産業日本語と文書検索
類似文書検索技術の現状と産業日本語への期待

株式会社日立製作所　システム開発研究所
間瀬　久雄

PROFILE 平成2年(株)日立製作所入社、システム開発研究所に配属。以来、特許や新聞記事、Webページ等を対象とした、分類自動付与、検索、文章要約、テキストマイニング等の日本語処理の研究に従事。2007年度から特許版産業日本語委員会／技術用日本語プラットフォーム委員会委員。2007年度から特許版産業日本語委員会／技術用日本語プラットフォーム委員会委員

hisao.mase.qw@hitachi.com

1 はじめに

我々の社会生活において、文書検索技術の果たす役割はこの10年で飛躍的に高まった。インターネット検索エンジンや、特許電子図書館IPDLなど、大量の文書の中から所望の文書を検索する環境は普及してきたと言える。

しかし、検索対象となる文書件数が多くなるほど、また、文書の内容が多種多様化するほど、所望の文書を見つけるのに、より多くの時間がかかっているのが現状である。この最大の原因は、人間が作成した文章を計算機が正確に解析・理解し、処理することができないために、検索結果に漏れが生じたり、ノイズが混入したりしてしまうことにある。人間が作成する文章は、その人の背景知識や文章作成に係るスキルや嗜好に依存する。どのような構成の文章とするか、どの語彙を選択してどのような構文で文を記述するかによって、意味的に同じ内容でも、全く異なる文章が作成される。人間はこれらの文章の意味を無意識のうちに同定できるが、機械にはそれができない。

この課題を解決すべく、産業日本語では、構文や意味に関する言語的曖昧性を排除し、人間だけでなく計算機にとっても解析・理解しやすい文章の作成を促進することにより、計算機の処理精度を向上させることを目指している。そこで本稿では、文書検索の視点から、産業日本語に対する期待について、私見を述べることとする。まず最初に、産業日本語の視点から見た、文書検索と機械翻訳の違いについて述べる。次に、文書検索技術の現状と課題について簡単に触れる。そして、文書検索の精度向上の観点から、産業日本語に対する期待と想定される技術課題について述べる。

なお、文書検索には、検索条件をキーワードと論理演算子（and/or/not）から構成される論理式として記述し、この論理式を満たす文書を検索するブーリアン検索（全文検索）と、任意の文や文章、あるいは文書全体を入力とし、内容の類似する文書を類似度の高い順に検索する自然言語文検索（概念検索）がある。本稿では、産業日本語の効果がより顕著に現れると期待している自然言語文検索に焦点を絞って考察することとする。

2 産業日本語から見た文書検索と機械翻訳の違い

産業日本語では、想定している計算機処理アプリケーションとして、機械翻訳と文書検索を挙げている。ここでは、産業日本語を適用した時の効果という観点から、両者の違いについて考察する。

機械翻訳は、基本的には文単位の処理である。翻訳対象となる文（原文）に、構文的または意味的な曖昧性があれば、翻訳精度は低下する。しかし逆に言えば、産業日本語によって明晰な原文を記述できれば、その分だけ機械翻訳精度は確実に向上する。このことから、産業日本語と機械翻訳の間の親和性は高いと言える。

これに対して、文書検索（自然言語文検索）は、基本的には文章単位の処理である。文章と文章を照合してその類似性を判定する。したがって、理想的な自然言語文検索システムでは、単語単位や文単位で照合するだけでなく、文章構造や文間のつながりといった文章単位で照合することも必要となる。同じ内容の文を一文共有しているだけでは、文章が類似しているとは言えない。その文がどういう文脈の中で記述されたのか（文章の中でその文がどういう位置付け・役割にあるのか）を理解した上で文を照合し、内容の類似性を判定しなければならない。しかし残念ながら、現実の自然言語文検索システムでは、文単位や文章単位の照合はなされていない。単語単位での照合によって類似性を判定するに留まっている。

単語単位の照合で見ると、産業日本語による構文レベルの明晰化は、機械翻訳の精度向上には非常に有効であるが、自然言語文検索の精度向上にほとんど寄与しない。

例えば、以下の二つの文を考える。

（文1）「文と単語の意味を解析する手段」

（文2）「文の意味と単語の意味を解析する手段」

文1では、「意味」に係るのが「単語」だけなのか、「文」と「単語」の両方なのかが曖昧である。その結果、機械翻訳では誤った翻訳結果を出力する可能性がある。一方、文2では、この曖昧性が解消された記述となっているため、正確な機械翻訳結果を出力できる。

