Modern Linguistics 现代语言学, 2024, 12(5), 310-318
Published Online May 2024 in Hans. https://www.hanspub.org/journal/ml
https://doi.org/10.12677/ml.2024.125363

Hans汉斯

自然语言处理视角下日语复合动词的语义计量方法探索

——以“V1-あげる”和“V1-あがる”为例

高晗瑜[1]，常云翼[2]

[1]西安外国语大学日本文化经济学院，陕西 西安
[2]西安外国语大学经济金融学院，陕西 西安

收稿日期：2024年3月28日；录用日期：2024年5月15日；发布日期：2024年5月29日

摘　要

本文在自然语言处理视角下，利用Doc2vec句向量工具，以“V1-あげる”、“V1-あがる”为例，就日语复合动词的语义计量方法进行了探索。结果表明，语义分类的平均正确率达90%，利用句向量技术对日语复合动词的语义计量研究具有可行性。同时，对于同一复合动词的多个语义，该工具可为大规模自动判断实际语境中的具体语义提供可靠手段。

关键词

自然语言处理，句向量，复合动词，语义计量

Study on Semantic Measurement of Japanese Compound Verbs from the Perspective of Natural Language Processing

—Taking “V1-あげる” and “V1-あがる” as Examples

Hanyu Gao[1], Yunyi Chang[2]

[1]School of Japanese Culture and Economics, Xi’an International Studies University, Xi’an Shaanxi
[2]School of Economics and Finance, Xi’an International Studies University, Xi’an Shaanxi

Received: Mar. 28th, 2024; accepted: May 15th, 2024; published: May 29th, 2024

Abstract

From the perspective of natural language processing, this paper uses Doc2vec sentence vector tool

文章引用: 高晗瑜, 常云翼. 自然语言处理视角下日语复合动词的语义计量方法探索[J]. 现代语言学, 2024, 12(5): 310-318. DOI: 10.12677/ml.2024.125363

and takes “V1-あげる” and “V1-あがる” as examples to explore the semantic measurement of Japanese compound verbs. The results show that the average accuracy of semantic classification is 90%, and it is feasible to use sentence vector technology to study the semantic measurement of Japanese compound verbs. At the same time, for multiple meanings of the same compound verb, the tool can provide a reliable means for large-scale automatic judgment of specific meaning in the actual context.

Keywords

Natural Language Processing, Sentence Vector, Compound Verb, Semantic Measurement

1. 引言

日语中存在大量以“动词 1 + 动词 2”形态出现的复合动词，且复合动词通常具有多义性，成为日语学习和研究的难点之一[1]。由于其前后项结合条件复杂，结合后语义变化多样，在一定语境下复合动词语义判断一直是研究中的关键问题。

教学领域的日语复合动词研究积累较为丰富，多为就其语义扩展、语法或语用特征进行细致描写，或对学习者的理解和产出进行实证调查的研究[2]。语义的大规模自动判断可为教学和习得提供借鉴，而目前尚缺乏具体语境下自动判断复合动词语义的有效手段。

近年来，词向量和句向量的出现为语言学界研究语义提供了抓手。Mikolov 等(2013)推出词向量工具 Word2vec，使得词向量在自然语言处理中得以广泛应用[3]。Word2vec 是一种用于将词语表示为向量的技术，它可以将每个词语映射到一个固定长度的向量空间。这些向量可用于许多自然语言处理应用程序，如词语相似度计算和分类。Word2vec 的主要优势在于它可以将相似的词语映射到接近的向量空间中，因此在许多情况下可以更准确地表示语义相似性[4]。Le 等(2014)作为对 Word2vec 的扩展，提出了 Doc2vec 工具，克服了词向量缺乏语义区分的缺点，可以较好地解析上下文词语的语义[5]。

2021 年以后，词向量的应用才逐渐被引入到日语研究中，但利用这项技术开展大规模的研究较少，且多集中于汉日同形词对比这一研究领域。施建军(2023)通过利用 BERT 词向量技术，探索了词向量在中日通用汉字词汇语义计量研究中的使用方法和有效性。实验表明，词向量能够在该领域的研究中发挥作用，同时也发现了一些影响其效果的因素，为今后进一步加深这方面的研究提出了课题[6]。由于复合动词是由前项动词和后项动词组成，经过测试，将其作为一个词语提取词向量或将前后两个动词的词向量进行简单的相加无法达到较好的相似度匹配效果，因此取平均值或前后词向量相加的方式在日语复合动词语义计量研究中无法使用。因其特殊性，使用 Word2vec 或 BERT 词向量工具进行日语复合动词语义计量研究存在一定的局限性。

