特許文書を活用するための文書診断技術

Document Diagnosis Technology for Utilizing Patent Documents

東芝ソリューション株式会社　プラットフォームセンター参事　**熊野　明**

PROFILE　1982年東京工業大学卒業。同年東京芝浦電気(株)（現(株)東芝）入社。2010年から東芝ソリューション(株)。自然言語処理システムの研究開発に従事。アジア太平洋機械翻訳協会理事。AAMT/Japio特許翻訳研究会委員。Japio特許版・産業日本語委員会委員。

kumano.akira@toshiba-sol.co.jp

1　はじめに

特許文書、一般に特許明細書は、多くの人が読むことのできる技術文書の一例である。新しい技術を開示し、その権利を主張するために、さまざまな技術分野で作成、公開されている。

特許として登録されるための条件の一つは、権利主張する技術に対して、当業者が正しく理解できることである。これは、技術的に実現可能な手段を開示することを意味するが、そのためには、主張する権利が明確でなければならない。すなわち、読み手によって権利内容に異なる解釈が生じてはいけない。技術文書一般の性質であるが、特に特許は、この性質が重要である。

特許文書には、その性質上、他の技術文書と比較して際立った特徴がある。

（1）　新規用語の新規性

新技術を説明するために、これまでにない新しい用語が使われることが多い。その多くは、既存の構成語を複数連続した複合語である。

新用語のうち、これまでにまったくない概念を表すものもある。新概念を説明するための造語であることが多い。

（2）　文の複雑さ

新しい手段、装置、概念を表現するために、多くの用語を用いる必要がある。結果的に、文の構成を複雑にしてしまう。

年々大量の特許文書が公開される状況では、人による解釈だけでなく、計算機による自然言語処理を用いた解釈が必須である。このようにして、科学技術の発展のために、幅広く活用するべきである。

自然言語処理による文書の解釈には、検索、情報抽出、要約、機械翻訳などがあり、現在までに多くのシステムが開発され、広く利用されている。しかし、実際の特許文書は必ずしも現在の自然言語処理で完全に解釈できるものではないことも事実である。

これを解決し、各種のシステムが精度を向上させ、利用者の満足度を上げるには、自然言語処理技術の発展とともに、文書作成のルール化が必要である。

日本特許情報機構が提案している特許版・産業日本語の取り組みはその一つであり、これまでに、特許文書のライティングマニュアル[1]を作成・公開するとともに、いくつかのサポートツールを評価・検証している[2]。また他の実験では、機械翻訳の精度向上につながることが示されている[3]。

本稿では、このような背景をもとに、特許文書の活用を推進するための文書診断技術について述べる。

2　特許文書が満たすべき性質

前章で述べた背景をもとに、特許文書として望ましい性質を整理する。

2.1　文単位での性質

（1）　用語の正確さ

専門用語を含む多くの用語が、正しい表記で正しい用法で使用されていることである。

(2)　曖昧性の軽減

1文が表現する内容に曖昧性がなく、読み手によって異なる解釈を生じないことである。

2.2　文書全体での性質

(1)　用語の一貫性

文書全体において、同じ概念を表す用語は同じ表記で、また区別すべき概念を表す用語は異なる表記で表現することである。

(2)　符号や番号の対応

本文と図面とで対応する要素に対して、同じ符号が付与されていること。同様に、本文とフローチャートで、対応する処理に同じ番号が付与されていることである。

3　文書を検証するための技術

検証のための基本技術は自然言語処理技術であり、その主なものは、形態素解析と構文解析の結果を応用したものである。

3.1　大規模な専門用語辞書

用語の正確さを確保するためには、大規模な専門用語辞書が必要である。機械翻訳システムをはじめとする自然言語処理システムには専門用語辞書を搭載しているものが多い。

しかし、特許文書では日々新しい用語が生まれている。これらの新語を効率的に収集し、辞書として利用可能にする手段が必要である。

3.2　形態素解析技術

1文に含まれる用語、表現を解析する技術である。一般に、文の長さが長くなると、曖昧性が増し、正しい解釈が困難になる。そこで、一定文字数を超える分に対して警告を与えることが有効である。

しかし、同じ文字数の文でも、用語の構成は様々である。(1)個々の用語が短く、全体として多くの用語を含む場合。(2)長い用語を多く含み、全体の語数は多くない場合。当然、(2)より(1)のほうに、曖昧性が多い。したがって、文字数だけでなく、語数を正確に解析する手段が必要である。