しかし、単語単位の照合に基づく自然言語文検索では、文1から抽出される単語と、文2から抽出される単語はどちらも、「文」「単語」「意味」「解析」「手段」の5種類であり変わらない。その結果、文2の方が明晰な文であるにもかかわらず、自然言語文検索結果はほとんど変わらない。このように、産業日本語の適用効果を、自然言語文検索に対して最大限に反映させるためには、単語単位で行っている現状の検索方式を見直す必要がある。

また、自然言語文検索と機械翻訳のもう一つの大きな違いとして、大量の検索対象文書の存在がある。産業日本語の適用によって自然言語文検索の精度を改善させるためには、検索される側の文書集合についても明晰な文章で記述しておく必要がある。入力側の文章だけをいくら明晰に記述しても、検索される側の文章が明晰でなければ、検索精度の改善は期待できない。

3 現状の文書検索の実力と課題

本章では、自然言語文検索の現状の実力について、公開特許公報を検索対象とした類似特許検索に焦点を絞って簡潔に述べる。

3.1 現状の類似特許検索技術の精度

表1のデータは、国立情報学研究所が主催している情報検索の国際ワークショップである第6回NTCIR（NTCIR-6）[1]の特許検索タスク[2]において、筆者を含む日立チームによる検索結果の精度を評価した結果[3][4]である（この検索結果は、NTCIR-6で報告された結果の中で最も精度の良い方式によるものである）。公開特許公報を入力文章とした時に、現状の自然言語文検索技術によって、その発明内容に類似する過去の特許を、どのくらいの精度で検索できるかを示している。なお、検索対象は公開特許公報10年分（約350万件）である。また、特許庁審査官が入力特許を拒絶する際に引用した特許を正解として評価している。

表1は、評価データ1,189件（延べ正解件数2,065件）について、正解特許が検索結果の何位に出力されたかの

表1　自然言語文検索による特許検索精度の現状

#	検索結果特許件数	含まれる正解特許	
		件数	割合
1	上位　1件	207件	10.0%
2	上位　10件	570件	27.6%
3	上位　20件	726件	35.2%
4	上位　50件	973件	47.1%
5	上位 100件	1191件	57.7%
6	上位 200件	1382件	66.9%
7	上位 300件	1484件	71.9%
8	上位 500件	1601件	77.5%
9	上位1000件	1750件	84.7%

件数割合を示している。これによると、上位20位（主要なインターネット検索エンジンの出力結果1ページ分に相当）までに出力された正解特許件数の割合は35.2%、上位200位（特許庁審査官が検索によって絞り込む目安と言われている件数）まででは66.9%である。

このように、現状の自然言語文検索技術によって、検索結果の上位に出力可能な正解特許もあるが、その割合はまだまだ低い。

3.2　現状の類似特許検索技術の課題

現状の自然言語文検索では、文章中に現れる単語の出現傾向を重み付けし、重み付き単語集合の類似性に基づいて文章間の類似性を判定する。各単語の重みは、その単語がその文章中に現れる頻度と、その単語が検索対象となる文章集合の中に現れる頻度の2種類に基づいて算出されることが多い。

このように、現状の自然言語文検索では、文章単位や文単位で類似性を判定する代わりに、文章中に現れる単語を単位として類似性を判定する。この場合、どの単語を検索に使用するか、各単語の重みをどのような基準に基づいて算出するかによって、検索精度が変動する。したがって、検索精度を向上させるためには、検索に使用する単語の抽出精度と重みの算出精度を向上させることが技術課題となる。

この技術課題を解決すべく、特許文書を対象とした検索においては、単語を抽出する範囲を特許明細書タグ単位で絞り込んだり、複数のタグに出現する単語の重みを高くしたり [3] [4]、請求項の内容に関連する記載がされている本文中の文章に出現する単語を重要視したりするなど、特許の文章構造や特許固有の構文に着目した方式が多く提案されている。しかし、単語を単位として類似性を判定するという枠を超えた高精度の類似特許検索方式は、ほとんど実現されていない。

また、単語の表記の違いを吸収するためには、同義語辞書が必要となる。特許については技術分野が多岐にわたっており、技術分野固有の用語も日々増えていくので、静的な辞書を整備するだけでなく、特許文章の中から同義語を自動抽出する、動的な辞書構築技術の確立も技術課題の一つである。

4　産業日本語への期待

4.1　検索精度改善への貢献

2章で述べたように、単語単位での自然言語文検索では、産業日本語の適用による検索精度向上はほとんど期待できない。産業日本語が自然言語文検索の精度向上に貢献するためには、単語集合を用いた現行の自然言語文検索技術を超えた「次世代自然言語文検索技術」として、文または文章を単位とした類似度判定に踏み込んでいく必要があると考える。

