第53卷增刊
2021年10月

南京航空航天大学学报
Journal of Nanjing University of Aeronautics & Astronautics

Vol. 53 No. S
Oct. 2021

DOI:10.16356/j.1005-2615.2021.S.025

基于语义相似度计算的航天标准关联度评价

张　嵩，杨晓明，田　露
（中国航天标准化研究所，北京 100071）

摘要：随着航天领域各级各类标准编制数量的逐年递增，相关标准之间的交叉重复制定问题随之而来。为解决该问题，文中通过计算语义相似度的方法开展航天领域标准名称、标准范围及主要内容的关联度评价。算例分析结果表明，该方法可以有效评估标准间的关联程度。

关键词：相似度计算；语义；依存结构；标准
中图分类号：TP391　　**文献标志码：**A　　**文章编号：**1005-2615(2021)S-0153-04

Evaluation of Aerospace Standard Relevance Based on Semantic Similarity Calculation

ZHANG Song, *YANG Xiaoming*, *TIAN Lu*
(China Academy of Aerospace Standardization and Product Assurance, Beijing 100071, China)

Abstract: With the increasing number of standards at all levels in the aerospace field, the problem of overlapping and repeated formulation of relevant standards follows. In order to solve this problem, the relevance evaluation of standard names, standard scopes and main contents in aerospace field is carried out by calculating semantic similarity. The results of experimental analysis show that this method can effectively evaluate the correlation degree between standards.

Key words: similarity computation; semantics; dependency structure; standard

航天标准关联度是指两个或多个标准之间在对象、内容及范围等方面的关联程度或相似程度。随着传统技术的稳步成熟与新技术的快速发展，航天领域各级各类标准编制数量也逐年递增，相关标准之间的交叉重复制定问题也随之而来。这对标准选用的唯一性与权威性带来挑战，同时也对存量标准的精细管理与增量标准的立项把关提出了严峻考验，因此，有必要开展航天标准关联度评价，以标准之间的关联程度为核心，提出立项标准审核与重复标准制修订等的合理化建议，确保航天领域标准质量。

1　语义词典介绍

自然语言具有海量性和歧义性的特点，海量性体现在不同的词语可以通过组合形成无穷无尽的句子和篇章，歧义性体现在相同的词语甚至相同的句子在不同的语言环境中所表达的意思存在分歧。为了让计算机能够按照人类的思维和知识背景理解和产生自然语言，比较有效的方法是尽可能地将语句离散成词，进而建立词之间的属性关系，而词也是能够表达语义的最小单元。

Hownet是中国著名机器翻译专家董振东等人创建的语言知识库[1]，库中将“义原”（不可再分的基本语言结构单位）作为基本描述单位，以汉语词汇和英语单词所代表的概念为描述对象，建立了各对象之间关系的义原层次体系树形结构。以“打”为例[2]，这个词有一项描述为：DEF＝exercise|锻炼，sport|体育。通过“锻炼”和“体育”两个义原以

收稿日期：2021-05-10；**修订日期：**2021-06-25
通信作者：张嵩，男，工程师，E-mail：18686848519@163.com。

引用格式：张嵩，杨晓明，田露. 基于语义相似度计算的航天标准关联度评价[J]. 南京航空航天大学学报，2021，53(S)：153-156. ZHANG Song, YANG Xiaoming, TIAN Lu. Evaluation of aerospace standard relevance based on semantic similarity calculation[J]. Journal of Nanjing University of Aeronautics & Astronautics，2021，53(S)：153-156.

及 exercise 的属性关系进行了解释描述。目前，Hownet 共有 1 618 条义原信息，以网状关系覆盖了对数万条中文词语的描述，如表 1 所示。

表 1 Hownet 义原信息数量表

Table 1 Quantity of sememe information of Hownet

义原分类	义原条数
Event|事件	813
Entity|实体	142
Attribute|属性/aValue|属性值	433
Quanlity|数量/qValue|数量值	13
Secondary feature|次要特征	100
Syntax|语法	41
Event role & features|动态角色和属性	74

哈尔滨工业大学社会计算与信息检索研究中心研发的语言技术平台(LTP)[3]也提供了类似的服务，包括中文分词、词性标注、命名实体识别及语义角色标注等。以“固体火箭发动机”为例，描述信息见表 2，描述关系见图 1。

表 2 针对“固体火箭发动机”的 LTP 描述信息

Table 2 LTP description information for “Solid rocket motor”

Tag	关系类型	Description
FEAT	修饰角色	Feature
Root	根节点	Root
ATT	定中关系	Attribute
HED	核心关系	Head
n	名词	General noun

