文章编号：1003-0077(2018)12-0057-10

航空术语语义知识库辅助构建方法

王思博，王裴岩，张桂平

（沈阳航空航天大学 人机智能研究中心，辽宁 沈阳 110136）

摘 要：语义知识库是自然语言处理任务的基础性资源，广泛应用于语义计算和语义推理等任务。现有的大规模语义知识库基本都是通用型知识库，缺乏特定领域的语义知识。为了弥补这种不足，该文基于 HowNet 的语义理论体系，提出了一种辅助构建航空术语语义知识库的方法。该方法根据航空术语的特点将辅助构建分成四个关键过程，构建了 2 000 条术语概念描述(DEF)。最后通过对人工标注的术语间相似度与根据术语 DEF 计算的术语间相似度结果的对比，验证了该构建方法的有效性。

关键词：航空术语；语义知识库；知网；概念描述

中图分类号：TP391 **文献标识码**：A

A Semi-automatic Construction Method of Semantic Knowledge Base of Aviation Terms

Wang Sibo, Wang Peiyan, Zhang Guiping

(Human-computer Intelligence Research Center, Shenyang Aerospace University, Shenyang, Liaoning 110136, China)

Abstract: Semantic knowledge base is a basic resource of natural language processing. The existing large-scale semantic knowledge base is basically generic knowledge base, lacking the domain specific semantic knowledge. This paper proposes a semi-automatic method of constructing the semantic knowledge base of aviation terms by HowNet. It consists of four key processes of construction, resulting altogether 2 000 descriptions of the term concept (DEF). Finally, the validity of the method is verified by comparing the term similarities obtained by manual annotation and those obtained according to the term DEF.

Keywords: aviation terms; semantic knowledge base; HowNet; DEF

0 引言

语义知识库是一种重要的基础性语言资源，可以为自然语言处理任务提供丰富的语义知识，常被广泛应用于词义消歧、机器翻译、信息检索以及自动问答等任务。目前，国内外研究者已经构建了许多大规模语义知识库。其中，国外被广泛应用的语义知识库主要有 WordNet、FrameNet、MindNet、OpenCYC 等。国内较为成熟的语义知识库有 HowNet(知网)[1]、CCD(the Chinese Concept Dictionary，中文概念辞书)[2]、CFN(Chinese FrameNet)[3]、《现代汉语述语动词机器词典》[4]等。这些语义知识库大多都面向通用领域，但在特定领域下则无法满足自然语言处理任务对语义知识的需求。而垂直领域下的语义知识库可以填补通用型知识库的不足。

本文以 HowNet 为基础，按照 HowNet 的 KDML 语法体系、义原体系与动态角色/特征体系构建航空术语语义知识库。因此，该语义知识库继承了 HowNet 全部特点与优势，便于计算机使用[5]，能够作为语义信息加入系统中，支撑面向航空领域文本理解任务的相似度计算、相关度计算等语义分析。

1 相关研究

HowNet 是一个以汉语和英语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

收稿日期：2018-01-30 **定稿日期**：2018-03-13

基金项目：教育部人文社会青年科学研究基金(17YJC740087)

库[1]，是公认的规模最大、收录词语最多、提供信息最多的语义词典。KDML（knowledge database mark-up language，知识系统描述语言）[6]是HowNet所使用的一种知识描述语言，具有明确的语言语法规则，规范了概念描述方式。最关键的一点是KDML是面向计算机的形式化描述方法，便于进行相似度、相关度和情感倾向性计算等。正如文献[7]所指出的："知网的知识表达模式是针对计算机的信息处理特点而制定的。"此外，HowNet秉承还原论思想，认为词语可以用更小的语义单元来描述。这种语义单元被称为义原（Sememe），即最基本、不宜再分割的最小语义单元，并构成了一套义原体系。

文献[8]和文献[9]先后进行了面向航空领域的术语语义知识库构建的相关研究。它们都基于HowNet的义原体系、动态角色/特征体系以及KDML语法理论，一定程度上了扩大了HowNet的覆盖范围。文献[8]根据HowNet的7条总规定延伸出针对航空术语知识库构建的5条基础规则，主要包括义原和动态角色/特征的使用规则与规范，对接下来的研究起到一定的指导作用。然而，根据文献[7]所提出的知识库构建规则，若仅凭手工构建，则需要巨大的时间和人力成本。为了提高构建效率，文献[9]在文献[8]的基础上提出了一种基于核心词框架的知识库构建方法，即利用统计与规则相结合的方法对核心词框架进行获取与补充，相比于手工构建大幅提高了构建效率，一定程度上实现了半自动化构建。但这种基于核心词框架的构建方法固定了术语核心词与术语内部其它词语之间的语义关系，忽略了术语非核心词语之间的语义关系。

本文考虑到术语内部词语之间具有一定的依存结构，并利用这种依存结构信息进行词义消歧和术语DEF的生成。同时，本文也提出了一种术语内部动态角色关系辅助判断方法，明确了术语内部核心词与非核心词之间以及非核心词语之间的语义关系。这使得术语DEF能够更充分地表示术语内部词语之间的语义关系，进一步提高航空术语语义知识库构建的自动化程度。

2 构建框架

针对术语DEF构建任务的特点，本文将整个构建任务分成四个关键过程，分别为术语内部依存结构分析、术语内部词语义项辅助选择、术语内部动态角色关系辅助判断以及术语DEF生成。其整体框架如图1所示。