産業日本語の適用により、構文的または意味的に曖昧のない文が計算機処理の対象になると想定すると、その文を解析して、主語・目的語・述語といった格関係や、その間の係り受け関係を、計算機によって高精度に抽出できるようになる。また、主節と従属節を区別したり、文と文との間のつながりを特定したりする解析も精度良くできるようになる。

これらの構文解析を高精度でできるようになると、以下で挙げるような、自然言語文検索の精度を向上させる方式の実現が期待できる。

（1）格関係を用いた類似性判定

図1（a）に示すように、現在の自然言語文検索では、単語を単位としてその出現頻度から重みを算出して単語ベクトルを生成し、ベクトルがどれだけ類似しているかによって文書間の類似性を判定する。

この考え方を単文単位に拡張すると図1（b）となる。ここでは、個々の単語をベクトルの要素とする代わりに、単文を構成する主語・述語・目的語の三つ組を要素とする一つの「概念」とみなす。そして、この概念が文章中に現れる頻度を算出して、概念の重みとする。類似性を判定する際には、この概念を単位として照合する。この際、概念の完全一致および部分一致を考慮した柔軟な概

念照合アルゴリズムを導入する。

このような単文を単位とした照合方式は、既にいくつか研究がなされている。産業日本語はこれらの処理精度を向上させるものとして期待できる。

(2) 文のタイプの選別による類似性判定

曖昧性の排除された文では、その文のタイプを特定しやすくなると考える。ここでいう文のタイプとは、事実を表す文、原因・理由を表す文、結果を表す文、結論を述べている文などを指す。また特許文書について言えば、発明の課題、目的、効果などを表す文を指す。

産業日本語を適用することによって、文のタイプを容易に識別できるようになると考える。その結果、例えば、「PCの筐体を軽くする」という、目的を表す文を入力文章とした場合に、検索対象となる個々の特許文章の中で、目的を表す文を特定し、さらにその中から「PCの筐体を軽くする」という記載の特許を選定することが低ノイズでできるようになる。

文のタイプの選別に関する研究としては、特許マップの自動作成において、特許文章から課題や解決手段に相当する記載箇所を抽出する研究がある。産業日本語はこれらの抽出精度を改善し、検索精度を改善する役割を果たすものとして期待できる。

(3) 請求項の構造解析の精緻化による類似性判定

特許のように、固有の文章構造を持つ文書を検索対象とする場合、入力特許の請求項の文章構造を解析して、各々の構成要素、構成要素間の階層関係、発明の特徴（新規性または進歩性）に相当する記載箇所などを高精度に特定できれば、類似する過去の特許を検索する精度を改善できると考える。そのためには、計算機によって請求項文章を正確に構造化する機能が不可欠である。産業日本語の適用によって、構成要素の特定が容易となり、発明内容をより正確に把握することが可能となる。

4.2 想定される技術課題

4.1節で挙げたような検索精度向上方式を実現するためには、いくつかの難しい技術課題を克服しなければならない。産業日本語によって、曖昧性のない文で文章が記載された場合、形態素解析や構文解析は高精度でできるようになる。しかし、単語辞書の整備や複合語の解析、類似性判定のロジック、現実的な検索速度を実現するための検索アルゴリズムなど、自然言語文検索の精度や性能に直接影響を与える処理において技術課題が存在する。以下、これらの技術課題の各々について考察する。

(1) 単語辞書の整備

産業日本語では、語彙レベルの曖昧性についてはあま

＜入力文章＞

【発明の名称】 単語重み算出方法
【請求項1】 入力文章を形態素解析し、文章に出現する単語を特定し、前記特定された単語の出現頻度を算出し、当該単語が検索対象文書集合を構成する文書の何件に出現するかを算出し、前記出現頻度と前記出現文書数から当該単語の重みを算出することを特徴とする単語重み算出方法。

（a）単語単位でのベクトル

単語	6	算出	5	重み	3
出現	3	前記	3	文書	3
方法	2	文章	2	特定	2
出現頻度	2	当該	2	入力	1
形態素解析	1	検索	1	対象	1
集合	1	構成	1	何件	1
数	1	こと	1	特徴	1
する	1				

（b）格情報を用いたベクトル

算出（ー，重み（単語））	2
形態素解析（ー，入力文章）	1
特定（ー，単語）	1
算出（ー，出現頻度（単語））	1
出現（単語，ー）	1
構成（ー，検索対象文書集合）	1
・・・	