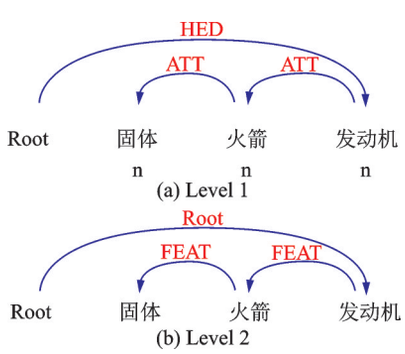

图 1 “固体火箭发动机”LTP 描述关系

Fig.1 LTP description relationship of solid rocket motors

图 1 中以“发动机”为核心概念，通过“固体”和“火箭”两层修饰角色逐级建立描述关系，形成了包含中文分词和属性关系在内的网状结构。

以上方法均是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思想，建立了包含常用分词在内的语义词典库，为计算机识别和处理语义相似性提供了最优的组织方式与基本数据库。

2 语义相似度计算

语义相似度指的是两个词语或句子在形式不完全相同的前提下，在上下文中可以相互替换而不改变语义结构的可能性大小。文献[4]从信息论的角度给出了表征任意两个事物(A,B)相似度的通用公式

$$\mathrm{Sim}(A,B)=\frac{\log p(\mathrm{common}(A,B))}{\log p(\mathrm{description}(A,B))} \tag{1}$$

文献[4]中认为，任何两个事物的相似度取决于其共性信息量的大小和个性信息量的大小。式(1)中分子为描述A,B共性信息量的大小，分母为描述A,B完整信息量的大小。由此引出义原相似度的计算方法[5]：假设两个义原S_1和S_2之间的路径长度为$\mathrm{Distance}(S_1,S_2)$，义原相似度为$\mathrm{Sim}(S_1,S_2)$，则$S_1$和$S_2$之间存在反向对应关系，即$\mathrm{Distance}(S_1,S_2)$越大$\mathrm{Sim}(S_1,S_2)$越小。这里假设两种极端的情况：(1)当$\mathrm{Distance}(S_1,S_2)$为无穷大时，此时两个义原距离无穷远，表示两者完全不相关，$\mathrm{Sim}(S_1,S_2)$值为 0；(2)当$\mathrm{Distance}(S_1,S_2)$为 0 时，此时两个义原之间不存在距离，表示两者完全一致，$\mathrm{Sim}(S_1,S_2)$值为 1。义原相似度可表示为

$$\mathrm{Sim}(S_1,S_2)=\frac{\lambda}{\lambda+\mathrm{Distance}(S_1,S_2)} \tag{2}$$

式中λ为可调参数。

在基于 Hownet 体系下的实际分析过程中，义原相似度的计算只是最基本的操作，最终目的是要通过对词语相似度的计算进而对整句或整段文字的相似度展开评价。文献[6-7]中提出了词语相似度的计算公式

$$\mathrm{Sim}(C_1,C_2)=\max_{j=1,\cdots,m}^{i=1,\cdots,n}\mathrm{Sim}(p_{1i},p_{2j}) \tag{3}$$

式中p_1、p_2分别代表词语C_1、C_2中的义项(属性描述项)。

基于以上计算过程开展短文本相似度计算，假设T_1、T_2为两个语义完整的短文本，判断两者是否相似的关键有以下 3 方面：(1)核心关键词是否一致；(2)义项属性描述信息的重合程度；(3)相同的属性描述信息对应的关键词的重合程度。3 方面影响因素按照相应的比例呈逐级递进关系，综合得分相对客观、准确，其计算公式如下

$$\begin{aligned}\mathrm{Sim}(T_1,T_2)=&\alpha\times\mathrm{Sim}(C_{\mathrm{root1}},C_{\mathrm{root2}})+\beta\times\\&\left(\frac{2\times R_{\mathrm{same}}(T_1,T_2)}{R(T_1)+R(T_2)}\right)+\gamma\times\\&\left(\frac{\sum_{1\leqslant i\leqslant j}\mathrm{Sim}(C_{1i},C_{2j})}{k}\right)\end{aligned} \tag{4}$$

式中：$R(T)$表示具有属性描述项的个数；C_1、C_2表示两段文字中具有相同属性描述项的词对；k表示属性描述项个数的最大值；α、β、γ为可调参数且$\alpha+\beta+\gamma=1$。

3　评价方法与算例分析

3.1　通过标准名称开展关联度评价

算例1　判断“载人飞船返回舱着陆冲击方法”和“弹箭产品爆炸分离冲击试验方法”。通过LTP分词得到两个名称对应的语义依存关系如图2，3所示。

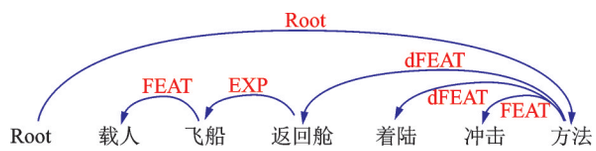

图2　“载人飞船返回舱着陆冲击方法”语义依存关系图

Fig.2　Semantic dependency diagram of name of “Landing impact method for return capsule of manned spacecraft”

