

(55) 大規模テキストを用いた 土木用語を含む形態素解析用辞書の作成

荒木 光一¹・吉田 龍史²・田中 裕二²・広瀬 允佳²
・瀧本 圭介²・高橋 浩貴²

¹正会員 五大開発株式会社 技術研究所（〒921-8051 石川県金沢市黒田 1-35）

E-mail: kouichi.araki@godai.co.jp

²非会員 五大開発株式会社 システム技術部（〒921-8051 石川県金沢市黒田 1-35）

E-mail: {yoshita, yuji.tanaka, hirose, takimoto, takahashi}@godai.co.jp

土木分野においてテキストマイニングは幅広く行われている。日本語のテキストマイニングでは文書を単語に分割する形態素解析が前処理として行われるが、このときに用いられる形態素解析の辞書には土木用語が十分に含まれていない。そのため、土木用語を加味したテキストマイニングが十分にできない。そこで、本研究では、土木用語を含む多量のテキストを用いて、土木用語を含む形態素解析の辞書を作成する。作成した辞書を用いてトピック分析をした結果、従来の形態素解析の辞書には含まれていない土木用語を抽出できたことを確認した。一方で、土木用語の抽出は人手でなく、プログラムによる自動抽出であるため、意味を持たない単語も含まれていることが課題として浮き彫りとなった。

Key Words: civil engineering, technical term, morphological analysis, text mining, topic model

1. はじめに

アンケートや資料などに対してテキストマイニングを行うことで、様々な事柄や局面に関する知見と知識を得ることができる。土木分野においても、様々な目的のためにテキストマイニングは幅広く行われている^{1,2,3}。

日本語のテキストに対してテキストマイニングを行う場合、メインの分析の前処理として形態素解析が行われる。形態素解析では、形態素解析の辞書を元に、日本語のテキストを単語単位に分割する。しかしながら、従来

の形態素解析の辞書には土木用語が十分に含まれていない。そのため、土木分野のテキストマイニングを行う場合、土木用語を十分に加味した分析ができない。

そこで、本研究では、土木用語を加味した分析ができるように、土木用語を含む形態素解析の辞書を作成し、この有効性を示す。形態素解析の辞書に追加する土木用語は、土木用語を多く含む多量のテキストから専門用語抽出ソフトウェアで取得する。

2. 形態素解析と既存の辞書

形態素解析とは、単語と品詞が登録されている形態素解析の辞書を元に、意味のある最小限の単語（形態素）にテキストを分割する処理である。形態素解析はトピック分析や類似語検索などのメインの分析の前処理として行われる。例えば、図-1 のように「今日は晴れです。」に対して形態素解析を行う場合、形態素解析の辞書を参照して「今日／は／晴れ／です／。」と複数の形態素に分割する。そして、これらの形態素を元に、メインの分析を行う。本研究は、図中の形態素解析の辞書を対象としており、この辞書に土木用語を追加する。

