複雑な文に対応した意味構造検索システムの開発

大倉 清司¶　　潮田 明¶¶

¶（株）富士通研究所
¶¶ 奈良先端科学技術大学院大学

1. はじめに

自然文を入力して検索する自然文検索が特許検索などで実用化されているが、その精度は十分とはいえない。その理由の 1 つは、検索対象中に含まれる単語をベースに検索する手法に由来すると考えられる。この手法では単語間の係り受け関係や意味的な関係までとらえられないため、文章にこめられたユーザーの意図を十分に反映した検索は難しい。本稿では単語をベースとした自然文検索をキーワードベース検索と呼ぶことにする。これに対し、文章における単語と単語の関係を表した意味構造を用いて検索する意味構造検索が研究されている[1,2]。しかし、特許文や論文のような長くて複雑な文章を対象とした意味構造検索システムは未だ実用化されていない。今回、意味構造検索において複雑な文にも対応できる技術を開発し、検索性能の向上を図った。その結果、複雑な文を多く含む特許を対象とした検索において、従来のキーワードベース検索より高い精度を実現した。ランキング検索において、正解対象が検索結果上位 200 件に含まれる比率がキーワードベース検索と比較して約 1.5 倍となり、意味構造検索の有用性を確認した。今後は論文・社内文書などの検索に応用していく。

2. 従来の意味構造検索システム

“意味検索”と呼ばれるシステムは数多く存在するが、そのほとんどがキーワードベースのシステムである。しかし意味構造を使って検索をするシステムも存在する[1,2]ため、本稿では、意味構造を使った検索手法をキーワードベースの意味検索と区別して”意味構造検索”と呼ぶことにする。

意味構造検索とは、係り受け構造やグラフ構造などで表される意味構造を検索する手法である。本稿の意味構造検索技術は、当社がこれまでに開発した先行研究[2]の技術をベースに開発した。意味構造は格文法[3]に基づいており、１文の意味構造は、単語の概念を表すノードおよびノード間の関係を表すアークからなる有向グラフにより表される。図 1 は「太郎は花子に本をあげた。」の文の意味構造を有向グラフにより表した 1 例である。

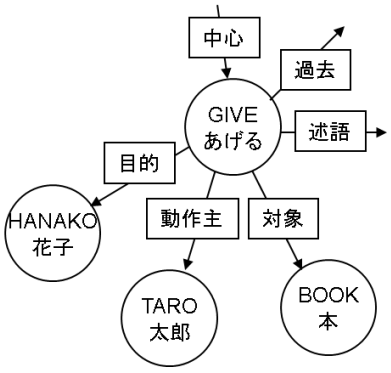

図 1. 意味構造

図 1 において、○で囲まれたものがノードを表し、□で囲まれたものがアークを表す。ノードは、”GIVE”,”HANAKO” ,”TARO”,”BOOK”の 4 つである。これらのローマ字表記は各形態素の意味概念を表す記号であり、意味記号と呼ぶ。アークは”中心”,”目的”,”動作主”,”対象”,”述語”,”過去”の 6 つである。アークはノード間の関係を表す。向かう先にノードがないアークはそのノードの属性を表す。

文書の意味は文書中に含まれる文の意味を足し合わせたものと考えることができる。文の意味を有向グラフとして表現したとき、１文の意味構造検索はグラフマッチングとして、１文書の意味構造検索は文書中に含まれる文の意味構造を検索することと考えることができる。

先行研究において、検索を単純なグラフマッチングとした場合、グラフの複雑性、意味解析処理の精度の問題があり、十分な検索精度が得られないと考えた。そこで、意味構造をアークとそのアークにつながるノード(1 つか 2 つ)の部分グラフに分解し、この部分グラフにより検索する手法を採用した。正確性を重視して、部

分グラフが完全一致するものを検索するようにした。

意味構造の抽出には、日英翻訳エンジンATLAS[4]の翻訳過程から意味構造を取り出すことにした。ATLAS はルールベースの翻訳方式を採用しており、原文を辞書と文法規則に基づき解析して意味構造を計算する。

3. 自然文検索システムの目的

ここで、検索システムが目指すところについて触れておきたい。一般的に広く普及しているキーワードを数個入れて検索するシステムにおいては、もれなく検索するよりも、ランキングの上位 n 位までにユーザーが欲する結果が入ればよい。キーワードの代わりに自然文を入力するときも、もれなく検索したいというよりは、入力した自然文と関連の深いものが見つかればよいと考えられる。

本稿では複雑な文を多く含む特許文書を使い、本手法の検証を行った。特許調査においては、特許専門家が使用するブーリアン検索は、専門家以外にはロジックを立てづらく、十分な公知例調査ができないことが多い。一般の技術者にとっては、自然文検索は、発明の要点を記述したクエリーをつくるだけでよいので、使いやすい。自然文検索により公知例をもれなく調査するのではなく、有力な手がかりとなる特許を見つけることができれば、そこから連鎖的に公知例を探すことができる。なお、特許専門家が公知例調査をする際にも、発明のポイントを自然文で記述することが多い。それをクエリーとして、公知例の有力な手がかりを見つけることができれば、公知例調査の工数を削減することができる可能性がある。

