文章编号:1000 - 582X(2005)03 - 0069 - 03

机器翻译中词义的常识排歧*

段 绮 丽
(中国电信 成都分公司,四川 成都 610051)

摘 要:提出了一种机器翻译中多义词词义排歧的新方法。首先对翻译过程中多义词的词义选择是否符合常识给出了一条形式化的标准,然后将人们在翻译过程中排歧时所进行的逻辑推理归结为一种机械的集合运算,使之易于机器操作。在此基础上建立了义项多元组的概念,利用此多元组引入词义的语境相关限制信息,以改进现有电子词典,使之更加利于排歧。并从方向上指明了这种词典知识获取的途径。

关键词:常识排歧;义项义素集;义项多元组;语境相关元
中图分类号:TP181 **文献标识码**:A

1 问题的背景

语言歧义的排除是机器翻译中长期存在的一个难题。要想较好地解决这个问题,可以参照人在自然语言翻译中排除歧义的处理过程:首先人脑中要存有大量反映自然语言词汇、语法等基本语言特性的语言知识(这在人们的语言交流中已成为常识),在翻译时便利用这些常识进行推理,得出符合逻辑的判断、作出恰当的选择,进而排除歧义。因此机器翻译的排歧过程,也应该是首先获取充足的自然语言常识,再利用这些常识进行逻辑推理来完成的过程。

国内外在利用常识排歧方面已经开展了不少的工作。例如普林斯顿大学的 Wordnet、中科院董振东先生的 Hownet,香港的《同义词词林》、中科院黄曾阳先生的 HNC(概念层次网络)等系统中已日益完备地反映了自然语言词语的意义类别(或义原、义素)和概念体系,提供了极为丰富的语言常识,而围绕利用这些资源进行常识排歧的研究也正在广泛开展[1-3]。

当前的问题是:1)如何利用上述词典中提供的词义常识进行同样符合常识,即符合日常逻辑的推理,以实现合符常理的词义筛选? 2)如何将上述类型的词典改造成为新型的词典,将自然语言理解过程中制约多义词词义筛选的语境相关限制信息纳入词典,以提高排歧的准确性?

为了解决以上 2 个问题,笔者首先对翻译过程中多义词的词义选择是否符合常识给出了一条形式化的标准。然后将人们在翻译过程中排歧时所进行的逻辑推理归结为一种机械的集合运算,使之易于机器操作。在此基础上建立了义项多元组的概念,利用此多元组引入词义的语境相关限制信息,以改进现有电子词典,使之更加利于排歧。并从方向上指明了这种词典知识获取的途径。

2 常识排歧的形式化

Wordnet 这类词典中提供的词义常识,大约可以归结为语言学中的"义素"或"词素"。所谓义素,是指一个词义项所代表的概念之内涵的若干基本特性。例如,husband 作"丈夫"讲时,这一义项显然至少包含以下义素:"man(人)""adult(成年的)","male(雄性的)","married(已婚的)"。形式地,义项"丈夫"可以看作是以上 4 个集合的交集。现在的问题是,如何利用词典中的词义常识——义素来进行多义词词义义项的筛选。显然,首要的问题是要对这种筛选之合理性给出一条可靠的形式化标准。

笔者提出的标准,主要是基于罗素的摹态词理论。大家知道,自然语言可以用谓词逻辑形式化地加以表示,例如,按罗素的摹态词理论[4]句子"A panda is a beast."应当形式化为表达式:$\exists x$ (Panda(x) $\wedge$ Beast(x))。要使上式能够成立,须存在 x 既满足谓词 Panda(x) 又满足谓词 Beast(x),因此,谓词 Panda(x) 和 Beast(x) 的论域(分别记为 D_p、D_b)的交集非空,是上式成立的必要条件,也即是句子合符常理,合符逻辑的必要条件。将论域中元素的诸特性用二词语之义项所包含的义素的集合 $E_p = \{p_1 \cap p_2 \cap \cdots \cap p_n\}$、$E_b = \{b_1 \cap b_2 \cap \cdots \cap$

* 收稿日期:2004 - 10 - 22
作者简介:段绮丽(1974 -),女,四川宜宾人,工程师,主要研究方向:计算机软件与理论。

$b_m\}$(其中 p_i、b_i 分别表示 D_p、D_b 中元素的特性)来表示,要使 $D_p \cap D_b \neq \Phi$,显然必须有 $E_p \cap E_b \neq \Phi$。因此,可将相邻相关词的义项义素集之交非空作为判断词义义项之筛选是否符合常理与自然语言日常逻辑的标准("相邻相关"是指短语或句子之相邻实词所蕴涵的判断中的主谓相关),按此标准,若相邻相关词的义项义素集的交集为空,则对当前义项的认定不符合常理与自然语言的日常逻辑,反之便合符常理。这样,对一个多义词的义项的筛选是否合理的判定,亦即词义排歧过程中进行的日常逻辑推理,便已经归结成了机械的集合运算。要利用以上标准,通过集合运算,选择出正确的义项,必须构造足够精细的义项义素集词典(精细常识),使得只有符合文本上下文的词义的义项义素集才能与相邻相关词的义项义素集的交集非空,而其它词义的义项义素集与相邻相关词的义项义素集的交集为空。在获得了足够的精细常识后,就可以使用判断相邻相关词的义项义素集之交是否为空的标准进行排歧,剔除那些交集为空的义项,选择出正确的词义。

