区分在义原前出现与后出现的义原。令 S-(A)、S+(A)分别代表 A 的前、后同现集合。而对于属性类义原则无需作此区分。

算法中义原的同现集合获取过程是无指导的，语料库不需要做任何的语义标注，但是我们事先要对其进行自动分词及词性标性。算法主要处理动词及形容词，及这些词的上下文窗口中的动词、名词、形容词及副词。本文选取的窗口的大小为前后 6 个词。

在算法中我们只对单义词进行处理。因为我们无法确定语料库中一个句子里某个词语的词义究竟是什么，如果把它的所有的义项都参与同现统计，势必会使统计结果含有较大的噪音，因此选取的词语应是无词义歧义的。即只对单义词进行处理，而不考虑多义词。注意这里的“单义”是相对于词性而言的，词语可能具有多个词性，如果词语 W 在某个词性 C 中的意义是唯一的，我们就称 W 在词性 C 下是单义的。

算法 1：义原的同现集合获取算法

对于语料库中的一个输入句 S，对每个单义词语 $W\in S$, $Category(W)\in\{v,adj\}$

设 W 的义原定义集合为 Atoms(W)，W 的上下文窗口为 Winow(W)，则对每一个 $a\in Atoms(W)$在当前窗口的同现集合为

$$Concurrent(a)=\bigcup_{k\in Window(W)} Atoms(k) \qquad (2)$$

令 Total=Total+1。

同时对所有的义原 $b\in Concurrent(a)$ 都做

COUNT[a,b]=COUNT[a,b]+1;

其中 COUNT[a,b]是 b 对于 a 的同现次数。

在对语料库中的句子处理完毕后，可以计算出每个义原的同现义原集

S(A)=

$\{b.freq(a,b) \mid freq(a,b)=10^3*count[a,b]/Total, b\in ATOMSSET$ 且 $freq(a,b)>THRESHOLD\}$ **(3)**

算法中要注意动作类义原应分别根据上下文窗口的前后两部分计算前、后同现集合。

我们的统计语料来源是《读者 20 年文集》，规模大小为 1,100 万字。经自动获取得到的义原规则为 718 条。

2．2 产生义原语义限制规则

利用用算法 1 可以得到每个动作类及属性类义原的上下文同现集合。实际上到这一步我们已经可以根据同现集合与测试句中词语的上下文语境进行相似度的计算，将相似度最高的义项作为该词的在当前句子中的词义。这种方法对于简单的句式结构有比较高的正确率，但对于具有特殊的语法性质的词语则排歧结果不是很理想。如对动词来说，带复杂宾语（如小句宾语和兼语宾语）的多义词的词义排歧结果会差于带简单宾语的多义词[2]。这是由于在复杂语式里中心词语与搭配词语距离较远，搭配词语或是超出中心词语的上下文窗口范围，或是与中心词语之间有过多的干扰词语。

在机器翻译的语法分析阶段成生了源语句的中间语言，它可以更为准确地描述词语所在上下文的语境；同时我们也为义原定义语义限制规则，它描述了含有该义原的词语的期望出现的语义环境。这样我们可以根据词语实际所处的语义环境与义原规则中描述的语义环境进行相似度的计算，将比较结果作为词义排歧的依据。

在本文中考虑采用格结构的中间语言模型。在格结构中动词或形容词的语义环境是由 Agent、Theme、Result、Location 等格信息来描述。由于缺少语义信息，在分析阶段给出的语义格信息并不一定正确，我们可以在词义排歧阶段进行自动的语义格调整。

在义原的语义限制规则中定义含有该义原的词语的语法格应该出现义原的逻辑关系。这些逻辑关系用 AND、OR、NOT 等符号来表示。AND 表示在格中希望限定的义原都出现；OR 表示格中只要出现任意一个指定义原即可；而 NOT 表示不希望出现指定的义原。

义原的语义限制规则的模型表示为：

Logic-Rule = ({(Sem-Case Logic-Item)}*)

Logic-Item = (Logic-Op {Logic-Item}*) | {(SenseAtom Prob)}*

Sem-Case = Agent | Theme | CO-THEME | Result | Clause | . . .

Logic-Op = AND | OR | NOT

为了将 2.1 中得到的义原同现集合的信息充分运用到义原的限制规则定义中，我们采用了以下的转换模板，用于自动根据同现集合构造出义原语义规则。

a)：对于属性类义原，同现集合中实体类的义原做为 THEME 格。

b)：对于动作类义原，前同现集合做为义原的 AGENT 格；

c)：对于属性类义原，后同现集合中，属性类的义原做为 RESULT 格；

d)：对于属性类义原，后同现集合中，实体类的义原做为 THEME 格；

e)：对于属性类义原，后同现集合中，动作类的义原做为 THEME 格；

此外，对于如“URGE|促使”、“ EXPECT|期望”、“ GIVE|给”等具有特殊语法性质的词语的义原，，需要手工地为其定义 CO-THEME、CLAUSE 等格。

例如对于动作类义原 eat 吃，义原语义限制规则为：

```
(eat|吃
 (AGENT ( OR (HUMAN|人   0.056) (THIRDPERSON|他   0.041) (MALE|男   0.028) (MASS|众   0.023)
             (PLACE|地方   0.022) (BIRD|禽   0.020) (PROPERNAME|专   0.020) …  ).
       )
 (THEME (OR (MEDICINE|药物   0.052)(PART|部件   0.038) (HUMAN|人   0.020)
            (FOOD|食品   0.015) …  )
 (RESULT (OR   ((ATTRIBUTE|属性   0.026)   (DESIRED|良   0.020) )
)
```