3.3　構文解析技術

1文を構成する用語の間の、係り受けや修飾関係など、構造に関する情報を解析する技術である。

文の構成する要素が多くなると、係り受け、修飾関係の解釈に曖昧性が生じる。さらに、表現によっては、並列関係の解釈にも曖昧性が生じる。

意味的な制約を利用して、機械による解釈を絞り込むことも可能だが、人間の行うレベルの解釈を実現することは難しい。

したがって、複数の可能性がある場合、それを何らかの方法で利用者に提示できる手段が必要である。

3.4　一貫性検証技術

文書全体での情報の一貫性を調べる技術である。

表記や表現の揺れを認識し、その結果を提示する。3.2の形態素解析結果を利用して実現する。

3.5　特許文書解析技術

特許文書の構成を利用した分析・検証である。

請求項と本文との対応、本文と図やフローチャートとの対応など、人手による特許文書作成で生じやすい誤りを、可能な限り機械的に検証する。

さらに、複数の請求項の依存関係を分析し、提示する機能を実現したものもある。

4　文書診断技術

2章で示した性質を文書の可読性とみなし、3章で示した技術を利用したものが、可読性診断技術[4]である。本稿では文書診断技術として紹介する。

この文書診断技術は、1文単位での可読性を診断するものである。文書全体の整合性を診断するものは、他のシステム[5]を参考にしてほしい。

4.1　診断機能

文書診断技術として実現している診断項目には、様々なレベルがある。その中から、主なものを以下に示す。

(1) 長文

一定文字数を超える長い文を指摘し、短文化を促す。

(2) 曖昧な係り受け

係り受け関係の解釈が複数ある箇所を指摘し、解消する書き換えを促す。

(3) 主語の省略

主語が省略されている述語動詞を指摘し、主語の明示化を促す。

(4) 述語の省略

述語動詞が省略されている箇所を指摘し、述語の明示化を促す。

(5) 一定数を超える述語

1文に述語動詞が多く含まれていると、解釈の曖昧性が大きくなる。このような文を指摘し、分割などの書き換えによって曖昧性を解消することを促す。

(6) 助詞「は」

格の曖昧性がある助詞「は」を指摘し、格を明確に示す書き換えを促す。

これらの診断は、いずれも次のような手順で実現している。

(1) 対象の文を形態素解析・構文解析する。

(2) 各診断項目に対応して準備した診断基準に対して、(1)の解析結果が適合しているか調べる。

(3) 診断基準に適合した場合、その内容を提示する。

(4) 可能な場合は、入力文に対する言い換え候補を提示する。

4.2 特許明細書に対する実験

実際の特許明細書を用いて、文書診断の実験を行った。診断結果として多く指摘されたものは、長文と、係り受けの曖昧性であった。

今回の実験では、長文として診断した文に対しては、その他の診断は行わない。したがって、長文を短く分割して書き換えた場合、その結果に対して新たに係り受けの曖昧性があると診断されたものもあった。

これらの実験結果から、特許文書を幅広く活用するための方針を、以下に示す。

(1) 1文には1つの関係や処理を基準する。

特許に記述される文には、多くの用語や処理を含んで長いものが多い。これらはほとんど、複文構造をとっている。主文と従文に分割することで、それぞれを単文化する。2文に分割した場合、因果関係、依存関係は、接続詞などで表現する。

[例]

〈…関係 A…〉なので、〈…処理 B〉を行う。

(単文化 1)

〈…関係 A…〉である。

したがって、〈…処理 B〉を行う。

(単文化 2)

〈…処理 B〉を行う。

その理由は、〈…関係 A…〉である。

(2) 複数の関係をすべて含まないと成り立たない場合は、箇条書きを利用する。

[例1]

次の順に処理する。

(1)…。(2)…。

[例2]

特徴は以下の3点である。

(1)…。(2)…。(3)…。

このように、長文の単文化が有効である。単文化を重視した結果、過剰な言い換えと感じる場合があるだろう。しかし、特許文書の活用を目的と考えるなら、むしろ好ましいと考えるべきである。

5 ライティング支援システム

これまでに述べた特許文書の性質を実現し検証するためには、特許ライティング支援システムが必要である。特許ライティング支援システムとは、利用者が特許文書を作成する際に、専用の機能で望ましい文書の作成をサポートするものである。支援システムが提供するべき機能を考える。

(1) インタラクティブな診断

システムの診断に基づいて、利用者は文書を書き換える。その書き換え結果に対しても、システムはさらに診断を行うべきである。このように、診断と書き換えを繰り返し行うことで、効率的に文書を仕上げることができる。このためにも、インタラクティブな診断インタフェースが必要である。

(2) 診断の根拠を表示

診断の根拠は、入力文の解析結果である。形態素解析、構文解析などを活用するが、その結果を利用者に分かりやすく提示する必要がある。単に結果を列挙しただけでは、専門家以外の利用者には理解できない。

曖昧性を検出した場合、解釈が複数存在する。複数の解析結果には、人間の解釈と一致しないものがありうるが、計算機の解釈の結果を利用者に伝えることが重要である。

(3) 書き換え候補の提案

システムの解釈を参照し、書き換えの候補を提案することができれば、より有効である。

(4) 利用者による編集・選択の操作

書き換えを行ったり、複数候補から一つを選択するのは、最終的に利用者である。言語処理の専門家だけが利用するものではないので、簡易で効率的なインタフェースが望まれる。

(5) システムによる学習

システムの提示に対して、利用者が書き換えや選択を行うが、システムが同じ指摘を何度も行って、利用者が繰り返し同じ操作を強いられるのは望ましくない。利用者の操作履歴をもとに、システムが可能な範囲で学習を行い、利用者の操作回数を減らすことで負担を軽減することが望ましい。

6 おわりに

毎年国内で大量に作成･出願される特許文書に対して、その活用を進めるための手段として文書診断技術を紹介した。また、特許作成に利用できるライティング支援システムの実現すべき仕様を提案した。

今後は、実用的なライティング支援システムの実現を目指すとともに、ライティングルールを含めたプラットフォーム全体の普及を推進していく予定である。

参考文献

[1] 日本特許情報機構“特許ライティングマニュアル(初版)”(2013)
[2] 熊野　明 他“機械翻訳精度を向上させる可読性診断技術”Japio YEAR BOOK 2012 (2012)
[3] 熊野　明 他“産業日本語の構想と特許文の言い換え実験”情報処理学会　第 190 回自然言語処理研究会 (2009)
[4] 祖　国威 他“構文的特性に着目した可読性診断技術”東芝レビュー Vol. 66 No. 4 (2011)
[5] 第５回産業日本語研究会・シンポジウム　発表資料 (2014)