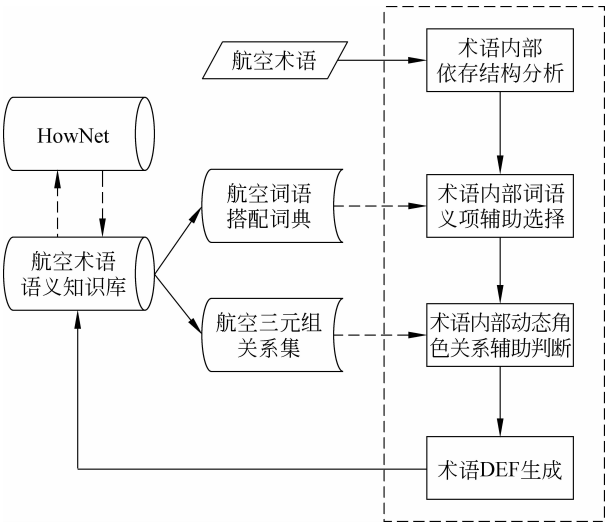

图1 航空术语语义知识库构建框架图

（1）术语内部依存结构分析

本文对术语内部的依存结构分析，考虑更多的是术语内部概念间语义层面上的依存关系。因此，可以通过术语内部依存结构分析，确定术语内部词语间的依存结构，从而得到具有语义依存关系的关联词对。本文将此依存关系表示为三元组，其中包括关联单位、关系方向以及关联类型（关联单位是具有依存关系的词对；关系方向是依存与被依存的方向；关系类型被表示为HowNet的动态角色/特征）。

（2）术语内部词语义项辅助选择

由于多义词所处的上下文一定程度上决定着该词语义项的选择。因此，本文提出了一种基于依存结构的词义消歧方法，它将术语内部的关联单位视为词语组合上具有相互搭配关系的词对，并根据这种词语间搭配的同现关系进行词义消歧。

例如，航空术语"空气循环冷却系统"中的"空气"一词，在HowNet中对应如下两个义项：

① DEF＝{gas|气：{contain|含：OfPart＝{～}}，{inhale|吸入：agent＝{AnimalHuman|动物}，patient＝{～}}}

② DEF＝{Occasion|场面：host＝{group|群体}{place|地方}}

第一个义项的第一义原是"gas|气"，它所描述的"空气"是一种物质，即气；第二个义项的第一义原是"Occasion|场面"，所要描述的是一种场面。因为当前术语中的"空气"是"循环"的对象，表示"气"的"空气"与"循环"的同现更易存在，所以选择"空气"的第一个义项更符合当前术语内部的语义环境，以

使这里的“空气”语义表示得更准确，进而将“gas|气”作为“空气”的 DEF。

(3) 术语内部动态角色关系辅助判断

针对术语内部关系类型的表示问题，本文提出一种术语内部动态角色关系辅助判断方法。HowNet 应用动态角色/特征来标注概念间的语义关系，每种动态角色/特征关联着无计其数个具有语义关系的关联词对，其中 HowNet 包含 100 种不同的动态角色/特征，面向通用领域涵盖了较为全面的语义关系类型，反映了丰富的语言现象。在航空术语中常用到的动态角色/特征大约有 20 几种，这些动态角色/特征表示了航空领域语义空间所出现的各种语义关系。

通过(1)(2)两步以及本过程(3)，可以完成术语“空气循环冷却系统”DEF 的结构分析(如图 2 所示)，得到如下 3 个三元组，(空气，patient，循环)、(循环，means，冷却)、(冷却，instrument，系统)，以及术语内部词语 DEF。对于三元组“(空气，patient，循环)”可做如下解释：三元组的关联单位是“空气”和“循环”，它们的关系方向是“空气”依存于“循环”，其中的关系类型是“patient”。

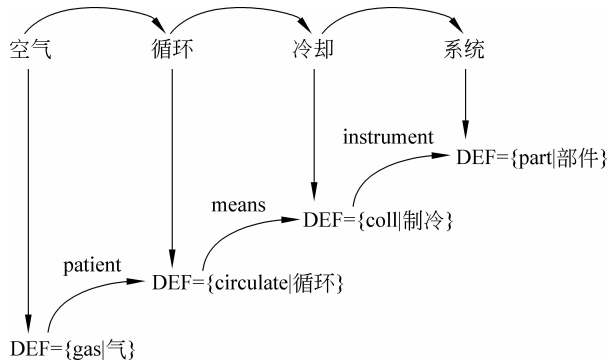

图 2　术语“空气循环冷却系统”DEF 结构分析

(4) 术语 DEF 生成

本文基于术语内部依存结构提出一种术语 DEF 生成算法，基于术语内部依存结构，将术语内部所有关系三元组映射成术语 DEF，提高了航空术语语义知识库的构建效率。

3　构建方法

本节将详细阐述术语语义知识库的构建方法。

3.1　术语内部依存结构分析

对术语内部依存结构的分析是本文构建方法的基础，下文将进一步介绍本文所指的术语内部依存结构以及本文所采用的术语内部依存结构分析方法。

3.1.1　术语内部依存结构

一般认为，依存语法的理论研究始于法国语言学家特思尼耶尔(Lucien Tesnière)。他提出了依存语法的基本论点[10]，即利用词语之间的依存关系刻画文本的语法结构。依存语法提出至今，一直深远地影响着语言学的发展。本文的研究对象是术语，术语内部依存结构是描述术语内部词法结构的，也就是通过词语间的“依存”使得术语内部的词语关联起来。