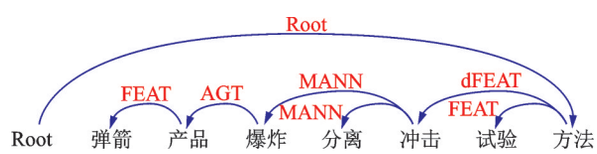

图3　“弹箭产品爆炸分离冲击试验方法”语义依存关系图

Fig.3　Semantic dependency diagram of name of “Pyroshock test method for missle and launch vehicle”

为降低冗余计算量，对关系图作适当简化，去除影响程度较小的深层次属性信息（当计算对象相似度较高且有必要进一步验证准确性时再补充深层属性计算）后，上述两项名称保留各自关系图的主线部分如图4所示。

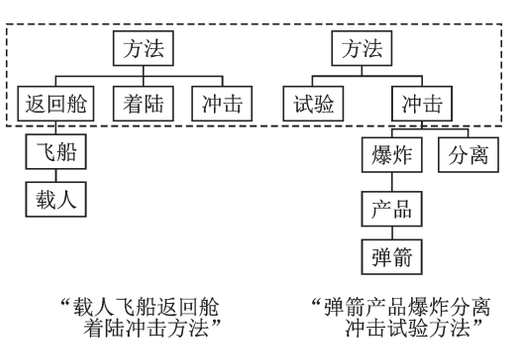

图4　算例1语义关系图简化版

Fig.4　Simplified version of semantic diagram of Example 1

令α、β、γ分别取值为0.2、0.3、0.5。根据式(4)计算得到两者相似度为0.487，可以较为直观地判断两项标准名称相似度不大，关联程度较低。

算例2　判断“航天器钛合金贮箱及气瓶清洗工艺规范”和“钛合金贮箱与气瓶净化工艺规范”。通过LTP分词得到两个名称对应的语义依存关系，其简化图如图5所示。

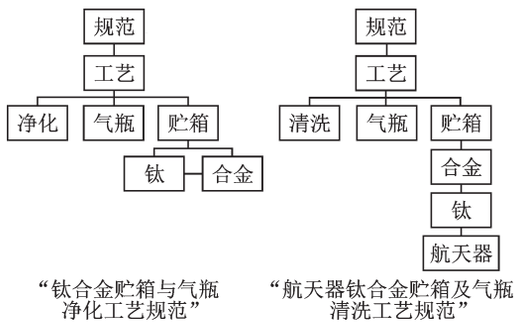

图5　算例2语义关系图简化版

Fig.5　Simplified version of semantic diagram of Example 2

从图5中可以直观看出两个名称的语义依存关系中，关键词根均为“规范”且其下层关系均为“工艺”，第三层通过“贮箱”“气瓶”“清洗”及“净化”等词语与“工艺”建立关系。至此，两个名称的基本框架已基本一致，差距主要体现在更深层次的名词差异上，对整体相似度对比的结果影响不大。按照式(4)计算两者的相似度为0.905，验证了观察的结果，两项标准名称相似度较高，有较强的关联程度，应重点关注两项标准的内容是否存在交叉重复的现象，以保证标准的有效性与唯一性。

算例3　上文算例1和算例2均代表了正向验证，即两项标准名称可以通过观察的方式进行相关与否的判断并通过公式计算验证观察的结果。在实际应用过程中会出现观察结果和实际计算结果相反的情况，此时，仅通过观察就会有局限性，而通过公式计算可以较为准确地得到文字间的相似程度。如判断“固体火箭发动机静止试验参数测试方法”和“固体火箭发动机试车力学环境测试方法”。通过LTP分词得到两个名称对应的语义依存关系，其简化图如图6所示。

观察图6中语义依存关系可以看出两项名称均是针对“测试方法”的，但是测试内容有着明显的区别，而实际工作过程中，“试车”为“静止试验”的口语化表达，“力学环境参数”与“参数”也存在关联，通过同义词替换后，根据式(4)计算得到两者相似度为0.832，说明两项名称之间的相似性很高，具有较大的关联度，需要对两项标准的内容开展详细的比对分析以确保标准的编制质量。同时，也进一步验证了此方法的有效性与可靠性。