図1 産業日本語の適用による精度向上が期待できる、格情報に基づく文章解析

り言及されていない。一般に、文に含まれる単語が一般語なのか、専門用語なのか、著者が独自に定義した語なのかによって、形態素解析や構文解析の精度は大きく変わる。したがって、検索対象となるドメインに関する単語を収集した単語辞書を予め整備しておく必要がある。

また、単語を正確に特定できるだけでなく、同義語や類義語を特定できなければ、検索精度に悪影響を及ぼす。4.1節（1）で述べたように、文単位や文章単位といった、単語よりも上位のレベルで照合を行うためには、その構成要素である単語単位での意味的同定が正確になされることが大前提となる。単語単位での同定ができなければ、文や文章単位の同定においてノイズがノイズを呼ぶことになるため、検索精度は改善しない。

一般に、単語辞書や同義語辞書を構築・維持していくには多大なコストがかかる。そこで、検索対象文書の中から辞書に未登録の単語や同義語を自動抽出する技術が注目されている。特許からの同義語抽出では、文章中の括弧表現や、単語の共起関係を手掛かりとして、同義語や関連語を自動抽出する研究がある。産業日本語によって、同義関係にある単語対を抽出する精度が改善できるかについては、未知数の部分が大きいが、学術的には非常に興味深い。

（2）複合語の解析

日本語の文章の中で使われる複合語の扱いが、検索精度に与える影響は大きい。特に特許では、複合語が多用される傾向が強い。4.1節（1）で述べた、格関係に基づく類似性判定を行う場合、例えば「情報検索」が「情報を検索する」と同じ意味であることを解釈する必要がある。現在でも、単語の持つ品詞や意味属性から、複合語を構成する単語間の格関係を推定することがある程度可能となっているが、精度の高い検索精度を実現するためには、複合語の解析にもより高い精度が要求される。産業日本語では、複合語による記載について言及しているが、自然言語文検索の観点から見た場合、計算機がその構成を解析できないような複雑な複合語については、使用を避けるようにすべきであろう。

（3）類似性判定のロジックと処理性能

一般に情報検索では、精度を向上させようとすると、処理性能が劣化する。産業日本語の適用によって、検索精度の向上が期待できるが、処理性能の維持についても並行して検討していく必要がある。

類似性判定のための照合単位を、単語から文・文章に拡張することにより、文章解析にかかる時間および類似性判定にかかる時間が長くなることが予想される。特に、類似性判定では、部分一致を含めた照合を高速に行うためのロジックが必要となる。例えば、4.1節（1）で挙げた格関係に基づく類似性判定では、主語・述語・目的語の三つ組同士の照合を柔軟かつ高速に行わなければならない。また、実際の文では主語・述語・目的語を修飾語する単語が多く存在するが、類似性判定においてこれらの修飾語をどのように扱うかによって、検索精度と検索速度の両方に多大な影響を与えることになる。

5 おわりに

本稿では、文書検索（特に自然言語文検索）の観点から、産業日本語に対する期待について私見を述べた。文章記述の多様性に起因する自然言語処理の技術的限界を一歩踏み越えた「次世代文書検索システム」を実現するにあたり、産業日本語は重要な役割を担ってくると考える。4.1節で挙げた、産業日本語の適用に基づく検索方式はあくまで例であり、これらの他にも検索精度を向上させる施策はいろいろあるだろう。

産業日本語が文書検索にどのように貢献できるかの具体内容については、筆者もメンバーとなっているJapio主催の「特許版・明晰日本語策定委員会」及び「技術用日本語プラットフォーム委員会」においても、これまであまり深く議論がなされていない。今後は、この議論を深めるとともに、産業日本語をベースとした次世代文書検索技術に対して、産官学が一体となって取り組むため

の組織体制を早急に組み、研究開発を推進する必要があるだろう。

参考文献

[1] Kando, N. : Overview of the Sixth NTCIR Workshop, Proceedings of the Sixth NTCIR Workshop Meeting, pp.i-ix（2007）.

[2] Fujii, A., Iwayama, M. and Kando, N. : Overview of Patent Retrieval Task at NTCIR-6, Proceedings of the Sixth NTCIR Workshop Meeting, pp.359-365（2007）.

[3] Mase, H. and Iwayama, M. : NTCIR-6 Patent Retrieval Experiments at Hitachi, Proceedings of the Sixth NTCIR Workshop Meeting, pp.403-406（2007）.

[4] 間瀬久雄，岩山真：特許の文書構成と分類情報を用いた類似特許検索方式の精度評価，Japio 2007 YEAR BOOK, pp.166-171（2007）.