3.1.2　术语内部依存结构分析方法

本文术语内部依存结构分析借鉴了文献[11]对术语的依存分析模型。它利用了模型选择策略为结构化风险最小的 SVM，在训练语料不十分充分的情况下模型依然能够取得不错的效果。

在特征选择上，选取了术语基本特征、术语内部任一词对之间的点互信息以及术语内部词语的 HowNet 义项的第一义原。模型根据词对的依存强度输出依存强度值，若为正值说明预判断的词对存在依存关系，当输出值越大则说明词对的依存强度越大；若输出值为负值说明词对不能构成依存关系，当值越小则说明词对越不可能存在依存关系。其中术语基本特征包括术语内部词、词性、词对之间的距离以及术语内部上下文窗口为 1 的词。点互信息度量的是变量间相互依赖的程度，在该模型中度量了术语内部词间的依赖度。术语内部词在 HowNet 中的第一义原作为特征的加入，有效减缓了数据稀疏的问题。由于一条术语不仅是一个由词语组成的序列，更是一个由语义依存关系连接而成的树。因此对术语进行依存结构分析可得到术语内部所有的关联单位。

3.2　术语内部词语义项辅助选择

术语内部词语义项选择是为术语内部词语选择合适的概念描述(DEF)，即采用 HowNet 的最小语义单元(义原)来表示。此辅助选择过程将本文所提出的基于搭配词的词义消歧方法融入到词义选择的任务中来。以下两小节将详细介绍本文所指的搭配词、搭配词集、搭配词典的相关概念，以及本文所提出的词义消歧方法。

3.2.1　搭配词、搭配词集以及搭配词典

(1) 搭配词

所谓搭配词，是指与多义词同在一个关联单位

的词语，这些词语与多义词在语义层面上具有依存关系。在领域语义空间中多义词与其搭配词共现，对应多义词确定的某一义项。

（2）搭配词集

顾名思义，搭配词集是由多义词的搭配词构成的集合。多义词在领域语义空间中所出现的每个义项对应一个词语集合，即该多义词的搭配子集，也意味着当前多义词的搭配子集对应多义词的某一义项，多义词的搭配子集构成了搭配词集。

（3）搭配词典

搭配词典是由语料库中所有多义词、搭配词集以及多义词各个义项构成的集合。

3.2.2 基于搭配词的词义消歧

基于搭配词的词义消歧是根据多义词的搭配词所属搭配子集选取该多义词的义项。术语内部的多义词在特定的语义约束下其表示的语义相对稳定，符合术语单义性[12]的特点。本文从已有的航空术语语义知识库中分析并抽取出航空术语内部词语的搭配词典。由于搭配词与多义词的共现对应多义词确定的某一义项，只需判断在搭配词典中多义词的搭配词收录于哪个搭配子集，搭配子集所对应的义项即为该多义词在当前术语中表示的义项。若当前搭配词不在当前多义词的搭配词集里，则将搭配词与多义词的搭配词集的每个词语进行相似度计算，取与搭配词最相似的词语所属搭配子集的对应义项作为该多义词义项。

综上所述，本文将基于搭配词的词义消歧方法融入术语内部词义辅助选择的任务中。术语内部词语义项辅助选择方法的具体算法过程如下所示：

词义辅助选择算法

输入：待确定义项的词语 w 及其搭配词 c

输出：当前词语 w 的义项 S

1. $S=\varphi$，配置资源初始化
2. Case1：$w\notin \mathrm{Dict}_{\mathrm{HowNet}}$
   S←人工标注
3. Case2：$w\in \mathrm{Dict}_{\mathrm{HowNet}}$ AND $w\notin \mathrm{Dict}_{\mathrm{ambig}}$
   S←$\mathrm{Dict}_{\mathrm{HowNet}}[w][0]$
4. Case3：$w\in \mathrm{Dict}_{\mathrm{ambig}}$ AND $w\notin \mathrm{Dict}_{\mathrm{match}}$
   S←人工从 $\mathrm{Dict}_{\mathrm{ambig}}[w]$选择
5. Case4：$w\in \mathrm{Dict}_{\mathrm{ambig}}\cap \mathrm{Dict}_{\mathrm{match}}$
   if $c\in \mathrm{Dict}_{\mathrm{match}}$
   then S←c 所属的 subsetmatch 对应的义项
   else 计算得到在 w 的搭配词集 $\mathrm{set}_{\mathrm{match}}$ 中与 c 语义最相似的 c'；
   S←c' 所属的 subsetmatch 对应的义项
   Endif
6. EndCase
7. 返回词语 w 的义项 S

3.3 术语内部动态角色关系辅助判断

本文基于 HowNet 将动态角色/特征应用到术语概念的描述中，使得术语内部的简单概念通过动态角色有机关联起来，从而构成表示术语本身语义知识的复杂概念。其中，对于术语内部的词语 w_1 和词语 w_2 之间存在语义关系，可以表示为某种动态角色/特征，用三元组的形式表示：

$$(w_1,\mathrm{EventRole/EventFeature},w_2)$$

其中关联单位是 Relation（w_1，w_2），关系类型为 EventRole/EventFeature，关联方向为 w_1 依存于 w_2。术语内部所有的三元组表示了术语内部词语结构。