このような背景から、本稿における検索システムの目的を、上位 n 件までにユーザーが意図する文書がランクされることとした。特許の公知例調査の場合、通常は上位数百件を目視するので、ここでは上位 200 位以内の正解文書の再現率を 1 つの指標とし、これを本稿における検索精度と定義することにした。

4. 複雑な文を対象とした検索精度の向上

4.1. 先行研究の手法の問題点と解決手法

先行研究においては、クエリー文の意味構造と正解文書中の文の意味構造のそれぞれの部分構造を完全一致で検索していた。この方法では、一致したときは正解文書である可能性が高いものの、同じ意味を表しているが意味構造が違う場合、正解文書がマッチしない。このため、意味構造の様々なバリエーションに対応できなかった。例えば単純な文の例だが、「肝臓癌に関して、治療成績が向上した年は。」という自然文クエリーを解析したときの意味構造が図 2 のように計算されたとする。

しかし、正解文書中の記述は「…癌の治療成績について…」とあり、その意味構造は図 3 のようになっている。先行研究においては、直接つながる 2 つのノードとその関係（アーク）を検索キーとしていたため、図 2 の意味構造からは、図 3 のような、CANCER と ABCXYZ が直接つながる部分グラフは検索キーとして生成されない。このため、正解文書は検索されない。

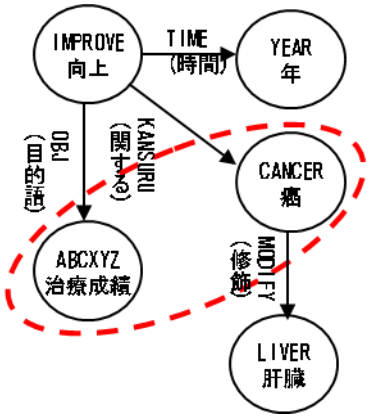

図 2. クエリーの意味構造

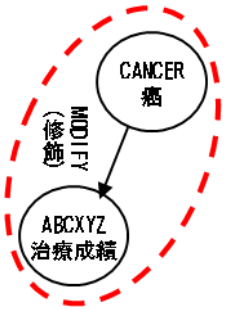

図 3. 正解文書中の意味構造（一部）

また、1 文で 1 つの意味構造を持つため、複数文からなるクエリー（例：「肝臓癌に関する文献を検索したい。その治療成績が向上した年は？」）だと、例えば 1 文目の意味構造中の意味記号と 2 文目の意味構造中の意味記号からなる検索キーは生成されないことになる（例：「CANCER(癌)、ABCXYZ(治療成績)」）。これが正解文書に含まれる可能性はあるだろう。

この例と逆に、クエリーの意味構造中では直接つながっているノードが、正解文書の意味構造中では直接はつながっていない、ということもある。

上に挙げたのは単純な例だが、文が複雑で長文になればなるほど、1 文中でも多くの語が複雑な関係をもつため、同じ意味であってもその意味構造は多様であり、部分グラフが一致しない、すなわち検索結果として検出されないという検索もれの問題が起きてしまう可能性が高くなる。複雑な文においても検索もれがなく意味構造を検索するには、同じ意味を表す意味構造のバリエーションに対して頑強に検索できなければならない。今回、再現率を向上させるため、以下の 2 つの技術を開発した。

1. 検索キー生成：文章中で離れた意味記号間の関係も加味することで複雑な文章の検索に対処する。この際に意味記号の組み合わせのうち不必要なものを自動的に除去し、ノイズを最大限に抑える。
2. マッチング：ノードが直接つながる構造だけでなく、間接的につながる構造も検索対象とする。

以下、これらの技術について説明する。

4.2. 検索キー生成

検索クエリーでヒットしなかった正解文書について調査したところ、検索クエリーが複数文からなり、別々の文中の意味記号が、正解文書の（1 文の）意味構造においてつながっていることがしばしばあることがわかった。そこで、検索キーを、クエリーの意味記号の部分構造とするのではなく、クエリー中に出現する意味記号を組み合わせたものとした。検索キーの重みは、意味記号の IDF×意味記号のクエリー中の出現頻度、とした。アークについては、別々の文中の意味記号を組み合わせるため、アークはマッチさせない方法や、複数のアークにマッチさせる方法などが考えられるが、今回は再現率向上を第一の目的として、アークはマッチさせず、意味記号が接続していればよいことにした。