本文使用 Doc2vec 工具，以“V1-あげる”和“V1-あがる”为例，就日语复合动词的语义分布及使用情况进行分析。旨在探索 Doc2vec 句向量工具在日语复合动词语义计量中应用的可行性及有效性，为日语复合动词语义计量研究提供借鉴。

2. 日语复合动词的语义计量研究方法

本文以“V1-あげる”和“V1-あがる”为例，使用 Doc2vec 句向量工具，对其语义分类及其使用情况进行研究。

具体思路如下：选取姬野昌子(1999)中对“V1-あげる”和“V1-あがる”的语义分类作为标准义项分类[7]，为提高语义分类的有效性及准确率，为每个义项选取 20 个标准例句进行训练，义项标准例句均来源于《新明解国语辞典(第五版)》《国语大词典》以及新闻语料；选用『現代書き言葉均衡コーパス』语料库作为日语语料来源，分别抽取“V1-あげる”和“V1-あがる”的相关语料；对上述语料进行语料清洗，删除无关信息，进行标准化格式整理；使用SpaCy工具对语料进行日语分句分词以及词形还原；使用Doc2vec模型将目标词句在语料库和标准例句中进行训练并抽取目标句向量。以标准例句向量为参考，使用基于余弦距离的K近邻算法对语料库中的目标语句进行分类；对分类结果进行统计，并计算其准确率。

综上所述，本文使用“Doc2vec + KNN”的算法模型，构建了一个基于上下文的复合动词语义分类器。该分类器具有算力需求小、用时短、准确率高的特点。

2.1. 标准义项及标准例句选取

为方便探索句向量工具在日语复合动词语义计量研究中的使用方法及其有效性，本论文选取了语义较为丰富，且使用频率较高的日语复合动词“V1-あげる”和“V1-あがる”为研究对象。

国内外关于“V1-あげる”和“V1-あがる”的语义研究较多，主要有姬野昌子(1999) [7]、森田良行(1977) [2]、和杨晓敏(2018) [8]等。森田良行(1977) [2]将“V1-あげる”和“V1-あがる”的语义分为“向上方移动”、

Table 1. Classification of standard meanings of “V1-あげる” and some example sentences
表 1. “V1-あげる”的标准义项分类及部分例句

义项	例句
上昇	こぶしを振り上げる 涼風が川から吹き上げる 袖をまくり上げる
下位者/上位者から上位者/下位者に対する社会的行為	お答え申し上げます 民有地を借り上げる 謹んで初春のお慶びを申し上げます
体内の上昇	不安が胸を突き上げる 怒りがこみ上げて来る 悲しみが胸を突き上げる
完了•完成	全体をまとめ上げる 月に 100 万部を売り上げる 政治資金を一挙に調べ上げる
強調	父はぼくの息子をほめあげた 喉元を締めあげた 男を後手に縛り上げる
その他(a)	素直にそれを読み上げる 短い文章を読み上げる 彼は声を張り上げた
その他(b)	人間の自由と独立を高らかにうたいあげた 国歌を朗々と歌い上げた 世界の融和と平和をうたいあげる

注：根据姬野昌子(1999) [7]分类，将“その他”分为その他(a)和その他(b)两类，另外三种分类“引き上げる”、“切り上げる”、“入れ上げる”由于使用过少，此次未将其列入研究范围内。その他(a)为“読みあげる，(声を)張りあげる，(声を)絞りあげる”等。その他(b)：“歌いあげる”、“描きあげる”等与文学或创作相关的表达。

“动作的程度”以及“完了”三类。但该分类过于宽泛，如“褒め上げる”等例外情况较多。姬野昌子(1999)将“V1-あげる”分为“上升”、“对上级或对下级做出的行为”、“体内的上升”、“完了・完成”、“强调”和“其他”；将“V1-あがる”分为“上升”、“完了，完成”、“强调”、“尊敬语”等[7]。杨晓敏(2018)分析了“V1-あげる”的多义性，指出“V1-あげる”的 11 种用法通过同一核心图式下的七种不同图式得以实现，并考察了前项动词语义特征对“V1-あげる”语义的影响[8]。