本过程采用最大熵分类器与基于相似度的动态角色判断方法相结合的方法辅助推荐动态角色，以人工选择标注三元组的关系类型。

3.3.1 基于最大熵分类器的动态角色判断

本方法将动态角色关系判断转化成一种对于关联单位的分类问题，并且将关联单位所对应的动态角色/特征作为分类的标签。其中，最大熵分类器以最大熵模型为理论基础，其基本思想是将所有满足已知约束条件的概率模型中熵最大的模型视为最好的分类模型[13]。最大熵分类器能够较容易地对多分类问题进行建模，并对各个类别输出一个相对客观的概率值[14]。与此同时，最大熵的训练效率相对较高，相比于 SVM，最大熵模型可以较容易地对多分类任务建模。其中最大熵分类器选择以上两过程获得的结果作为特征，如表 1 所示。

表 1 最大熵分类器所选用的特征

特征	说明
关联词对	(w_1,w_2)；w_1 依存于 w_2
关联词对义项第一义原	S_1，S_2；S_1 为 w_1 义项首义原；S_2 为 w_2 义项首义原
关联词义项第一义原类别	包括事件类/实体类/属性类/属性值

最大熵分类器为每种动态角色给出概率值。因此，本方法基于概率值对候选动态角色排序，得到概率值从大到小的动态角色排序表，并从此排序表中选择排序最高位的动态角色。

3.3.2　基于相似度的动态角色判断

基于相似度的动态角色判断方法是将待判断关系类型的关联单位与训练集中每一个三元组的关联单位进行相似度计算，并将此相似度值作为三元组的分值，从而在训练集中出现的每个动态角色都对应一个分值列表，如下所示：

$$\begin{cases} \text{EventRole}_1:[\text{score}_{11},\text{score}_{12},\cdots] \\ \text{EventRole}_2:[\text{score}_{21},\text{score}_{22},\cdots] \\ \vdots \\ \text{EventRole}_n:[\text{score}_{n1},\text{score}_{n2},\cdots] \end{cases}$$

其中“EventRole_1、EventRole_2、...、EventRole_n”为表示三元组关系类型的动态角色；“$[\text{score}_{11},\text{score}_{12}\cdots]$”为动态角色“$\text{EventRole}_1$”的分值列表，“$\text{score}_{11}$，$\text{score}_{12}$”是关系类型为“$\text{EventRole}_1$”的三元组的分值。

本方法取动态角色分值列表的最大值作为候选动态角色的分值。根据分值从大到小对动态角色从高到低排序，从而得到动态角色排序表。按照预先设定的优先级从排序表中选取未在答案集中排序最高的动态角色。其中待判断关系类型的关联单位 $U_1(w_{11},w_{12})$ 与训练集中三元组的关联单位 $U_2(w_{21},w_{22})$ 间的相似度计算如式(1)所示，$\text{Sim}_w(w_1,w_2)$ 详见文献[15]，此处不再赘述。

$$\text{Sim}_u(U_1,U_2)=\text{Sim}_w(w_{11},w_{21})*\text{Sim}_w(w_{12},w_{22}) \tag{1}$$

3.3.3　最大熵分类器与基于相似度方法相结合

最大熵分类器利用使概率模型的条件熵趋于最大值的统计信息，给待判断关系类型的关联单位的可能动态角色关系打分；而基于相似度的方法，则利用词语的语义信息，通过度量待判断关系类型的关联单位与在训练集中关联单位之间的相似度，为动态角色打分。二者分别从统计和语义两个不同层面进行动态角色判断，存在一定的互补。

因此，本文采用最大熵分类器与基于相似度方法相结合的动态角色判断方法，从两者生成的动态角色排序表中按照预先设定的推荐优先级顺序依次向答案集添加动态角色，以供人工选择。并在实验中证实了本方法的可行性，详见第 4.2 节。

3.4　术语 DEF 生成

术语 DEF 生成是本文方法的最后一个过程，它根据 KDML 语法规则将以上三个过程分析得到的语义信息表示成 HowNet 的语义知识。以下两小节将详细介绍 KDML 的规定和本文所提出的术语 DEF 生成算法。

3.4.1　KDML 规定

本文对航空术语语义知识的描述遵从 KDML 的规定，一定程度上保障了语义知识描述的复杂度、一致性以及准确性。按照 KDML 的描述概念的主要规定[6]：

① 任一概念的描述都以“DEF＝”为开始。任一概念中出现的所有义原或符号必须是在 HowNet 的 Taxonomy 中定义的义原或符号或者由知网知识系统描述语言所规定的特定标识符。

② HowNet 概念描述的第一个义原必须指出该概念最基本的意义，并用事件、实体、属性和属性值这四类义原中的一个标注出来。

③ HowNet 利用动态角色/特征标注复杂概念，表示简单概念之间的语义关系。

例如，本文所构建的航空术语“空气循环冷却系统”DEF 表示为：

DEF＝{part|部件：{cool|制冷：means＝{circulate|循环：patient＝{gas|气}}，instrument＝{～}}}

它的第一义原是“part|部件”，是一个实体类概念，对应术语核心词“系统”，反映了该术语最基本的意义。术语 DEF 中出现了“means”、“patient”和“instrument”三种动态角色。“patient”说明了空气(gas|气)是循环(circulate|循环)的对象(patient)；“means”说明了系统冷却的方式(means)，即空气循环；“instrument”说明了“系统”这个部件的功能，即冷却(cool|制冷)的工具(instrument)。其中“～”特殊指示符代替了前一层的义原“part|部件”。

