氏　　名	田中　るみ子
学位の種類	博　士（情報学）
学位記番号	博甲第　10419　号
学位授与年月日	令和　4　年　3　月　25　日
学位授与の要件	学位規則第４条第１項該当
審査研究科	図書館情報メディア研究科
学位論文題目	日本語文章からの化学物質名の抽出に関する研究 ：特許公開公報を対象にした検討

主　査	筑波大学	教授	博士（学術）	中山　伸一
副　査	筑波大学	教授	博士（教育学）	芳鐘　冬樹
副　査	筑波大学	教授	博士（情報科学）	真榮城哲也
副　査	筑波大学	准教授	博士（情報学）	高久　雅生
副　査	豊橋技術科学大学	教授	博士（理学）	後藤　仁志

論　文　の　要　旨（2,000 字程度）

本学位論文で述べられている研究は、日本語で書かれた様々な化学文章から化学知識を自動抽出するための基本的要素となる、日本語文章からの化学物質名の抽出方法を確立することを目的としている。

著者は、化学知識の自動抽出の重要性を指摘し、その中核となると考えられる化学のファクトデータベースの自動構築において、文章からの化学物質名の抽出は欠かせない要素であると主張している。そして英語文章を対象とした化学物質名の抽出についての研究が、BioCreative のようなワークショップにおいて大規模なコーパスを提供して活発に行われているのに比べて、日本語の文章から化学物質名を抽出する研究は、単語切り出しの困難さ、文字種の膨大さやコーパスの未整備などの理由により、余り行われていないと分析している。また、日本語で書かれた化学領域の論文や特許などは多くあり、そこから新たな知識の抽出が期待できると提唱している。

このことから著者は、日本語の文章から化学物質名を抽出する方法の確立が重要であると考え、日本語の文章から化学物質名を含めた単語を切り出す方法と、得られた単語群から化学物質名を識別する方法について提案している。そして日本語で書かれた特許公開公報を対象に化学物質名をタグ付けしたコーパスを作成し、提案した方法を適用して、その妥当性を明らかにしている。

本学位論文は 5 章からなり、第 1 章では化学知識の自動抽出の必要性を述べたうえで研究目的の設定を行い、英文による研究事例を含めた関連する先行研究を紹介している。そして第 2 章では先行研究を踏まえた上で、日本語から化学物質名を抽出する流れとして、化学物質名を含む単語の切り出し、得られた単語群からの化学物質名の識別という二段階の方法を提案し、合わせて対象とする日本語文章として特許公開公報を用いることを述べている。

第 3 章では、提案した流れにそって詳細な方法論の提案とその妥当性の検討結果につ

いて述べている。まず、方法の適用対象とする公開特許公報 50 件について化学物質名にタグ付けをしたコーパスの作成方法を述べ、得られたコーパスには 15,834 個の化学物質名のタグがあったことを報告している。

次に化学物質名を含む単語の切り出しの段階の方法として、形態素解析による形態素の分離と得られた形態素の連結を提案している。そして一般的に使われている形態素解析システムを特許公開公報に適用した際、正確に形態素が生成できていない場合があるという問題を示し、日化辞 Web から得たデータを使ってユーザー辞書を作成することによりその問題をクリアしたことを報告している。さらにタグ付けした化学物質名を分離しないひとまとまりの形で得る方法として、形態素の品詞に着目し、記号、接頭詞、名詞が連続して出現した場合にそれらを連結して一つの単語とすることを提案し、作成したコーパスの全ての化学物質名が分割されないことを確認している。ただし、提案した連結方法では化学物質名の前後に余計な文字列がつながってしまう場合があることを示し、どのような文字列がつながるかを分析して、それを分離するための 12 の方法を提案している。その結果、総出現数 14,486 個のうち 80.4%の化学物質名を切り出すことができたと報告している。

