

机器翻译中应用广义格框架进行语义消解

李堂秋 倪子伟 骆 阳

(厦门大学计算机科学系, 厦门 361005)

摘要：本文阐述机器翻译中的语义消解问题，着重说明如何应用人工智能的知识表示方法解决机器翻译中的语义表示和歧义消解问题。这种方法是基于 Fillmore 提出的格语法为基础的，为形成一个实用的系统以便应用于机器翻译源语言与目标语言的转换，我们对之进行扩充，形成我们称之为“广义格框架”的语义表示框架 (Extended Case Frame)。通过在 XMMT 英汉机译系统中的应用情况表明，这种方法对于语义的表示和语义的消解都有很好的适应性。

关键词：机器翻译，语义消解，广义格框架，语义特征，语义消解

Applying Extended Case Frame for Semantic Resolution in Machine Translation

Li Tangqiu Ni Ziwei Luo Yang

(Dept. of Computer Science, Xiamen University, Xiamen 361005)

Abstract : The paper discusses the semantic resolution problems in machine translation (MT), emphasis is put on how to apply knowledge representation in AI to resolve semantic representation and ambiguity resolution in MT. The method is based on Case Frame and has been expanded to Extended Case Frame. The method using in XMMT system has shown well adaptability for semantic representation and resolution.

Key words : Machine Translation, Semantic ExResolution, Extended Case Frame, Semantic Feature, Semantic Resolution

一 广义格框架

在自然语言处理中，特别在机器翻译中，人们已从早期采用语法转换的策略转到以语义处理为重点的策略。因此需要一种较强的语义表示机构，在此基础上进行分析与推理，以便生成较高质量

的目标语言。

Fillmore 所提出的格文法 (Case Frame) 给一个句子提供了深层语义上的描述。根据格文法的理论, 一个句子可表示为语态 (Modality) 与命题 (Proposition) 两个部分。Modality 是关于句子时态、情态等信息的集合。而 Proposition 是由一个动词与其相关的多个格组成的。其中动词是命题的中心, 而其他的名词短语则与中心动词形成格关系。

格文法为自然语言的处理提供了一个有效的表达方法, 然而 Fillmore 当初提出的格框架结构是过于简单的模型, 在不同的应用中必须进行相应的扩充和改造。例如机器翻译中我们要处理丰富多采的自然语言, 原有的格体系显然是不够的。为了满足机器翻译这一特定目标, 我们对它进行了必要的扩充和改造, 以发展成为一个实用的系统, 我们把这种表示方式称为广义格框架 (Extended Case Frame, 简称 ECF)。

首先, 我们将 ECF 扩展成为由三个要素组成的: 语态, 命题与生成规则:

$$S \rightarrow M + P + R$$

其中 R 表示生成规则, 它表达了如何根据框架的语态和命题信息, 生成目标语言的线性词序列。R 与动词的类型有关, 所有的动词可以分成一定数目的类别, 如在双解词典中将所有的动词分成二十五型, 每种对应着一个生成规则。有些动词的生成规则还与框架的其他信息结构有关, 必须根据不同的情况加以具体的构造。生成规则用于指导目标语句子的生成。改造以后的 EFC 可表示为:

$$S \rightarrow M + P$$
$$M \rightarrow \text{tense, aspect, } \dots\dots$$
$$P \rightarrow V + C_1 + C_2 + \dots\dots + C_n$$
$$V \rightarrow \text{run, break, put } \dots\dots$$
$$C_i \rightarrow K_i + NP_i$$

其中的 K_i 为格标识, 它与对应的名词短语 NP_i 中的中心词一起构成动词的一个格。

EFC 第二方面的扩充是在格的数目和种类, 根据 Fillmore 当初提出的格框架理论, 格有: 施事格、工具格、受与格、使成格、地点格和客体格等六种。而 EFC 的格数目要多得多, 它的格分为基本格和状语格两种型类。这两种格的作用是相同的, 只是得到的来源不同。基本格是从句子的主要成分, 即名词性短语分析得来的, 它反映句子中的主要成分的语义功能。而状语格则是从状语提取产生的。它反映各状语在句子中的语义功能。

考虑到从句子句法结构分析出格的方便性, 也考虑到语义分析的基本要求。由于所有的动词都可以分成有限个不同的类型, 而同一种动词所带的名词短语在句子中的语义作用是相对固定的。这样我们有可能通过有限个规则从句子句法结构分析出它相应的格。通过分析将基本格定为十二种, 它们分别是:

- (1) agent \rightarrow 施事格, 通常是句中的主语, 语义上是动作的施事;
- (2) theme \rightarrow 主题格, 非施事的主语;
- (3) dative \rightarrow 受与格, 对应于句中的宾语;
- (4) neutral \rightarrow 对应某些动词的间接宾语;
- (5) goal \rightarrow 对应某些动词的间接宾语;
- (6) purpose \rightarrow 目的格, 对应于间接宾语;
- (7) quality \rightarrow 品质格, 对应于宾语补足语;
- (8) state \rightarrow 状态格, 对应于非名词性表语;
- (9) substitute \rightarrow 同一格, 名词性表语;
- (10) value \rightarrow 数量格, 一些动词的补足语, 表示距离, 价值等;

[11] action --->动作格, 引出深层EFC;

[12] event --->事件格, 一些动词的子句;

状语格主要是由句子中的状语结构, 如句子中的副词、介词词组、名词词组、动词词组或从句, 转换而成的, 它们包括:

[1] disjunct --->主从格

[2] conj--adv --->并列格

[3] freq --->频度格

[4] time --->时间格

[5] place --->地点格

[6] field --->领域格

[7] manner --->方式格

[8] form --->形式格

[9] means --->手段格

[10] instrument --->工具格

[11] emotion --->情感格

[12] amount --->数量格

[13] accompaniment --->伴随格

[14] comparision --->比较格;

[15] degree --->程度格;

[16] reason --->原因格;

[17] aim --->目的格;

[18] condition --->条件格;

[19] concession --->让步格;

[20] result --->结果格。

由于广义格框架理论比其他文法理论为句子提供了更多语义表达成分, 因此它可用于表示句子的语义。例如在机器翻译中, 可以用这种语义结构作为源语言与目标语言的转换平台, 即源语言句子通过语法分析与语义分析生成这种框架结构, 然后在此基础上生成目标语言。这种方法的好处是, 它是一种命题结构, 可以进行一定的语义推理, 而这一点应用语法树是做不到的。又由于任何一种语言的特定语义结构有比较固定的语言表达模式, 因此这种结构可更好地用于目标语言的生成。

同时, 广义格框架可以做为表达动词所需的各种格的语义约束机制。在语言知识库中我们可以为每一个动提供一个格框架, 指明它在组成句子, 表达一个概念时可以具有的格, 以及各种格的语义约束, 其中有些格是强制性的, 即必须具有的, 另一些则是可选的。由于这种格框架可以表达动词与有关名词短语的语义搭配关系, 在句子分析中, 应用一定的算法, 可以用于语义的排歧。

二 转换与消解算法

下面给出算法Transfer, 它把从语法分析得到的语法树转换为易于生成中文句子的广义格框架结构。根据格文法的思想, 句子中的动词提供了句子的主要信息, 所以从句法结构分析出广义格框架的过程, 首先应找出句子的主动词, 取出该动词的语义单元中的 <语义约束>框架, 然后分析其他名词性短语成份。

Transfer:

对于句子的语法树T, 依次执行如下步骤:

1. 取出T中动词成份, 记下其中包含的语态信息、动词在句中的位置和句型。取出语言知识库中该动词语义单元中的<语义约束>格框架。
2. 依次取出T中的每一个成份, 执行如下步骤, 直至T中所有成份均处理完毕:
 - 2.1. 若该成份的内部结构包含动词, 则对动词及其连带结构所形成的子树执行transfer过程, 返回生成的EFC。
 - 2.2. 根据动词语义单元中的格框架, 将指定的成份转换为指定的格。
 - 2.3. 对其余成份, 作如下处理:
 - 2.3.1. 若为动词的句型所指定的名词短语成份, 则根据动词的类型和句子的基本成份与基本格的关系, 把该成份转换为对应的基本格。
 - 2.3.2. 若为句子的状语成份, 则:
 - 2.3.2.1. 若该成份是从句, 则根据关联词及与从句的关系, 把该成份转换为适当的状语格。
 - 2.3.2.2. 若是介词短语, 则分析介词所搭配的名词词组的中心词, 根据其词义单元的特征和语言知识库中该介词语境限制, 将之转换为适当的状语格。
 - 2.3.2.3. 若是副词词组, 则抽取该副词的词义单元中的语义限制, 并根据该副词在句中的位置, 将之转换成适当的状语格。
3. 返回生成的EFC。

为了体会这些格的作用及一个句子格框架的抽取, 这里给出两个例句的广义格框架, 例中忽略了一些次要信息, 如句子的语态和转换规则等, 仅给出EFC的命题部分:

[1] I told the servant to open the window.

```
ECF: {(v v3 "tell")
      agent (np (pron "i"))
      dative (np (det "the")(n "servant"))
      action {(v v1 "open")
              agent nil
              dative (np (det "the")(n "window"))
              }
      }
```

[2] They had come a long way.

```
EFC: {(v v20 "come")
      agent (np (pron "they"))
      degree (np (det "a")(adjp (adj "long"))(n "way"))
      }
```


四 结束语

格框架作为句子的一种表示方法比其他基于句法结构的表示方法具有较大的优越性。但是 Fillmore 格框架的原始形式要做为一个实用系统是过于简单的,它必须针对不同的应用领域加以适当的改造。本文讨论一种格框架表示方法的扩充形式——广义的格框架,它的结构设计既考虑了从语法树获得 EFC 的方便性,又考虑了语义推理的适当深度要求,在 XMMT 英汉机译系统中它既作为句子语义的内部表示方法,也作为语言知识库的重要的知识表示框架。这种方法具有如下的优点:第一,具有一定的语义表示深度,有利于语法和语义的排歧和推理;第二,它突出概念之间的联系,而不是句子各成分之间的线性顺序或句法的结构,从而使目标语言的生成可以直接从这种结构中导出,而不受源语言线性顺序的限制,有利于生成高质量的译文。这个模型在 XMMT 英汉机译系统中应用取得良好的结果。

参考文献

- [1] Charles j. Fillmore, The case for case, 1986
- [2] Lexical Ambiguity Resolution, Morgan Kaufmann Publishers, Inc.
- [3] Garrison W. Cottrel, A Modal of Lexical Access of Ambiguous Words.
- [4] 俞士汶, 自然语言的歧义与机器翻译的对策, 中文信息学报, 1989.2
- [5] 黄昌宁, 机器翻译与新的语法理论, 中国计算机用户, 1989.9