HowNet 的 KDML 对概念的描述是有一定结构的。按照 KDML 的规定，常用特定标识符如下所述：

① 左括号“{”表示一个概念描述的开始；

② 右括号“}”表示一个概念描述的结束；

③ 冒号“：”后面的内容是对冒号前面义原的具体描述；

④ 逗号“，”表示一个关系描述的结束；

⑤ 等号“＝”表示一个动态角色/特征所具有的值。

因此，从 HowNet 特定标识符标注的角度来看，HowNet 复杂概念的描述是通过大括号之间的嵌套与冒号、等号等特殊标识符的标注来表示的。因此，本过程将术语内部的三元组按照 KDML 对 HowNet 概念描述的规定解析成术语的 DEF。

3.4.2 术语DEF生成算法

本文基于术语内部依存结构提出一种术语DEF生成算法，按照术语内部依存结构，将术语内部所有关系三元组映射成术语DEF。例如，术语“$w_1w_2w_3w_4w_5$”生成DEF过程如图3所示，通过前3个过程，得到了所有完整的三元组，包括：(w_1，EventRole_{13}，w_3)，(w_2，EventRole_{23}，w_3)，(w_3，EventRole_{35}，w_5)，(w_4，EventRole_{45}，w_5)；以及术语内部词语DEF，表示为{w_1：DEF=$\{S_1\}$，w_2：DEF=$\{S_2\}$，w_3：DEF=$\{S_3\}$，w_4：DEF=$\{S_4\}$，w_5：DEF=$\{S_5\}$}。

本方法将术语内部依存结构表示成依存树的形式，如下所示：{w_5：[w_3，w_4]，w_3：[w_1，w_2]，w_4：[]，w_2：[]，w_1：[]}。其中当前术语的核心词是w_5，位于依存树叶子节点的词语为w_4、w_2以及w_1。

图3表示了术语依存树转换成术语DEF的映射过程，按照大箭头的指示依次变换。示意图中的起始框图表示了术语“$w_1w_2w_3w_4w_5$”依存树结构。其中依存树节点之间的实线边表示依存关系，由被依存对象指向依存对象；边上符号表示节点之间的动态角色关系，每个节点存储当前词语的DEF。可以看出，随着依存树的叶子节点向其父节点嵌入语义信息的过程演进，依存树的结构以及树节点信息也随之变化。其中节点之间的虚线表示将依存对象(子节点)的DEF以及两者之间的动态角色按照KDML的规定嵌入到被依存对象(父节点)；叶子节点完成嵌入语义信息后，被剪枝；依存树重复上一过程，每一次都是由当前依存树的叶子节点向其父节点嵌入语义信息，直至仅剩下根节点；当只剩下根节点时，术语依存树完成转换术语DEF的映射过程，输出根节点信息即可得到术语DEF。

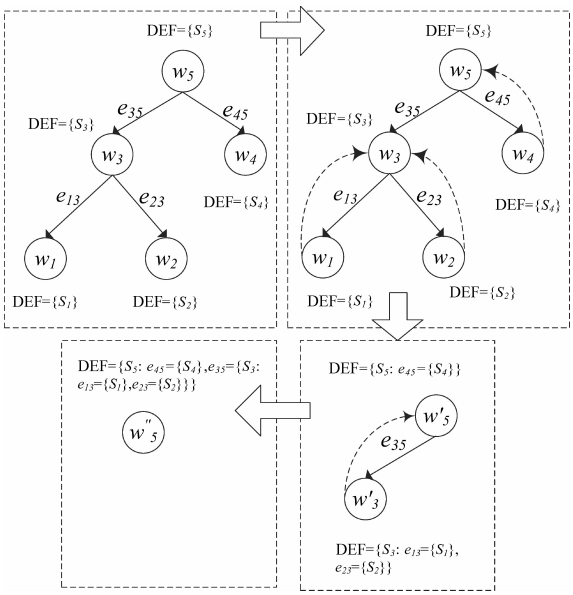

图3 术语DEF生成示意图

术语DEF生成算法如下所述。

术语DEF生成算法

输入：术语内部所有完整的三元组以及每个词语DEF
输出：术语DEF
1. 将三元组列表解析成依存树
2. 遍历依存树，找到当前依存树的叶子节点
3. 判断当前叶子节点是否为依存树根节点。若为是，进入步骤5；否则进入步骤4。
4. 将该叶子节点的DEF及该叶子节点和父节点之间的动态角色，以KDML的规定嵌入到父节点的DEF中，删除当前叶子节点，进入步骤2
5. 输出依存树根节点信息，即为术语DEF

4 实验结果与分析

本文对术语内部词语义项辅助选择、术语内部动态角色关系辅助判断分别进行了实验和实验结果分析；并通过相关性实验，验证了本文构建术语DEF方法的有效性。

4.1 术语内部词语义项辅助选择实验

本实验对1 000条术语进行人工词语义项标注，选取HowNet中最符合术语概念的义项，标注内容为已选义项的第一义原，将此作为实验语料。该实验语料的词表中一共有996个词语，其中的268个词语在HowNet中是多义词。这些多义词在搭配词典中大多只有一个义项，也有一些多义词只有部分义项出现在搭配词典里。对于那些不在搭配词典中的多义词，本实验无法给出该多义词义项的选择结果，记为选择错误。