其中，姬野昌子(1999) [7]的研究相对全面且影响较广。本文选取姬野昌子(1999) [7]中对“V1-あげる”和“V1-あがる”的分类作为基准。同时，每个标准义项分类均选取 20 个标准例句作为参考。具体分类及部分标准例句如表 1 所示。

2.2. 语料库提取与数据处理

本文使用的日语语料来源于『現代書き言葉均衡コーパス』语料库，首先对语料进行预处理，即语料清洗，剔除无关信息(如作者、出版时间、特殊字符、空格等)后，使用 SpaCy 工具及 ja_core_news_lg 词典模型进行分句分词和词形还原。

SpaCy 是基于 Python 编写的开源自然语言处理库，基于自然处理领域的最新研究，SpaCy 提供了一系列高效且易用的工具，用于文本预处理、文本解析、命名实体识别、词性标注、句法分析和文本分类等任务。由于日语涉及词形的变化，例如动词“上げる”，在提取向量时需要把“上げる”、“上げた”、“上げて”等形式一同作为研究对象，因此需要使用 SpaCy 工具将词形还原。

词形还原后，将所有包含“上げる”和“上がる”的句子进行提取，并且只提取“上げる”和“上がる”前为动词的情况(即满足复合动词的条件)，共收集“上げる”语料 13,002 条，“上がる”语料 8131 条。具体处理流程如图 1 所示。

Figure 1. Corpus extraction and data processing flow
图 1. 语料提取与数据处理流程

2.3. 向量提取与语义计量

本文使用 Doc2vec 模型抽取目标词在语料库和标准例句中的向量并进行训练，计算出语料库中目标句向量，以标准例句向量为参考，使用基于余弦距离的 K 近邻算法对语料库中目标语句进行分类，从而完成复合动词的语义分类，并计算每种语义在该词下的使用频率及使用情况。

2.3.1. Doc2vec 模型

Doc2vec 模型是基于 Word2vec 的一种文档向量表示[3]。Mikolov 等(2013)推出 Word2vec 工具，用于计算词向量[3]。Word2vec 模型是建立在词袋假设的基础上，利用词的上下文信息将一个词转化成一个低维实数向量，越相似的词在向量空间中越相近[5]。目前，词向量已被广泛应用于中文分词、情感分析、句法依存等诸多研究领域。Doc2vec 模型在 Word2vec 模型的基础上，将句中单词的词向量做线性组合，得到句向量并对句子进行编码。因此，Doc2vec 是可以将词、句子乃至文档表征为实数值向量的一种高效算法模型，可以把对文本内容的处理简化为多维向量空间中的向量运算。通常文本语义上高相似度的句子，其向量空间上亦呈现高度的相关性[9]。

2.3.2. 向量提取过程

本文使用 Python 中的 Gensim 库来实现句向量的训练和提取。首先，将词形还原后的语料文本数据

进行合并，共 21,133 条。之后，通过调用 Gensim 库中的 Doc2vec 模型来对文本数据进行处理。在运行算法时，Doc2vec 模型对每一个出现的词建立词向量，再对向量做加权线性运算，从而得到该句的句向量。如表 2 所示，设置模型训练时使用的滑动窗口为 5，即根据每一个词前后的五个词来计算词向量；设置输出向量为 600 维，即模型计算和输出的词向量、句向量均为 600 维。

Table 2. Training parameters of Doc2vec
表 2. Doc2vec 的训练参数

名称	数值
windows	5
vector_size	600
min_count	3
sample	1e−3
negative	5
seed	1
epochs	5

如表 3 所示，随着句子中词语的不同，句向量中每个维度上的数值亦会出现变化。训练模型后，使用该模型将标准例句和语料库中的句子均转化为句向量，对句子进行“编码”并开始训练和测试机器学习分类器。

Table 3. Calculation results of sentence vectors for some target sentences
表 3. 部分目标词句的句向量计算结果