そして最後に、得られた単語群から化学物質名を識別する方法として、ルールベースと機械学習による検討について述べている。まずルールベースの方法については、化学物質名に特有の文字を分析してその文字の有無で判別する 1-gram による方法について検討を行い、一部の化学物質名については有効に識別できることや官能基名が混入しやすいという特性を明らかにしたが、全ての化学物質名の識別法として用いるには難しいと考察している。機械学習の方法については、Word2Vec を用いて単語をベクトル化するという方法を提案し、ベクトルの次元数、最小単語出現数、ウインドウ幅を種々変化させて得た幾つかの単語ベクトルを用いた場合について、決定木、ランダムフォレスト、LightGBM による識別を検討している。その結果から、この方法により F 値が 1.0 と非常に高い精度で化学物質名を識別できることを明らかにしたとしている。さらに、大規模なデータを対象にした場合の識別能力について検討するため、コーパスを作成した 50 の特許公開公報を含む 507 公報を使って単語のベクトル化を行い、そのベクトルをコーパスを作成した 50 公報の単語に付与した場合について、化学物質名の識別精度を検討している。その結果から、このやり方でも、F 値が最低 0.68 とある程度の化学物質名の識別ができるとしている。そしてこのことから、大規模なコーパスを利用した場合も、この方法が適用できると主張している。

第 4 章では、提案した一連の方法についての検討結果から、この方法についての問題点を指摘し、その改善についての提案をしている。その上で第 5 章では、提案した一連の方法が妥当なものであると結論付け、海外で行われているような大規模なコーパスを用いたワークショップを開催することなどにより、この研究領域が活性化する必要があると主張している。

審　査　の　要　旨（2,000 字以上）

【批評】

著者は、化学知識の自動抽出という課題を大きな目標として掲げた上で、日本語文章からの化学物質名の抽出を目的として設定し、その方法論を確立することを研究目的としている。化学物質名の抽出は、データ駆動型科学の中心的役割を果たすファクトデータベースの自動構築につながるものであり、また日本語で書かれた化学に関する文章には英語文章では得られない知識も含まれると想定されることから、この研究目的には大きな意義が認められる。以下で、著者の提案する日本語文章からの化学物質名を含めた単語の切り出しの方法と、得られた単語群からの化学物質名の識別の方法を中心に批評を行う。

二つの方法の批評に先立って、この研究で用いた特許公開公報のコーパスについて述べる。対象を特許公開公報にしたことについては、電子的なテキストが手に入るうえに、BioCreative V の CHEMDNER タスクで特許文章が使われた理由にあるように、論文等に比べて使われる単語の意味にぶれが多いため、特許文章で確立した方法は多様な文章にも適用できるというメリットを持つ。よって、特許公開公報を方法論の適用対象とした選択は妥当であるといえる。コーパスの作成については、著者一人によるもので信頼性に若干疑問が残る。信頼性の疑義は、余計な文字列を除く際の検討の中で一割ほどのタグが除かれたことからも生じる。しかし、BioCreative のコーパス作成における人間のアノテータ間の合意比率が 91%であったことを考慮すると、これらの信頼性についての問題は、方法論の検討における結果に大きな影響を与えるほどのものではないと考えられる。

次に日本語文章からの化学物質名を含めた単語の切り出しの方法について論評する。著者は単語切り出しについて、形態素解析による形態素の分離と得られた形態素の連結という二段階による方法を提案している。単語の区切りにスペース等が無い日本語文章において、形態素解析によって形態素に分離することは必然であると考えられる。しかし、実際に適用すると不適切な形態素が出現するという結果が得られた。著者は、それが化学物質名に特徴的な単語が辞書に無いことによるものと気づき、日化辞 Web から得たデータを使ってユーザー辞書を作成することによりこの問題を解決しようとしている。ユーザー辞書は更新の必要があり、特に商品名などの命名法に従わない化学物質名では依然同様の問題が生じる可能性も考えられるが、今回対象とした化学物質名については問題が解消されたことから、妥当な改善策であると評価できる。形態素の連結については、記号、接頭詞、名詞が連続していた場合に一つの単語とすることを提案し、その方法で化学物質名が分割されないことを確認している。しかし、余計な文字列がつく場合がタグ付けした化学物質名の四割ほどで起こるという結果が得られた。形態素の連結という方法を用いると、この結果は容易に予想されることである。著者は余計な文字列の分析を行い、その分離の方法を 12 個提案して、二割ほどまでその割合を減らしている。提案された分離の方法は妥当なものであり、評価できる。しかし、それでも二割の化学物質名について分離できていない点については、さらなる検討が求められるところである。