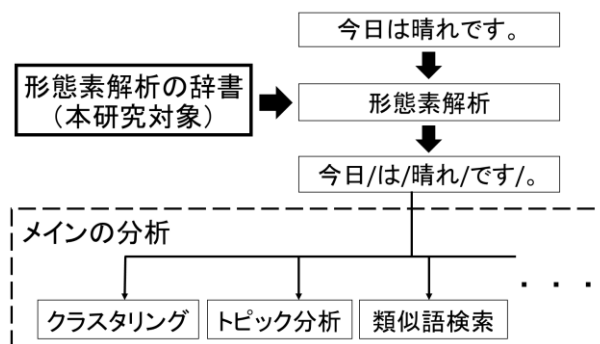


図-1 テキストマイニングにおける
形態素解析の立ち位置

既存の形態素解析の辞書として、NAIST-jdic⁴⁾や mecab-ipadic-Neologd⁵⁾が存在する。NAIST-jdic は IPAdic の後継であり、表記ゆれ情報や複合語の構造を付加した辞書である。しかしながら、NAIST-jdic は現在メンテナンスされていない。mecab-ipadic-Neologd は SNS やネットニュースなどの Web 上の言語資源から単語を取得し、週に2回更新している。そのため、非常に幅広い単語を収録しており、新語や固有表現への対応も早い。一方で、専門書や指針書などは対象としていないため、土木用語は十分に登録されていない。実際に、土木用語である「斜面安定計算」と「補強土」を mecab-ipadic-Neologd で形態素解析すると「斜面／安定／計算」と「補強／土」となり、1つの単語として認識しない。そのため、mecab-ipadic-Neologd を用いた場合、土木用語を十分に加味して分析ができない。

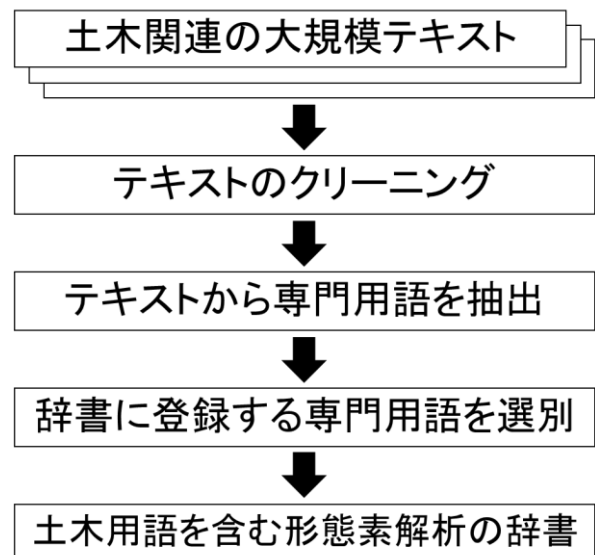


図-2 本研究で作成する形態素解析の辞書の作成フロー

3. 土木用語を含む形態素解析の辞書

(1) 作成フロー

土木用語を多量に含む大規模な土木関連のテキストから専門用語抽出ソフトウェアによって専門用語を抽出することで、土木用語を多く取得できることを前提に、本研究では形態素解析の辞書を作成する。

図-2 に、土木用語を含む形態素解析の辞書の作成フローを示す。入力土木分野に関する多量のテキストである。初めに、入力されたテキストをクリーニングする。具体的には、全角文字のアルファベットと記号は半角文字に変換し、半角文字のカタカナは全角に変換する。また、タブと数値を削除し、スペースが連続した場合は1つのスペースに変換する。

次に、専門用語抽出ソフトウェアを用いて、クリーニングされたテキストから専門用語を抽出する。一般的に、専門用語抽出ソフトウェアは連続する名詞の関連性から専門用語を抽出しているため、土木用語のみを対象とした抽出は行っていない。したがって、土木用語以外の専門用語も抽出される。

抽出された専門用語から登録すべき専門用語を選別する。ここでは、文字数が2以上であり、既存の形態素解析の辞書に含まれていない専門用語をピックアップする。そして、ピックアップされた専門用語を、本研究で作成する形態素解析の辞書に登録する。

(2) 辞書の作成

水工学や構造工学など土木分野全般の土木用語を網羅するために、大規模テキストとして、国や自治体が2018年6月から2019年1月までに公開した約9万件の入札テキストを用いた。入札テキストにはPDFや図面などの

表-1 作成した形態素解析の辞書の特徴

単語数	2,968,234 単語
最小文字数	2 文字
最大文字数	4,506 文字
平均文字数	11.99 文字

添付資料のテキストも含まれる。

図-2 のテキストのクリーニング、テキストからの専門用語の抽出、辞書に登録する専門用語の選別は全て python3 で実装した。テキストからの専門用語の抽出では専門用語抽出ソフトウェア termextract⁶⁾を利用した。termextract には、入力が異なる複数の実装がある。ここでは、森山ほか⁷⁾の研究において高い精度として示された、パープレキシティによる需要度で専門用語を抽出する実装を利用した。この実装の入力は形態素解析の結果であるため、クリーニングした入札テキストを MeCab⁸⁾と mecab-ipadic-Neologd で形態素解析した結果を termextract に入力した。そして、この termextract の出力に対して、前述した専門用語の選別を行った。