义项	目标词句	句向量
上昇	湧き上がる歓声	[0.017532473, 0.035555243, …, 0.014534176]
	ロケットも打ち上げた	[−0.007243161, 0.00690245, …, 0.005806065]
	花火を打ち上げる	[0.012157571, 0.019556284, …, 0.005722438]
完了•完成	最近は日本でも情報という怪物に振り回されて、悪戦苦闘してモノをつくり上げる町工場的精神を軽んじる風潮がみられる	[−0.00083498, 0.020250814, …, 0.036870282]
体内の上昇	不安が、胸を突き上げる	[0.021420369, 0.041710556, …, 0.02897873]
	船が河岸を離れ、次第に遠ざかってゆき、見り送った平七郎の心には、<ああ、ついに別れてしまったのか>という絶望感が突き上げていった	[0.039991673, 0.021966556, …, 0.0843959]

2.3.3. 复合动词语义使用情况计量

本文使用余弦距离来衡量句向量之间的差异大小。以$[x_1, x_2, \cdots, x_n]$和$[y_1, y_2, \cdots, y_n]$来表示语句 A 和语句 B 的句向量，则两者句向量的夹角余弦值为：

$$\cos\theta = \frac{\sum_{i=1}^{N} X_i Y_i}{\sqrt{\sum_{j=1}^{N} X_j^2}\sqrt{\sum_{k=1}^{N} Y_k^2}}$$

简单来看，两个句向量之间的夹角余弦值与两者向量模的大小无关，而与向量各个维度上数值比例

有关。如果两个句子中词语语义越相近，则句向量中各个维度的比例也就越相近，两个句向量的夹角就越小，余弦值就越大。我们取$1-\cos\theta$作为余弦距离，余弦距离取值在 0 到 2 之间，余弦距离越小，两个句向量各维度数值占比越接近，语义就越相似[10]。

$$1-\cos\theta=1-\frac{\sum_{i=1}^{N}X_iY_i}{\sqrt{\sum_{j=1}^{N}X_j^2}\sqrt{\sum_{k=1}^{N}Y_k^2}}$$

如表 4 所示，“花火を打ち上げる”与“ロケットも打ち上げた”的余弦距离为 0.580631324，而与“その為、基本単語の意味だけを覚え、品詞別に覚えていない事と、頭から訳さず文の後ろから訳し上げるクセがついているからです”的余弦距离为 0.823746422，可以判断该句与前者的相似度更高。

Table 4. Calculation of cosine similarity between some sentences
表 4. 部分句子间的余弦相似度计算

句子	余弦距离
花火を打ち上げる ロケットも打ち上げた	0.580631324
花火を打ち上げる その為、基本単語の意味だけを覚え、品詞別に覚えていない事と、頭から訳さず文の後ろから訳し上げるクセがついているからです	0.823746422
不安が、胸を突き上げる 船が河岸を離れ、次第に遠ざかってゆき、見り送った平七郎の心には、<ああ、ついに別れてしまったのか>という絶望感が突き上げていった	0.590004923
不安が、胸を突き上げる ロケットも打ち上げった	0.699240118