目前,不少义素词典已提供了许多现成的义素资源,充分利用这些资源,已不难构造相应的义项义素集,并解决不少词义排歧的问题。

例如,判断"A panda is a beast."中 beast 的词义,可以利用现有义素层级词典"Wordnet"[5],从词义的层级关系中搜得如下的义项义素集:

panda(n. 熊猫)——P = {physical thing, organism, animal, mammal};

beast 1(n. 野兽)——B_1 = {physical thing, organism, animal, brute};

beast 2(n. 凶残贪婪的人)——B_2 = {physical thing, organism, person, bad person, savage, brute};

利用这几个义项义素集,按上述标准,并通过相应的集合运算进行推理,便可得出正确的词义选择。方法如下。

1) 由于 Wordnet 的义素是按层级编定的,不同义项的顶级义素往往相同,故应删除待排歧词 beast 两个义项义素集中共同的特性,得到仅包含区分特性义素集:

beast 1(n. 野兽)——B'_1 = {animal},

beast 2(n. 凶残贪婪的人)——B'_2 = {person, bad person, savage};

2) 计算 beast 两义项义素集与 panda 义项义素集的交集得:$P \cap B'_1$ = {animal} $\neq \Phi$, $P \cap B'_2 = \Phi$;

3) 由上面的计算,根据义项义素集之交非空的标准可知句中的 beast 应选义项 1(n. 野兽),因此句子应译为"熊猫是野兽",而非"熊猫是凶残贪婪的人",歧义得以排除。

3 词义更广泛的语境相关限制信息

利用从上述词典获得的义项义素集虽然可以处理不少的词义排歧问题,但是由于这样的常识没有反映足够的义项语境相关信息,尚不能做到将所有词的各义项完全区分开来,因此有时仍会导致排歧失败。

如判定"Leaves fall in autumn."中 fall 的词义时,仍可以从 Wordnet 词典获得如下的义项义素集:

leaf(n. 树叶)——L = {physical thing, natural object, plant part};

fall 1(v. 落下)——F_1 = {change location, move, go, travel};

fall 2(v. 进入,陷入,掉进)——F_2 = {change state, turn};

fall 3(n. 秋天)——F_3 = {abstraction, measure, time, time of year};

autumn(n. 秋天)——A = {abstraction, measure, time, time of year};

计算相邻相关词的义项义素集的交集得:$L \cap F_1 = \Phi$, $L \cap F_2 = \Phi$, $L \cap F_3 = \Phi$, $F_1 \cap A = \Phi$, $F_2 \cap A = \Phi$, $F_3 \cap A \neq \Phi$,因而得出 fall 的词义应选"秋天",而非正确的词义"落下"。

分析一下,可以发现,因为不会有人因"树叶"可以"动"而将 move、go 一类的次要特性收入 leaf 的义素集,所以词义的逻辑特性中"树叶"与"落下"这类相去甚远的词语一般不大可能收入相同(通)的义素。因而需要在语言常识中纳入更多词义的语境相关信息,以提高排歧能力。为此,引入义项多元组 $S_i(c_{i1}, c_{i2}, \cdots, c_{ij}, \cdots, c_{in})$来反映词义的语境相关限制信息,其中 $i \geq 1$, $j \geq 1$, S_i 表示某一词语的不同义项,c_{ij}表示词语选择该义项时所要求的各语境相关元(即 Fillmore 格语法中的"格")应有的义素集[6];同时用 D_j 表示实际文本中待排歧词要求的各语境相关元对应位置上实际出现的词语的义项义素集。当实际文本中各语境相关元对应位置上词语的义项义素集与待排歧词在词典中的某义项多元组中要求的各语境相关元应有的义素集匹配时,该义项就是待排歧词在实际文本中应选的词义。即当对任意 j 有 $c_{ij} \cap D_j \neq \Phi$ 时,义项 S_i 为文本中待排歧词应选词义。

