G-1

建築作品データベース構築に伴う概念ベースの有効性の検証

Verification of the Validity of the Architecture Database by Natural Language Processing Technology

登川 幸生[1]，○片江 優[2]，小島 紗綾[3]，松浦 佑介[2]，宇野木 隆宏[4]

Sachio Togawa[1],Suguru Katae[2] ,Saya Kojima[3],Yusuke Matsuura[2], Takahiro Unoki [4]

This research aimed at using for search or grasp of the characteristic by constructing the database from the document about the architecture.The knowledge base was constructed from the dictionary using natural language processing technology, and the architecture database using the free description sentence by architects, such as the characteristic and the intention, was constructed.In order to raise the feature and the accuracy of search and extraction, accuracy is improved as compared with the conceptual base which newly built the Japanese dictionary and the architecture term dictionary.

1.　研究背景・目的

日常的に行われている情報収集の手法として「検索」がある．しかし「検索」は作品名や設計者名などの直接的な情報のみでしか情報を検出することができず，建築物の持つ全体像などの抽象的な情報では必要としている情報を検出することができない．さらに建築雑誌に掲載されている画像や図面，設計者が記述した説明文書なども掲載されているが，それを検索する有効なシステムは未だ確立していない．

著者[1]らの研究では，自由記述された建築作品に関する文章に対し，自然言語処理技術を利用して建築作品の検索の他に作品の特徴を抽出・類似性の把握が可能な建築作品データベースを構築した．これは概念ベースを利用した手法であり，概念ベース構築時の知識源として国語辞書を用いている．しかし，国語辞書は収録されている語数や表記内容などに違いがあり，それにより記述された文章を概念ベースで表記しきれていない可能性があると考えられる．

そこで本研究では， 自由記述された建築に関する文章と国語辞書から構築した概念ベース，広辞苑(第 6 版)[2]とを照合することで，建築作品文書に対する概念ベースの有効性を検証することを目的とする．

2.　概念ベース

2.1　概念ベース構築手法

概念ベース構築フローをFig.1に示す．概念ベース構築の手順として，まず辞書データから概念と属性を獲得する．その際に見出し語は概念，語義文はその概念の属性となる．次に1次精錬で獲得した概念および属性を精錬し縮小を図る．概念と属性から名詞・形容詞・動詞の各自立語以外を除去する．更に再帰的観点から概念に存在しない属性を不要属性とし，属性種類数と概念数が同じになるまで除去作業を繰り返す．次に2次精錬として，再帰的参照・逆引き参照を行うことで得られた新たな属性を考慮した上で，重みづけの計算を行い，元の概念と線形結合する．最後に概念ベースの中の頻出属性の重みを下げるような係数を概念の属性それぞれの重みに加え，それぞれの概念の重みを調整する．

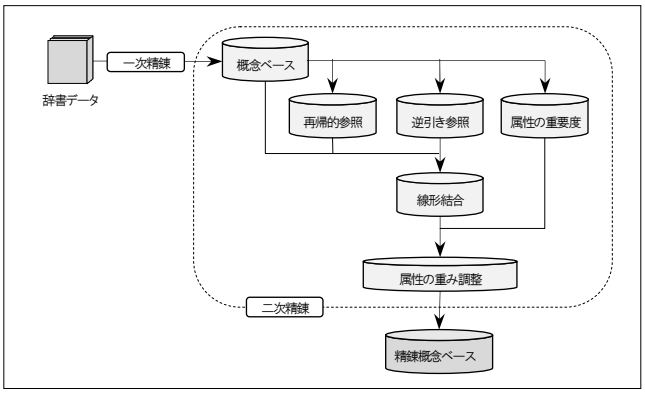

Fig.1 Construction of Knowledge Base

2.2　概念ベース構築結果

作製された概念ベースは約7万語を収録する明鏡国語辞典[3]を用い，属性抽出のための形態素解析には茶筅を用いた．概念ベースの構築の結果として，概念数及び属性種類数は4450語抽出された．

3.　建築作品の文章比較

3.1　対象文献

本研究では対象文献として新建築社『住宅特集』[4]における 2007 年 12 月から 2008 年 11 月号までの 1 年分 154 作品の住宅に関する解説文を用いることとする．

対象文献 154 作品の自由記述された文書を形態素解析にかけ，自立語以外の品詞を削除する．文書に多く含まれる助詞や副詞などは，それ自体では意味をなさないた

1 :日大理工・教員・海建, Department of Oceanic Architecture and Engineering, College of Science and Technology, Nihon University. Prof. Dr. Eng
2 : 日大理工・院・海建, Department of Oceanic Architecture and Engineering, Graduate school of Science and Technology, Nihon University.
3 :日大理工・研究員・海建, Department of Oceanic Architecture and Engineering,　Researcher student of Science and Technology, Nihon University.
4 : 日大理工・学部・海建, Department of Oceanic Architecture and Engineering, College of Science and Technology, Nihon University.