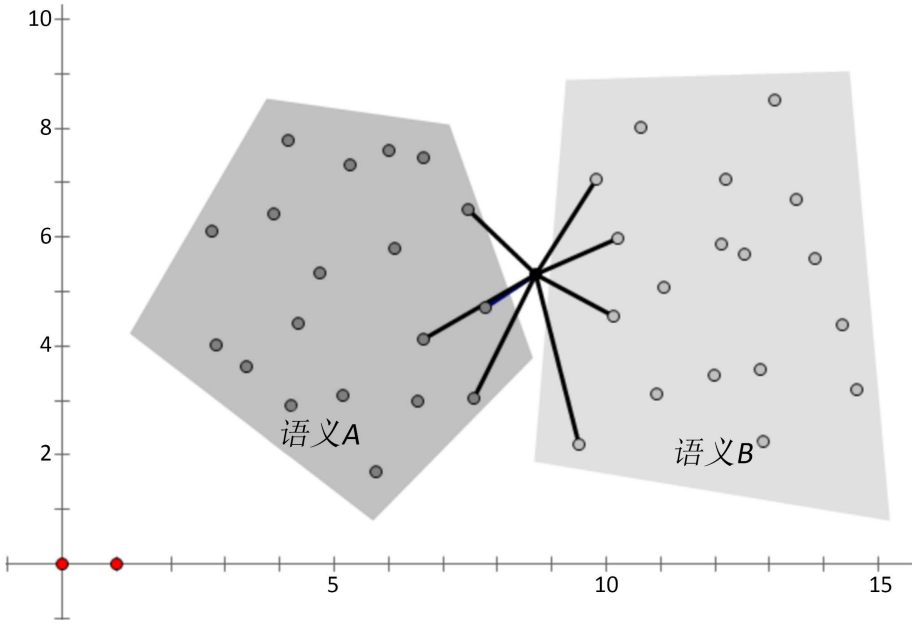

Figure 2. Schematic diagram of the principle of K-Nearest Neighbor algorithm
图 2. K 近邻算法原理示意图

本文采用 K 近邻算法(KNN 算法)对语料进行语义分类。KNN 是机器学习中比较经典的分类算法，最早是由 Cover 和 Hart 在 1968 年提出，其原理为：如果一个样本在特征空间中的 k 个最相似的样本中的大多数属于某一个类别，则该样本也属于这个类别。如图 2 所示，如果一个待分类向量的余弦距离最近的 15 个已分类向量大多属于语义 A，那么此待分类向量类别为语义 A。

在上述步骤中，每个语义下分别已选取 20 个标准例句。其中，“V1-あがる”例句向量共 100 个，“V1-あげる”例句向量共 140 个。设置 K 为 15，对所收集的 21,133 条语料数据的句向量按照 KNN“相对多数”原则进行分类。将每一个义项下的例句数相加，得到该义项的使用次数，即频数，同时得到该义项的使用频率。“V1-あげる”和“V1-あがる”的使用情况(频率)如表 5、表 6 所示。

Table 5. Semantic item distribution and usage of “V1-あげる”
表 5. “V1-あげる”语义项分布及其使用情况

义项	频数	频率
上昇	8518	65.5%
下位者/上位者から上位者/下位者に対する社会的行為	465	3.5%
体内の上昇	64	0.4%
完了・完成	3207	24.6%
強調	149	1.1%
その他(a)	468	3.5%
その他(B)	130	0.9%

Table 6. Semantic item distribution and usage of “V1-あがる”
表 6. “V1-あがる”语义项分布及其使用情况

义项	频数	频率
上昇	6157	75.7%
完了・完成	1636	20.1%
強調	78	0.9%
図々しさ	86	1.0%
尊敬語	173	2.1%

由表 5 和表 6 可知，“V1-あげる”和“V1-あがる”使用频率最高的都为其基本义，即“上升义”，其中“V1-あげる”的“上升义”使用频率达到 65.5%，其次是“完成义”，使用频率达到 24.6%；“V1-あがる”的“上升义”使用频率达到 75.7%，其次是“完成义”，使用频率达到 20.1%。

3. 使用 Doc2vec 进行语义计量的有效性分析

为验证 Doc2vec 句向量工具在日语复合动词语义计量研究中的有效性，本文使用人工标注的方法，从每个义项下随机选取 50 个例句，交由两位日语专业教师进行人工判断和标注(判断结果几乎一致，说明语义之间的区分度较大)，其中，“V1-あげる”抽取 350 句，“V1-あがる”抽取 250 句。将人工判断并标注的语义分配结果与使用 Doc2vec 句向量工具进行自动分类的匹配结果进行比较，判断使用 Doc2vec 句向量工具进行自动分类的正确率，并围绕比较的结果进行有效性分析。

本文使用统计学中常用的 Kappa 系数进行有效性分析。Kappa 系数是用于一致性检验的指标，也可

用于衡量分类的效果。P_0 是每一类正确分类的样本数量之和除以总样本数，也就是总体分类精度，假设样本总量为 n，每一类机器识别分类样本个数分别为 $a_1,a_2,\cdots,a_n$，而人工判断校对出来的每一类样本个数分别为 $b_1,b_2,\cdots,b_n$，P_e 为所有类别分别对应的“实际与预测数量的乘积”之总和除以“样本总数的平方”。Kappa 系数计算式如下所示：

$$\text{Kappa}=\frac{p_0-P_e}{1-P_e}$$

以“V1-あがる”为例，利用 Kappa 系数对机器分类的有效性进行分析，首先需要做混淆矩阵，如下表 7 所示。

Table 7. Confusion matrix of“V1-あがる”
表 7. “V1-あがる”混淆矩阵

人工 \ 机器	义项 1	义项 2	义项 3	义项 4	义项 5
义项 1	48	1	1	0	0
义项 2	7	42	0	0	1
义项 3	8	1	40	1	0
义项 4	5	2	1	42	0
义项 5	6	1	0	0	43