しかし、任意の意味記号の組み合わせをつくると、検索キーには大量のノイズ（＝多くの不正解文書にマッチする検索キー）も含まれる。この問題に対処するため、ノイズとなる検索キーの自動判定を 2 段階で行うようにした。まず、明らかにノイズと考えられる検索キーを除去する。これは、クエリー中の意味記号の頻度や IDF 値、意味記号の品詞や属性をもとに判定する。例えば、副詞と名詞の組み合わせは検索キーとしては適さない。次に、検索後の判定も行う。検索キー（組み合わせ）は多くの文書にマッチするほど、文書の分別能力がなく、ノイズになりやすい。そこで、検索キーごとにマッチ文書数を計算し、その降順にソートして上位 n%をノイズとなりやすい検索キーと自動判定する。それぞれの文書の評価値は、(マッチした検索キーの重み×そのマッチ数)の総和で表され、ランキングは文書の評価値をもとに行われる。ノイズと判定された検索キーについては、その重みを低くして文書の評価値を再計算する。

4.3. マッチング

先行研究においては、クエリーの意味構造の部分構造（2 ノードとそれらを結ぶアーク）と、データベース中の意味構造の部分構造を完全一致させる検索手法を採用していた。しかし上述のように、クエリー中の意味構造で直接つながるノードが、正解文書中の意味構造において直接つながっていない場合もある。1 つの他のノードを介して 2 つのノードがつながることを「間接接続」と定義し、間接接続するノードも検索するようにした。このとき、アークは何でもよいということにした。例えば、図 4 の検索

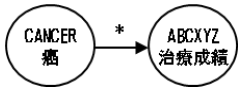

図 4. 検索キー

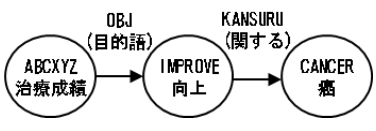

図 5. 検索される構造

表 1. 実験結果

86 クエリー	200 位以内に正解が入ったクエリー数	D)意味構造検索の勝敗（200 位以内）
A)従来の意味構造検索	27	25 勝 13 敗 2 分
B)キーワードベース検索（独自）	21	22 勝 13 敗 4 分
C)キーワードベース検索（製品）	25	25 勝 13 敗 5 分
D)本稿の意味構造検索	34	---

キーに対しては、”CANCER”と”ABCXYZ”が直接つながる構造以外に、図 5 のような間接接続の構造も検索することにする。

5. 評価実験

複雑な文を多く含む特許明細書を検索対象として、以下の 4 つのシステムを比較した。

A) 従来研究の意味構造検索システム
B) キーワードベースの検索システム（独自に開発）
C) キーワードベースの検索システム（製品）
D) 今回の手法を用いた意味構造検索システム

システム B)は、キーワードベース検索で一般的に使われているベクタースペースモデル[5]に基づく検索方式を独自に実装した。システム C)は市販されている製品[6]の 1 機能でキーワードベースのシステムであるが、検索対象が名称、要約、請求項、図と符号の説明のみである点で、明細書の本文全文を対象に検索する A), B), D)のシステムと異なる。

実験のための検索対象データとしては、公開されている特許明細書（限定された分野のもの、約 30 万件）を使用した。課題として与えられるのは、社内の公知例調査にかける前の特許アイデアを説明した書類（特許アイデア書類と呼ぶ）である。特許アイデア書類をもとに、社内で公知例調査を行った結果提示された案件（複数あることもある）を正解とする。被験者は、特許アイデア書類を読み自然文のクエリーを作成する。その後、4 つのシステムでそれぞれ、ランキング検索したときの正解文書順位を出す。正解文書の順位が 200 位以下の場合は検索できなかったものとみなす。クエリーは 1 つとは限らず、複数つくってもよいものとする。86 のクエリーで検索した結果を表 1 に示す。

200 位以内に正解がランクされたクエリー数については、従来の意味構造検索よりも約 26%向上した。また従来のキーワードベースの自然文検索と比較しても、従来の 21(B), 25(C)から 34(D)に、約 36%～62%向上した。200 位以内に正解文書がランクされた結果に対して正解文書順位が上か下かで勝敗をつけたところ、本稿の意味構造検索が優位であり、本手法による意味構造検索の有用性を確認した。

6. 今後の展望

この技術は特許検索だけでなく、論文や社内文書の検索など幅広い分野で使用でき、調査業務における効率化および調査の質の向上が期待できる。

参考文献

[1]意味構造を用いた情報検索システム「かもめ」AIST Today, Vol.3 (2003 年) 9 月号, p30.
[2]大倉清司,潮田明(2012) 意味検索のプロトタイプシステムの構築. 言語処理学会第 18 回年次大会予稿集. 2012.
[3] Fillmore, Charles J. (1968) The Case for Case In: E. Bach and R.T. Harms (eds) Universals in. Linguistic Theory. Holt, Rinehart and Winston, New York. pp. 1-88
[4] 富士通.英日・日英翻訳ソフト ATLAS. http://software.fujitsu.com/jp/atlas/
[5] Gerard Salton and M. J. McGill. Introduction to Modern Information Retrieval. McGraw Hill Book Co., New York, 1983.
[6] 富士通 知的財産ソリューション ATMS. http://jp.fujitsu.com/solutions/ip/