4. 評価

(1) 作成した形態素解析の辞書の分析

表-1 に、9万件の入札テキストから作成した形態素解析の辞書の特徴を示す。登録された単語数は約296万語となった。これらの単語における最小文字数は2文字、最大文字数は4,506文字となった。また、平均文字数は11.99文字となった。

登録された単語は、人手で1単語毎チェックしていないため、テキストマイニングには不要であり、意味を持

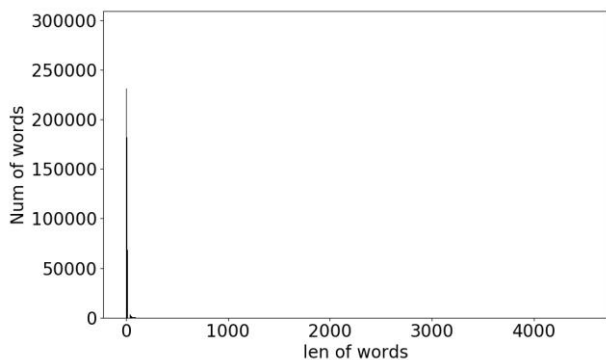


図-3 作成した辞書の各文字数における単語数

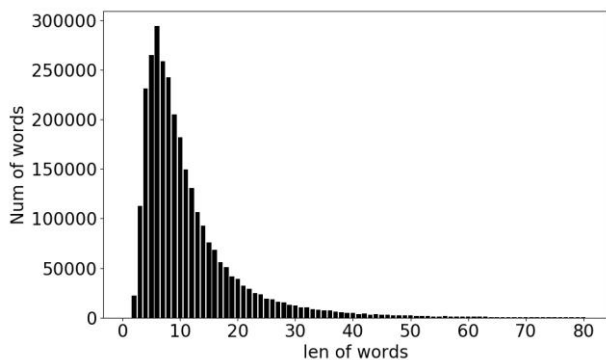


図-4 作成した辞書の2-80文字における単語数

たない単語が含まれていた。具体的には、「式測量業務式用地測量作業計画業務資料調査土地」や「委託料公園管理費課長文書取扱主任課長代理担当者様式」などである。このような単語が含まれてしまった原因は、PDFや図面からテキストへの変換が正確にできず、名詞が連続してしまったことが挙げられる。特に、図面においては、変換後のテキストを確認すると名詞の羅列が多くなっていた。このような意味を持たない単語の傾向として、文字数が多いということがあった。

図-3に、作成した辞書の各文字数における単語数を示す。本研究で登録された単語の多くは、文字数が少ないことが分かる。図-4に、2文字から80文字にフォーカスした単語数を示す。6文字が最も単語数が多くなり、294,226単語とであった。7文字以上は、文字数が増加するにつれて、徐々に単語数は減少している。

(2) トピック分析による適用事例

作成した形態素解析の辞書の適用事例として、トピック分析を行った。本トピック分析にはPythonのライブラリであるgensim⁹⁾を用いた。形態素解析にはMeCab¹⁰⁾を利用し、形態素解析の辞書にはmecab-ipadic-Neologdと本研究で作成した辞書を利用した。ただし、前節の傾向を踏まえて、本研究で作成した辞書の全単語を使用せず、20文字以下の2,625,523単語（全単語の88.4%に相当）を用

表-2 土木学会論文誌の沙録のトピック分析結果

クラスタ 1	クラスタ 2	クラスタ 3
*のり面	インバート	*定着長
溶質	*緊張力	*排水性舗装
*吸水性	*低減効果	ラボ
ブロック	防振	エネルギー
*コンクリート構造物	*健全度	大学院
ダンパー	座屈	性格
*固化処理土	岩盤	学習者
*付属施設	*地盤振動	道路
程度	まなごし	*設計法
亀裂	窯元	勉強会