Kappa 系数可分为五个不同级别：0.0~0.20 为极低的一致性(slight)、0.21~0.40 为一般的一致性(fair)、0.41~0.60 为中等的一致性(moderate)、0.61~0.80 为高度的一致性(substantial)、0.81~1 为几乎完全一致(almost perfect)。

通过一致性计算，可得 Kappa 系数为 0.825，即说明使用 Doc2vec 工具对“V1-あがる”进行语义计量具有极高的一致性。同理，可计算出 Doc2vec 工具对“V1-あげる”分类的 Kappa 系数为 0.861，同样具有极高的一致性。

总体上看，使用 Doc2vec 工具对“V1-あがる”和“V1-あげる”进行语义分类的平均准确率达 90%，其中表示“上升”的语义分类准确率最高，平均达到 96%，这可能与“V1-あがる”和“V1-あげる”的“上升义”为基础义，使用频率最高，词语搭配范围最广有关；表示“强调义”的语义分类准确率最低，平均为 80%。

4. 结语

本文在自然语言处理视角下，利用 Doc2vec 句向量工具，以“V1-あげる”和“V1-あがる”为例，就日语复合动词的语义计量方法进行了探索。

研究表明，使用 Doc2vec 工具对日语复合动词的语义大规模机器自动分类无需大算力和预训练模型，同时具有用时短、方便快捷的特点。机器分类与人工判断分类的结果具有极高的一致性，平均正确率达 90%，说明其具有一定的可行性。但仍会出现判断错误，其原因可能与以下几点有关：一是使用 SpaCy 工具分词发生错误；二是语料库中的部分“あげる”未完全统一格式，“上げる”和“あげる”被误判定为两个词；三是标准例句选取数量过少，无法提取到更加完整的特征；四是语料库(训练集数据)中出现不规范或者错误使用，导致无法进行有效匹配。

在今后的研究中，可以尝试使用更加精确的分词方式，统一日语语料格式，选取大量的标准例句来进行机器学习等方式，进一步提高语义分类的准确性和正确率。

参考文献

[1] 杨晓敏. 日语复合动词的多义性研究[D]: [博士学位论文]. 上海: 上海外国语大学, 2009.

[2] 森田良行. 基礎日本語 1－－意味と使い方[M]. 东京: 角川書店, 1997.

[3] Mikolov, T., Sutskever, I., Chen, K., *et al*. (2013) Distributed Representations of Words and Phrases and Their Compositionality. *Advances in Neural Information Processing Systems*, **26**, 3111-3119.

[4] 唐明, 朱磊, 邹显春. 基于 Word2vec 的一种文档向量表示[J]. 计算机科学, 2016(6): 214-217+269.

[5] Le, Q.V. and Mikolov, T. (2014) Distributed Representations of Sentences and Documents. arXiv: 1405.4053.

[6] 施建军. 基于词向量的汉日通用汉字词语义计量研究方法探索[J]. 外语教学理论与实践, 2023(1): 18-36.

[7] 姫野昌子. 複合動詞·「~あがる」, 「~あげる」および下降を表す複合動詞類[M]. 东京: ひつじ書房, 1999.

[8] 杨晓敏. 核心图式理论下日语复合动词后项“~上げる”多义性再考[J]. 复旦外国语言文学论丛, 2018(3): 33-41.

[9] Capel, A. (2012) Completing the *English Vocabulary Profile*: Cl and C2 Vocabulary. *English Profile Journal*, **3**, e1. https://doi.org/10.1017/S2041536212000013

[10] McCarthy, D., Apidianaki, M. and Erk, K. (2016) Word Sense Clustering and Clusterability. *Computational Linguistics*, **42**, 245-275. https://doi.org/10.1162/COLI_a_00247