3.2　通过标准范围与内容开展关联度评价

通过标准范围与内容开展关联度评价主要基

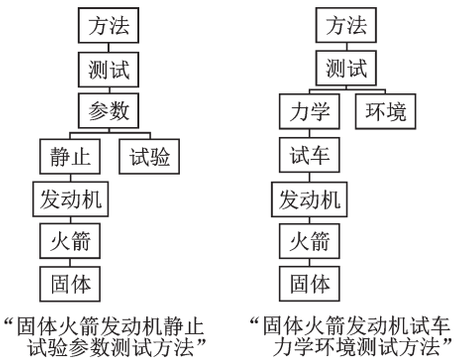

图6 算例3语义关系图简化版

Fig.6 Simplified version of semantic diagram of Example 3

于标准规定范围内所确定的描述性语句部分。例如,《固体火箭发动机离心试验方法》中明确该标准规定了固体火箭发动机离心试验的目的、条件、设备、程序和结果评定等。该标准适用于发动机的离心试验。再如,《固体火箭发动机振动试验方法》中明确该标准规定了固体火箭发动机振动试验的试验条件、系统组成、安装、控制及安全要求、试验程序、试验记录和试验报告。该标准适用于固体火箭发动机振动试验。

按照上述分析过程,从标准适用的范围来看,两项标准分别针对发动机的离心试验与振动试验,本质上两者的关联度不大,而从标准规定的内容来看,两项标准在试验环节的组成与关注的重点要求方面相似程度也不大,因此通过人工判定的方法可以直观地确定两项标准之间关联程度较低。按照式(4)对拆解后的依存关系进行计算,令α、β、γ分别取值为0.2、0.3、0.5,计算得到两项标准相似度为0.288,验证了观察的结果。

由于标准规定的内容与标准名称相比,文字内容较长、语法复杂,适用范围内写法并无严格的规定,在一定程度上受到编写人编制习惯的影响,同时权重因子值的不同也会导致计算结果的波动,综合考虑以上因素,应将评价标准范围与内容的关联度作为补充计算验证的过程,当两项标准名称相似度极高,需进一步检查确认标准内容时,再开展相关计算与评价。

4 结 论

开展标准关联度评价对于领域内存量标准的一致性梳理、重复性剔除以及增量标准的立项把关、相关性分析等工作意义重大。通过结合语义相似度计算的方法开展了航天领域标准名称、标准范围与主要内容的关联度定量评价,选取了典型算例验证评价方法的有效性与可靠性,结果表明此评价方法的有效性及准确性。在此基础上,后续研究工作应重点从以下两方面开展:(1)建立航天标准语义词典数据库,实现库内同义词替换功能并随标准中专有名词的增加而适当扩展数据库,以减少因词库不覆盖造成的计算不准确等问题,提高容错率。(2)构建标准关联自动评价程序。本文所引用的算例中均采用两两比较、手动计算的方式,不具备自动化查询的能力,后续通过LTP平台的api接口及程序语言可以实现数据库中多项筛选比较的自动化流程以提高分析效率。(3)据实调整公式中的参数值,结合航天领域标准名称中的专有名词,参考命名规范规则等内容合理调配参数值以确保相似度计算结果更接近真实值,减小偏差。

参考文献:

[1] HowNet. How Net's Home Page[EB/OL]. (2021-06-30). http://www.keenage.com.

[2] 李峰,李芳.中文词语语义相似度计算——基于《知网》2000[J].中文信息学报,2007,21(3):99-105.
LI Feng, LI Fang. An new approach measuring semantic similarity in Hownet 2000[J]. Journal of Chinese Information Processing, 2007,21(3):99-105.

[3] CHE Wanxiang, LI Zhenghua, LIU Ting. LTP: A Chinese language technology platform [C]// Proceedings of COLING 2010, 23rd International Conference on Computational Linguistics. Demonstrations Volume. Beijing, China: Association for Computational Linguistics, 2010.

[4] LIN Dekang. An information-theoretic definition of similarity semantic distance in WordNet[C]// Proceedings of the Fifteenth International Conference on Machine Learning.[S.l.]: ACM, 1998.

[5] 刘群,李素建.基于《知网》的词汇语义相似度计算[J].中文计算语言学, 2002(7):59-76.

[6] 郭炳元.基于语义树的短文本相似度算法研究与应用[D].湘潭:湘潭大学,2019.
GUO Bingyuan. Research and application of short text similarity algorithm based on semantic dependency tree [D]. Xiangtan: Xiangtan University, 2019.

[7] 李彬,刘挺,秦兵,等.基于语义依存的汉语句子相似度计算[J].计算机应用研究,2003(12):15-17.
LI Bin, LIU Ting, QIN Bing, et al. Chinese sentence similarity computing based on semantic dependency parsing[J]. Application Research of Computers, 2003(12):15-17.

(编辑:孙静)