表-3 入札テキストのトピック分析結果

クラスタ 1	クラスタ 2	クラスタ 3
*号表	*同種業務	*単価表
*当たり名称	暴力団	*会社等
*内業	*業務費内訳書	*申請書等
*参加表明書	*土整備部	明細書
*会社等	企業体	*監理技術者等
技師	活動実績	*社会保険等
*技術提案書	*別途協議	*紙入札方式
*設計共同体	積算基準	*技術提案書
*単位当り	技師	週休
*外業	*した業務	*長通知

いた。

トピック分析の対象テキストとして、土木学会論文誌の沙録と入札テキストを用いた。土木学会論文誌の沙録は8分野から各10件の沙録を、入札テキストは2019年5月1日から同年6月5日までの745件を利用した。

表-2と表-3に、トピック分析の出力である各クラスタにおいて重要となる上位10単語を示す。*印が付いた太文は、mecab-ipadic-Neologdには登録されておらず、本研究で登録された単語である。土木学会論文誌の沙録では本辞書で登録された単語は12個含まれていた。したがって、mecab-ipadic-Neologdのみによるトピック分析より、土木用語を考慮してトピック分析していることが分かる。

入札テキストにおいては、「内業」や「技術提案書」などをカバーできた。一方で、「した業務」や「長通知」など単語として成り立っていないものも含まれていた。特に、「等」を末尾に含む単語は、その前の「申請書」や「会社」が単語として認識されるべきであり、mecab-ipadic-Neologdでも認識できた。これらの対応は、今後の課題となる。

5. おわりに

本稿では、土木用語を十分に加味したテキストマイニングができるように、土木用語を含む形態素解析の辞書を作成した。トピック分析による適用事例の結果、従来の形態素解析の辞書に含まれていない土木用語を加味して分析していることを確認した。

今後は、課題となった意味を持たない単語の除去について研究を進める予定である。

参考文献

- 1) 岩見麻子, 宮下知己, 井手慎司: 大規模パブリックコメントの論点把握に対するテキストマイニングの有用性の検討, 土木学会論文集 G (環境), Vol.71, No.6, pp.Ⅱ_13-Ⅱ_21, 2015.
- 2) 傳田正利, 萱場祐一: テキストデータマイニングを用いた河川整備計画分析に関する研究—主要直轄河川を対象とした試行的分析—, 第54回土木計画学研究発表会講演集, pp.197-201, 2016.
- 3) 堀智仁, 吉川直孝, 大幡勝利: テキストマイニングによる掘削用機械の労働災害分析の試み, 土木学会第66回年次学術講演会, pp.203-204, 2011.
- 4) 奈良先端科学技術大学院大学 自然言語処理研究室: NAIST Japanese Dictionary, <<https://ja.osdn.net/projects/naist-jdic>>, (入手 2019.6.10) .
- 5) Sato, T. : Neologism dictionary based on the language resources on the Web for Mecab, <<https://github.com/neologd/mecab-ipadic-neologd>>, (accessed 2019.6.3).
- 6) 中川裕志, 前田朗, 小島浩之: 専門用語 (キーワード) 自動抽出システム, <<http://gensen.dl.itc.u-tokyo.ac.jp>>, (入手 2019.6.3) .
- 7) 森山聡, 吉田稔, 中川裕志: 複合語のパープレキシティに基づく重要語抽出法の研究, 言語処理学会年次大会発表論文集, pp.408-411, 2005.
- 8) 工藤拓: MeCab: Yet Another Part-of-Speech and Morphological Analyzer, <<https://taku910.github.io/mecab>>, (入手 2019.6.12) .
- 9) Řehůřek, R. : gensim, <<https://radimrehurek.com/gensim>>, (入手 2019.6.3) .