如上例,通过"格"语法的分析,并对 fall 的 3 个义项 fall 1(v. 落下)、fall 2(v. 进入,陷入,掉进)、fall 3(n. 秋天)不同"格"位置的语境相关信息进行统计,可得以下义项多元组:

fall 1(v. 落下): S_1 = (Ag, Av) = ({physical thing}, {time ∪ place ∪ manner})其中 Ag 为施事格,Av 为状语结构,{time ∪ place ∪ manner} 为状语结构的中心实词应有的义素集;

fall 2(v. 进入,陷入,掉进): S_2 = (Ag, Av) =

({physical thing}, {feeling ∪ state ∪ device}) 其中 Ag 为施事格，Av 为状语结构；

fall 3(n. 秋天)：S_3 = (Attr, Attr) = ({time ∪ colour}, {time ∪ place}) 其中 Attr 为修饰结构。

在翻译排歧时经过语法分析模块的分析可得到"Leaves fall in autumn."中 fall 的语境相关修饰关系可能为 R_1 = (Ag, Av) 或 R_2 = (Attr, Attr)，其中 Ag 为施事格，Av 为状语结构，Attr 为修饰结构。若句中 fall 的语境相关修饰关系为 R_1 则只能有义项 fall 1(v. 落下)、fall 2(v. 进入，陷入，掉进)对应的义项多元组与之对应，为 R_2 则只能有义项 fall 3(n. 秋天)对应的义项多元组与之对应。

再根据 leaf 和 autumn 的义项义素集：

leaf(n. 树叶)——L = {physical thing, natural object, plant part}；

autumn(n. 秋天)——A = {abstraction, measure, time, time of year}；

分别计算它们与 fall 1、fall 2、fall 3 义项多元组中对应语境相关元的义素集的交集。

对 fall 1：

$L \cap \{\text{physical thing}\} = \{\text{physical thing}\} \neq \Phi$,

$A \cap \{\text{time} \cup \text{place} \cup \text{manner}\} = \{\text{time}\} \neq \Phi$;

对 fall 2：

$L \cap \{\text{physical thing}\} = \{\text{physical thing}\} \neq \Phi$,

$A \cap \{\text{feeling} \cup \text{state} \cup \text{device}\} = \Phi$;

对 fall 3：

$L \cap \{\text{time} \cup \text{colour}\} = \Phi$,

$A \cap \{\text{time} \cup \text{place}\} \neq \Phi$。

根据交集非空的标准可以得出正确义项 fall 1(v. 落下)为句中 fall 应选的义项。由此可见利用义项多元组 $S_i(c_{i1}, c_{i2}, \cdots c_{ij}, \cdots, c_{in})$ 引入词义的语境相关限制信息的排歧方法确实较只使用义项义素集更为有效，能提高排歧能力。

4 结　语

将翻译中的逻辑推理归结为构造一种含排歧义项义素集与义项多元组的语言常识词典，并利用这些常识根据义素集交集非空的原则进行集合运算的过程的确有助于处理机器翻译中的词义排歧问题。其中义项义素集、义项多元组等知识的获取，则可通过大规模语料统计，粗糙集信息提取，并与人工分析相结合的方法来处理。根据在一次英汉机器翻译试验系统中的初步使用，提出的常识概念和排歧方法尚可推广应用于切词、词类兼类判定及句法结构排歧等其它问题的处理。

参考文献：

[1] 冯志伟. 自然语言的计算机处理[M]. 上海：上海教育出版社，1996.

[2] 姚天顺. 自然语言理解——一种让机器懂得人类语言的研究[M]. 北京：清华大学出版社，2002.

[3] 杨尔弘，郝秀兰，李盛. 基于粗糙集的汉语词语义项知识的获取[J]. 中文信息学报，2002,16(3):27-33.

[4] 罗素. 语言哲学[M]. 北京：商务印书馆，1998.

[5] Cognitive Science Laboratory at Princeton University. WordNet-1.7.1[EB/OL]. http://www.cogsci.princeton.edu/~wn/, 2004-05-10.

[6] CHARLES J FILLMORE. 格辩[A]. 语言学译丛(第2辑)[C]. 北京：社科院出版社，1980.

Word Meaning Disambiguation Using Everyday Language Knowledge in Machine Translation

DUAN Qi-li

(China Telecom Chengdu Branch, Chengdu 610051, China)

Abstract: A new method of polysemant meaning disambiguation is provided. First, a formalized criteria of selecting proper meaning of polysemant according to everyday language knowledge is presented. Then the logical reasoning in human's activities of word disambiguation is transformed into a simple and mechanical set operation which is easy to perform for computer. Basing this, the authors provides a concept ——"multi-place sequence of word meaning item" to improve existing electronic dictionaries by introducing more information relevant to language situation for more effective word disambiguation, and thus clarifies the orientation in getting the knowledge of those dictionaries.

Key words: disambiguation on basis of everday knowledge; semanteme set of word meaning item; multi-place sequence of word meaning item; context-related element

(编辑　张　苹)