さらに、得られた単語群からの化学物質名の識別の方法について論評する。著者は、ルールベースの方法と機械学習の方法を適用することを検討している。ルールベースの方法としては、化学物質名固有の文字を分析することによる 1-gram の方法を提案している。1 より大きな n-gram も考えられるが、文字種の多い日本語においては余り現実的ではなく 1-gram は適切な選択といえよう。検討の結果、1-gram の方法による結果は高い F 値が得られず、方法論として適用が困難であるという結果を得ている。これはある程度予想できるものであり、研究全体の中でのこの検討の意義は余り大きくないと言えよう。機械学習による方法は、CHEMDNER タスク等でよく使われ、高い精度を出していることから、それの適用を検討することは妥当といえよう。その際、単語を Word2Vec を用いてベクトル化する方法を用いているが、これも通常使われる方法で、妥当ではあるが新規性は余り認められない。機械学習の結果として 1.0 という F 値を得ているが、これはタグづけされた化学物質名の二割に余計な文字列が付随していて化学物質名として扱われていないことを考慮すると、高すぎる値と考えられる。これは、今回用いた決定木などの手法が、比較的少ない対象に対して過学習を起こした可能性を示唆する。著者はロジスティック回帰の適用を試みたがうまくいかなかったと述べているが、汎化能力の高い他の手法の適用についても検討するべきであったと考える。また、識別がうまくいかなかった事例の詳細な分析も今後必要であろう。

最後に、コーパスを作成した 50 公報を使った単語ベクトルに代えて、その 50 公報を含む 507 公報を使ってベクトルを作成し、それを 50 公報の単語に付与して機械学習の方法を適用するという検討について論評する。一般に Word2Vec による単語のベクトル化は、非常に大規模な文章を対象にして行われており、50 公報に含まれる単語数は少なすぎる懸念がある。その意味で、507 公報を使ったベクトル化は、単語のベクトルをより正確に作成していると考えられる。この検討である程度高い F 値を得ていることは、機械学習による識別の方法が妥当であることを示唆しており、大規模コーパスによる今後の方向性を考える上で、意義のある知見を与えているといえる。

著者は、日本語文章から化学物質名を抽出する方法を確立するという研究目的に対して、化学物質名を含む単語の切り出しと得られた単語群からの化学物質名の識別という二段階の方法を提案し、特許公開公報を用いて検討した結果から、その方法が妥当であることを明らかにしたとしている。そして、海外で行われているように大規模なコーパスを作成してワークショップを開催することなどにより、この研究領域が活性化する必要があることを主張している。コーパスの規模がさほど大きくないことや、余計な文字列の分離で二割ほどの化学物質名が切り出されていないことなど課題はあるが、困難だが重要な研究目的に対して適切な方法論を提案して、今後のこの領域の進展に有用な成果をあげていることから、学位論文として十分な内容を持つと評価できる。

【最終試験結果】

2022 年 2 月 18 日、図書館情報メディア研究科学位論文審査委員会において、審査委員全員出席のもと、本学位論文について著者に説明を求めた後、関連事項について質疑応答を行った。引き続き、「図書館情報メディア研究科博士後期課程(課程博士)の学位論文審査に関する内規」第 23 項第 3 号に基づく最終試験を行い、審議の結果、審査委員全員一致で合格と判定された。

【結論】

よって、本学位論文の著者は博士(情報学)の学位を受けるに十分な資格を有するものと認められる。