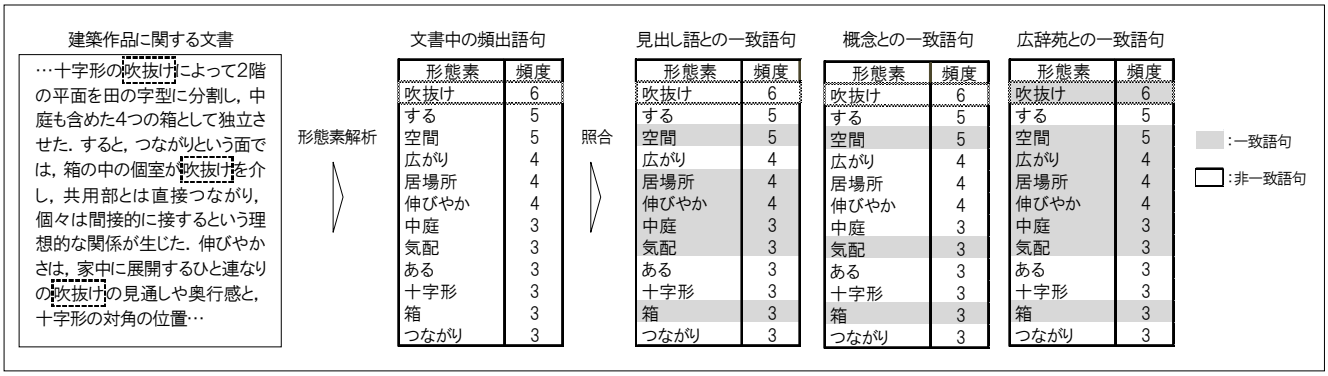

Fig.2 The Collation Result of The House of Nakahara

めである．次に文章中の頻出語句を文書特徴として抽出する．最後に辞書データの見出し語と一致した語句と照合をとり，同様に概念ベースの概念と文書の一致語句との照合をとる．Fig.2 に建築作品文書の形態素解析から見出し語・概念を照合するまでの手順を示す.

3.2　照合結果

154 作品と見出し語が一致した語句の総数は 18,912 語概念が一致した語句の総数は 9,936 語であった．1 作品に対する一致語句の平均数は，見出し語では約 122 個，概念では約 64 個であった．この結果から，記述文章に対しての概念が見出し語に比べ，平均的に語句が半分に減少していることが分かる．これは概念ベース作製時に、抽出される語句が減少してしまうためと考えられる.

具体例として，「中原の家」を挙げる．Fig.2 において見出し語と概念の，文章頻出語句との一致語句表を比較する．記述文書と概念との一致語句は「空間」「気配」など一般的な単語で語句数も少ないことから，一致する語句が少ない. それに対し見出し語では「伸びやか」「中庭」など，多くの語句が一致している．この結果は, 概念ベース構築の 1 次精錬段階で，自立語を削除する以外に属性にない概念などを削除した為, 語句数が減ってしまったことが原因であると考える.

また，Fig.2 における文章中の頻出語句の表から「吹抜け」という語句が頻出語句の上位に位置していることが分かる．実際に記述された文章を確認すると，この建築作品が特徴的な十字形の吹抜けを設け，部屋を田の字に分割した建物であることが記載されている．これらのことから，「吹抜け」がこの作品の特徴的な語句であることが挙げられる．Fig.2 の照合結果を見ると「吹抜け」という語句は，見出し語と概念の一致語句としては抽出されなかった．しかし，広辞苑と比較すると Fig.2 で示すように一致語句として抽出することができた．このことから，単一の辞書より編纂された概念ベースだけでは，特徴的な語句が抽出できない場合があると考えられる．今回の結果では，広辞苑では抽出することが可能であるため，

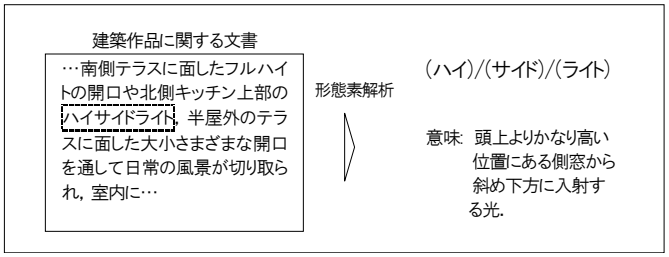

Fig.3 The Morphological-Analysis Result of a Technical Term

広辞苑の辞書データを追加することでより多くの語句を特徴として抽出することが可能であると考えられる.

一方, Fig.3 で示す対象文献中では「ハイサイドライト」という建築の専門用語が使用されている．この語句は広辞苑にも掲載されておらず, 形態素解析を行うと「ハイ」「サイド」「ライト」と分断されてしまう．このように 1 語として特徴を表す語句が 3 つの別々の語句として扱われてしまい，正しい特徴として抽出されていない．建築の専門用語を扱う建築学用語辞典[5]には掲載されていることから，広辞苑だけでなく専門的な辞書を辞書データとして追加することでより多くの特徴を抽出することが可能と考えられる.

4.　まとめ

建築作品に関する文書から特徴を抽出する際に，その作品のみに頻出する語句は重要な要素である. 照合の結果から現在の概念ベースでは抽出しきれない語句も，広辞苑や建築学用語辞典では抽出できることが分かった．このことから広辞苑や建築学用語辞典を有効に活用できる概念ベースを構築することにより，今まで抽出できなかった特徴も抽出が可能と考えられる.

参考文献

[1] 片江優 他: 自然言語処理技術を利用した建築作品データベースの構築，日本大学理工学部学術講演会，2010.11
[2]『広辞苑 第 6 版』: 新村出編，岩波書店発行，2008.1.11
[3]『明鏡国語辞典』: 北原保雄編，大修館書店発行，2003.4.10
[4]『新建築「住宅特集」』，新建築社，2007.12-2008.11.
[5]『建築学用語辞典 第 2 版』: 日本建築学会編, 岩波書店発行, 2008.5.20