本文将实验语料分成10等份，每份100条术语，进行10-fold交叉验证。采用平均准确率P作为评价指标，其中P的计算公式如式(1)所示，n为测试的次数。

$$P=\frac{\sum_{i}^{n}p_i}{n} \tag{1}$$

$$p_i=\frac{\text{判断正确的待测对象个数}}{\text{测试集/开发集中待测对象个数}} \tag{2}$$

通过10-fold交叉验证所得到的平均正确率，为90.68%，其中不在HowNet中的词语以及不在搭配词典中的多义词平均占测试集词语的7%。剩下

的接近 3%是由于本方法处理错误造成的。因此，为了使得知识库的语义标注结果更准确，对未在搭配词典里的多义词与未在 HowNet 中的词语进行人工义项标注。

4.2　术语内部动态角色关系辅助判断实验

本实验从人工标注的航空术语语义知识库[14]中抽取 475 条航空术语 DEF。人工将每条术语 DEF 分解成若干个三元组以及术语内部词语 DEF，一共有 1 550 个三元组(也意味着本实验数据集包含 1 550 个样本)，一共出现 27 种动态角色，其分布情况如图 4 所示。将 1 550 个样本平均分成 10 等份，进行 10-fold 交叉验证。

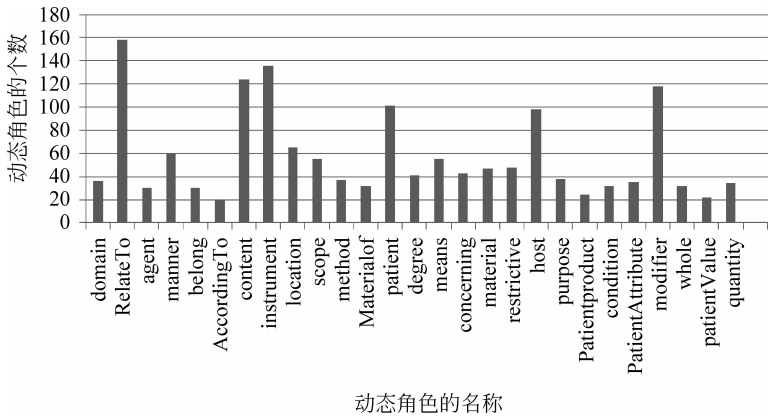

图 4　数据集中动态角色分布的情况

本次实验通过 10-fold 交叉验证，对最大熵分类器和基于相似度方法进行了对比实验。实验结果如表 2 所示，显示了每种方法 3-best(最有可能为正确答案的 3 个动态角色)中 Top1、Top2 以及 Top3 对应的三个不同排序位置上动态角色的平均准确率 P[见式(2)]，以及最大熵分类器和基于相似度方法推荐各自动态角色排序表中 3-best 的平均准确率 P(当待测三元组的正确动态角色出现在候选答案集(3-best)中时即为判断正确)。

表 2　两种方法的实验结果

类别	最大熵分类器/%	基于相似度方法/%
Top1	54.2	28.9
Top2	15.3	9.5
Top3	8.3	2.1
Total	**77.8**	**40.5**

从实验结果可以看出，当最大熵分类器和基于相似度方法分别从各自动态角色排序表中向答案集推荐 3-best 时，最大熵分类器所得到的准确率较高，而且它的动态角色排序表中排在前三位的每个位置上准确率均高于基于相似度方法。

通过对实验输出结果的统计，能够得到两种动态角色排序表 Top1～Top3 不同位置之间，正确动态角色的重复率。其中，两排序表中 Top1 上同为正确动态角色的重复率是 6.8%，基于相似度方法的排序表 Top1 与最大熵分类器的排序表 Top2 间正确动态角色重复率为 3.3%，可见当同时推荐最大熵分类器的排序表的动态角色和基于相似度方法的排序表的动态角色组成 3-best 时，能够得到更好的实验结果，因此从两种关系判断方法所得到的 3-best 动态角色中按照一定的优先级顺序选择动态角色组成 3-best 结果如表 3 所示。

由于无论是横向逐层(Top1～Top3 的顺序)依次从两排序表中按照不同优先顺序选择动态角色；还是纵向以不同的优先级顺序从两排序表中选择动态角色，最终都是要将如下两种情况与最大熵分类器推荐的 3-best 以及基于相似度方法推荐的 3-best 进行实验对比。这两种情况分别是，情况①：在最大熵分类器的动态角色排序表中选择 2-best(最有可能为正确答案的 2 个动态角色)以及在相似度方法的动态角色排序表 Top1～Top3 中选择一个未被选中(不重复)的动态角色，组成 3-best；情况②在基于相似度方法的动态角色排序中选择 2-best 以及在最大熵分类器的动态角色排序表 Top1～Top3 中选择一个未被选中的动态角色，组成 3-best。

因此本实验所按照推荐优先级顺序，分别为 S1->S2->M1->M2->M3、S1->S2->M1->M2->S3、M1->M2->S1->S2->S3、M1->M2->S1->S2->M3。这四种优先级顺序覆盖了上文所述的两种情况

(包括两排序表 Top1～Top2 组成的 2-best 集合相同的情况),并将其与"S1->S2->S3"以及"M1->M2->M3"进行对比实验。其中当待测三元组的正确动态角色出现在候选答案集(3-best)中时即为判断正确。

表 3 最大熵分类器与基于相似度方法相结合的实验结果

推荐优先级顺序	P/%
M1->M2->M3	77.8
S1->S2->S3	40.5
S1->S2->M1->M2->S3	81.5
S1->S2->M1->M2->M3	82.5
M1->M2->S1->S2->S3	85.5
M1->M2->S1->S2->M3	**88.3**