2．3 相似度的计算

1）义原间的相似度

根据《知网》中提供了动作类，实体类、属性类等义原分类树，可以定义义原之间的相度：

$$SIM-ATOMS(A,B)=\begin{cases}0 & \text{如果义原A,B不在同一棵分类树上}\\ 1-(MIN-SEM-DISTANCE(A,B)/MAXDISTANCE(TREE(A))) & \text{；否则}\end{cases} \quad (4)$$

其中 MIN-DISTANCE(A,B)是 A，B 在其分类树中的最小语义距离，MAXDISTANCE（TREE（A））是义原 A 所在分类树的最长的义原距离。

2)义项与格限制描述规则的相似度

义项的语义限制规则中定义了某一语义格可能出现的义原及其概率，我们要把当前词语的语义环境，即相应的语义格包含的词语的各个义项与限制规则中指定的语义格进行相似度的计算，以求得一个最能满足限制规则条件的词语义项。

设义项 Entry={e1,e2,…,ex}，义原 a 的规则中指定某格 Case 的限制条件是

CaseRule = (OP CR1, CR2, …,CRm,(b1 P_{b1}),(b2 P_{b2})…,(bn P_{bn}))

其中 CR1,CR2,..,CRn 是带有逻辑算符的义原集，b1,b2,…,bn 是义原。$P_{b1},P_{b2},\ldots P_{bn}$ 分别是 b1,b2,b3 对义原 a 的同现概率。则

$$SIM-CASERULE-ENTRY(Entry,CaseRule)=\begin{cases}\mathbf{MAX}\,Rs & \text{当OP = OR 时}\\ \mathbf{MIN}\ RS & \text{当OP = AND 时}\\ 1-\mathbf{MAX}\,Rs & \text{当OP = NOT 时}\end{cases}$$

其中

$$Rs=\{SIM\text{-}CaseRule\text{-}Entry(Entry,CR_j)\mid j<m\}\cup\{SIM\text{-}ATOMS(e_i,b_j)*P_{bj}\mid i<x,j<n\} \quad (5)$$

2．4 词义选择

1) 词语义项的评价分值计算

义项的评价是根据词语所在的上下文环境，与义项的构成义原的语义限制规则之间的相似度来计算。算法的描述如下：

算法 2：词语义项评价算法

a) 对于词语 W 的义项 E 的每个义原 A，令累计总分数 T=0；

b) 查找 A 的义原规则集 Rule(A)，设 Rule(A)中含有 N 条格限制描述。

c) 对于 Rule(A)中的每个格 C 的限制义原集合 SC，在词语 W 的格框架中查找格 C 的词语 W’，如果格框架中不存在对应的 C，则 T=T-100，则否则执行 d)

d) 对词语 W’的每个义项 E’，我们根据公式(6) 计算它与 SC 的语义相似度。设义项 E=E’时将得到的最大相似度为 M，令 T=T+M；

e) 义原 A 的最后评价分值为 T / (Rule(A)的格数目)。

f) 义项的评价分值为其构成义原的最大一个评价分值。

2) 词义选择

对于动词及形容词，我们可以根据算法 2 计算出其各义项的评价分值，选择具有最大评价分值的义项做为该词语在当前句子中的词义，而对于非动作类和非属性类的词语，在算法 2 的 d)中我们已经得到的 E'即为词语 W 在当前窗口中的选择义项。

2．5 调整分析结果的语义框架

在 2.2 中我们提到了分析的中间结果可能会存在语义格错误。如下面的例句：

"被子叠得整整齐齐"

在这句话中"被子"是做主语，语法分析很有可能就会误把它作为 AGENT，而实际上它却是 THEME。在词义排歧时获得的语义信息使我们能够对结果的语义框架进行适当的调整。首先在语法分析阶段的语法词典中为"叠"这样具有主语充当受事体的动词做上标志，在语义排歧时检查词语是否存在这个标志。如有，并且词语的 AGENT 格的评价分值低于某阈值，便将这个 AGENT 格标志成 THEME 格再做一次评价测试，如果评价分值高于阈值，我们就可以认为在此结果框架中的 AGENT 应调整为 THEME。

3．试验结果及总结

3．1 试验结果

在义原的同现集合获取中我们使用《读者 20 年文集》作为统计语料库，测试时也使用同类的语料。实验对词义排歧的正确率进行统计，计算公式为：

词义排歧的正确率=词义判断正确的词语数/测试语料中歧义词的总数。

通过对语料库中的 2，000 个测试句进行排歧试验，正确率为 92%。测试结果表明，利用语料库的同现义原来构造义原的语义限制规则，并以此进行词义排歧的思想是合理的。

3．2 总结

(1) 在《知网》中词语的义项由多个义原定义，不象传统的分类词典中义项只是一个类代码，这样对词语义项的意义描述更加全面，丰富。

(2) 利用义原的规则与当前词语所在语义环境进行相似度的比较进行词义排歧，可以提高复杂句式结构的词义排歧正确率。

(3) 将排歧知识建立在义原的基础上，义原的数目是有限的，这样避免了手工编制大规模词义排歧知识的繁重劳动。同时义原的排歧知识是参考义原的同现集合，而这一集合是通过对语料库无指导学习获取的。这样知识的获取的工作量进一步的减少了。

(4) 利用词语义项的评价算法，在词义排歧过程中可以对中间语言的语义框架做适当调整。

参考文献

[1] 赵铁军等.《机器翻译原理》：P300-317．哈尔滨工业大学出版社，2000。

[2] 李涓子，黄昌宁．一种基于无指导的词义排歧模型，JSCL'99。

[3] 董振东，董强．知网，http://www.how-net.com。

[4] 李涓子，黄昌宁．基于转换的无指导词义标注方法．清华大学学报（自然科学版），1999 年，第 39 卷，第 2 期。