从表 3 的实验结果可以看出,当推荐次序为 M1->M2->S1->S2->M3 时,答案集出现正确动态角色的准确率最高。M1->M2->S1->S2->M3 优先级顺序使得最大熵分类器的 2-best 能够优先加入答案集、基于相似度方法的 Top1 能够尽可能地加入答案集,使得两方法得到良好的互补。因此,执行此优先级顺序的准确率能够达到最高。

4.3 DEF 验证实验

为了说明术语 DEF 的有效性,本文进行了验证实验,其验证方法的基本思想是:计算机根据术语 DEF 对术语的区分度越接近于人对术语的区分度,则说明术语 DEF 越有效。计算机对术语的区分,一定程度上表现为术语间的语义距离,因此通过计算术语间的相似度得以实现。

因此,本文将人工标注术语间的相似度与根据术语 DEF 计算术语间的相似度进行相关性分析,即在本文方法构建的知识库中随机抽取 100 对术语。并运用文献[15]的概念相似度计算方法对此 100 对术语 DEF 进行相似度计算。另外,组织 5 个人对这 100 对术语的相似度进行人工判断,将术语间的相似程度分成 6 个等级,记为 0 到 5;取这 5 个人标注结果的平均值,并将其映射到 0 到 1 之间;从而得到两组相似度序列,这两组相似度序列折线的整体趋势对比如图 5 所示。另外,对两组序列进行皮尔逊相关系数计算。若皮尔逊相关系数等于零,则说明二者不相关;若皮尔逊相关系数越接近 1,则表明二者越趋近于正相关;若皮尔逊相关系数越接近－1,则表明二者越趋近于负相关。

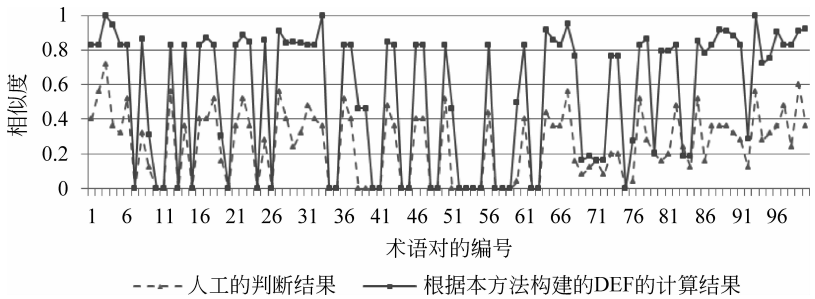

图 5 人工标注结果与根据术语 DEF 计算结果趋势图

根据术语 DEF 计算得到的术语间相似度 x 与人工标注的术语间相似度 y 之间的皮尔逊相关系数 r_{xy} 的计算,如公式(4)所示。

$$r_{xy}=\frac{\sum(X-\overline{X})(Y-\overline{Y})}{\sqrt{\sum(X-\overline{X})^2\sum(Y-\overline{Y})^2}} \tag{4}$$

通过实验得到的皮尔逊相关系数为 0.878 6,大于零接近 1,表明根据术语 DEF 计算术语间的相似度与人工判断术语间的相似度是接近相关的。

另外,从这 100 对术语中随机抽取 12 对术语,组成 4 组,每组有 3 对;表 4 显示了术语 DEF。每组相似度结果如表 5 所示,其中包括计算结果、人工标注结果、减去平均值的计算结果以及减去平均值的人工标注结果。

图 5 中两条折线的整体趋势基本一致,可见两术语相似度计算结果存在一定的正相关性;但图 5 根据术语 DEF 计算结果的折线普遍高于人工标注结果的折线,以及表 5 所示两方法得到相似度结果(相似度计算结果、人工标注结果)的绝对数值存在一定差异,这是由于两种方法的评价标准不同造成的。

然而从皮尔逊相关系数(0.878 6)以及表 5 所示两方法的相似度皆减去平均值的结果(减去平均值的计算结果、减去平均值的人工标注结果)来看,两种方法对不同术语的区分基本一致,验证了本文方法所构建术语 DEF 的有效性。

表 4　术语 DEF

术语	术语 DEF
第 1 组	
间隔标准	DEF＝{Standard|标准：RelateTo＝{Distance|距离}}
技术方案评价标准	DEF＝{Standard|标准：scope＝{estimate|评估：content＝{plans|规划：RelateTo＝{knowledge|知识}}}}
质量管理	DEF＝{manage|管理：content＝{Quality|质量}}
工程管理标准	DEF＝{Standard|标准：scope＝{manage|管理：patient＝{affairs|事务：domain＝{industrial|工}}}}
第 2 组	
通用飞机	DEF＝{aircraft|飞行器：modifier＝{general|共同}}
军用飞机	DEF＝{aircraft|飞行器：domain＝{military|军}}
超轻型飞机	DEF＝{aircraft|飞行器：modifier＝{NotHeavy|轻：degreer＝{over|超}}}
复杂气象飞行	DEF＝{fly|飞：RelateTo＝{weather|天象：modifier＝{complicated|繁}}}
第 3 组	
噪声干扰	DEF＝{obstruct|阻止：RelateTo＝{sound|声}}
红外干扰	DEF＝{obstruct|阻止：anner＝{lights|光}}
欺骗干扰	DEF＝{obstruct|阻止：manner＝{deceive|欺骗}}
大气腐蚀	DEF＝{damage|损害：RelateTo＝{gas|大气}}
第 4 组	
液压射流技术	DEF＝{knowledge|知识：scope＝{physical|物质：RelateTo＝{Strength|力量}}}
自动铆接技术	DEF＝{knowledge|知识：scope＝{fasten|拴连：manner＝{automatic|自动}}}
液体流量校准设备	DEF＝{implement|器具：instrument＝{check|查：content＝{Amount|多少：host＝{liquid|液}}}}
加速度计	DEF＝{tool|用具：RelateTo＝{Speed|速度}}

表 5　相似度结果

术语对		计算结果	人工标注结果	减去平均值的计算结果	减去平均值的人工标注结果
第 1 组					
间隔标准	技术方案评价准	0.6463	0.3600	0.1015	0.0944
间隔标准	质量管理	0.2955	0.0000	−0.2493	−0.2656
间隔标准	工程管理标准	0.5106	0.2400	−0.0342	−0.0256
第 2 组					
通用飞机	军用飞机	0.7376	0.5200	0.2425	0.2544
通用飞机	超轻型飞机	0.8713	0.6000	0.3265	0.3344
通用飞机	复杂气象飞行	0.0000	0.0000	−0.548	−0.2656
第 3 组					
噪声干扰	红外干扰	0.7173	0.4400	0.1725	0.1744
噪声干扰	欺骗干扰	0.6973	0.4000	0.1525	0.1344
噪声干扰	大气腐蚀	0.2968	0.0000	−0.548	−0.2656
第 4 组					
液压射流技术	自动铆接技术	0.5873	0.3200	0.0425	0.0544
液压射流技术	液体流量校准备	0.4591	0.2000	−0.0857	−0.0652
液压射流技术	加速度计	0.2940	0.0000	−0.2508	−0.2556

5 结束语

本文基于 HowNet 的语义理论体系[4]，全面阐述了一种辅助构建航空术语语义知识库的方法，从术语的语义层次，按照自底向上的思想构建航空术语语义知识库，并且将术语内部的依存结构信息，融入知识库构建中。基于术语依存结构，提出了基于搭配词的词义消歧方法和术语 DEF 生成方法。同时提出了基于最大熵分类器与关联单位相似度方法相结合的动态角色关系判断方法，从语义和统计的层面，判断术语内部词语间的关系类型。最后利用术语间相似度的验证方法，通过两相似度序列的皮尔逊相关系数以及人工标注结果与根据术语 DEF 计算结果的对比，验证了本文方法所构建术语 DEF 的有效性。

本文方法以构建航空术语语义知识库为导向，结合自身所具有的语料资源，初步完成了语义知识库闭环构建任务。为了保障知识库的准确性，本文方法采用人机协同的方式构建术语 DEF。面向未来，接下来的任务是：①按照本文方法构建更多高质量的术语 DEF；②从更加开放的语料资源中抽取航空术语以及航空词语间的语义关系，构建丰富、高质量的航空术语语义知识库。

参考文献

[1] 董振东，董强. 知网[EB/OL]. http://www.keenage.com/.

[2] 刘扬，俞士汶，于江生. CCD 语义知识库的构造研究[J]. 小型微型计算机系统，2005，26(8)：1411-1415.

[3] You L，Liu T，Liu K. Chinese FrameNet and OWL representation[C]//Prpceedings of the 6the International Conference on Advanced Language Processing and Web Information Technology. IEEE Computer Society，2007：140-145.

[4] 陈群秀，黄昌宁. 现代汉语述语动词机器词典研究初探[C]. 全国计算机语言学联合学术会议，1993.

[5] 董振东，董强. 建设中文词汇语义资源中的一些问题和我们的对策[EB/OL]. http://www.keenage.com.

[6] 郝长伶，董强. 知网知识库描述语言[C]. 全国计算语言学联合学术会议，2003.

[7] 董振东，董强. 面向信息处理的词汇语义研究中的若干问题[J]. 语言文字应用，2001(3)：27-32.

[8] 张桂平，刁丽娜，王裴岩. 基于 HowNet 的航空术语语义知识库的构建[J]. 中文信息学报，2014，28(5)：92-101.

[9] 王羊羊，等. 基于 HowNet 的术语语义知识库构建技术[J]. 沈阳航空航天大学学报，2016，33(4)：78-84.

[10] 冯志伟. 特思尼耶尔的从属关系语法[J]. 当代语言学，1983，(1)：63-65.

[11] 陈小芳，等. 基于统计和规则相结合的汉语术语语义分析方法[C]. 全国信息检索学术会议. 2010.

[12] 周其焕. 航空术语的构词分析[J]. 中国民航大学学报，2007，25(4)：60-64.

[13] 刘挺，车万翔，李生. 基于最大熵分类器的语义角色标注[J]. 软件学报，2007，18(3)：565-573.

[14] Berger A L，et al. A maximum entropy approach to natural language processing[J]. Computational Linguistics，1996，22(1)：39-71.

[15] 夏天. 汉语词语语义相似度计算研究[J]. 计算机工程，2007，33(6)：191-194.

王思博(1991—)，硕士，主要研究领域为自然语言处理、知识工程与知识管理。
E-mail：life121612@sina.com

张桂平(1960—)，博士，教授，主要研究领域为自然语言处理与机器翻译、知识工程与知识管理。
E-mail：zgp@ge-soft.com

王裴岩(1983—)，通信作者，博士，讲师，主要研究领域为自然语言处理与机器学习、知识工程与知识管理。
E-mail：wangpy@sau.edu.cn