特許文における長い名詞句表現の自動解析について

株式会社富士通研究所　メディア処理システム研究所主管研究員　**潮田　明**

PROFILE　1983年東京大学理学部物理学科卒業。同年株式会社富士通研究所入社。表面磁気光学効果、統計自然言語処理、機械翻訳などの研究に従事。1995年ＡＴＲ音声翻訳通信研究所研究員、2003年富士通研究所知能システム研究部部長を経て現職。電子情報通信学会論文誌Ｄ編集幹事、高度言語情報融合フォーラム（ALAGIN）幹事。

ushioda@jp.fujitsu.com

1　はじめに

産業日本語の代表格として取り上げられている特許文書や契約書文書においては、一般に１文が長く、複雑な専門用語や複合名詞、長い名詞句表現などを多く含むことなどが特徴として挙げられる。特に特許の請求項に関しては古くから１文で記載する慣習があり、そのため日常遭遇することがまずない程の長い文が出現することがしばしばある。長い名詞句表現を含む文は人間が読んでも時に理解が難しいものであるが、連体修飾節や連体修飾句を含む長い名詞句表現の構造解析は並列句の解析などとともに、構文解析における難題の１つである。長い名詞句表現を解析する際の問題点は昨年度も本誌で取り上げたが［1］、本稿では長い名詞句表現における係り受け解析を自動で行う手法について考える。

構文的曖昧性は語句の意味を考慮に入れないと一般的には解消できないが、名詞句の構造解析においても意味的情報の役割が大きいことが知られている。［2］は「ＡのＢ」型名詞句に対する連体修飾節の係り先の決定の際に、名詞の意味情報の組合せに基づく係り先の解析方式を導入している。６つの独立した個別解析方式の中で、名詞の品詞の組合せ情報に次いで、名詞の意味情報の組合せに基づく解析が有効であると報告している。［3］は連体修飾句と名詞の係り受け問題の代表的な例である「ＡのＢのＣ」型名詞句を取り上げ、係り先決定に有効な機械学習用素性として「ＡのＢ」の意味的分類と「の型連体句」の係り先との関わりを定式化して用いている。［4］は「ＡのＢのＣ」型名詞句を対象に、意味分類辞書における名詞の意味属性を用いた係り受け規則の自動生成法を提案している。

意味分類辞書など人手で編纂された辞書は、コーパスから収集される共起情報などに比べて分野や表現の偏りが少ない反面、量的に限られている上に、長い複合名詞などのエントリーは一般的に乏しく、特定の分野や文種のテキストに対して網羅的に適用するのは難しい。一方、特定の分野や文種のテキストコーパスが大量に存在する場合、コーパスから単語の共起情報や表現サンプルを抽出して係り受け関係の解析に適用する手法が有効である。しかしながら単語や表現サンプルを汎化なしに文字列レベルで扱う場合、通常、データのスパース性により適用すべき共起関係や表現サンプルが見つからないと言った問題に直面する。

本稿では、大量のテキストが入手可能で、かつ長くて複雑な名詞句表現が頻出する特許文書を対象とし、長い名詞句表現の代表例として「～によるＡのＢ」、「～におけるＡのＢ」、「～の（～する）ためのＡのＢ」と言った、連体形複合辞に修飾された名詞句表現を取り上げ、長い名詞句表現内の係り受け解析を自動で行う手法について考察する。係り受けの自動解析手法には大きく分けてルールを用いる方法と、人手による解析結果（アノテーション付コーパス）をもとに機械学習する方法が挙げられるが、本稿ではルールや人手による学習コーパスは用いず、その代わりベースになる既存の自動係り受け解析器（パーサ）を利用する手法を探る。ベースの自動係り受け解析器の出力から単純なケースにおける最も信頼できる部分解析結果を抽出し、抽出結果を汎化した上で統計的に組み合わせることにより、複雑な入力に対する解

析結果を、ベースの自動係り受け解析器よりも遥かに高い精度で得ることのできる手法を考察する［5］。汎化の手段としては大量のテキストから自動学習した表現のクラスを用いる。

2 「F ＋ A の B」型名詞句表現

本稿において曖昧性解消の対象とする「［連体形複合辞で終わる連体修飾句］＋［A］＋［の］＋［B］」（A, B はそれぞれ名詞）という型の名詞句を以下「F ＋ A の B」型名詞句と呼ぶことにする。まず「F ＋ A の B」型名詞句の例として、

（1）カラーフィルタ基板の各被膜と基板との密着性を定量的に評価するための被膜の密着性評価方法
（2）ピッキングロボットによる物品の自動移載

という表現について係り受け関係を考えてみる。構文的にはそれぞれ２つの解釈が可能である。

（1A）［カラーフィルタ基板の各被膜と基板との密着性を定量的に評価するための被膜］の密着性評価方法
（1B）カラーフィルタ基板の各被膜と基板との密着性を定量的に評価するための［被膜の密着性評価方法］

（2A）［ピッキングロボットによる物品］の自動移載
（2B）ピッキングロボットによる［物品の自動移載］

（1）の場合、構文上は（1A）、（1B）２通りの解釈が可能だが、密着性を定量的に評価するために用いるのは被膜そのものよりは被膜の密着性評価方法の方が自然であるという意味的な制約から、人間は（1B）の解釈を選択できる。（2）においても同様に、（2A）より（2B）の方が自然な解釈であると判断できる。しかしながら語句同士の意味的な関係を用いず品詞などの統語論的な素性のみに基づいて処理を行う多くの構文解析器にとっては、これらの構文上の曖昧性を解消することは難しい。

また［1］においても論じたように、機械だけではなく人間にとっても係り受けの判断が難しいケースも存在する。一般に、「F ＋ A の B」型名詞句における係り受け解析の曖昧性は以下のグループに分けて考えることができる。

（i）係り受け関係の解釈によって文の意味が変らないケース
（ii）係り受け関係の解釈によって文の意味が変わるケース（常識にそぐわない解釈も含む）
　（ii-1）文脈、分野知識に依存せずに係り受け関係が一意に定まるケース
　（ii-2）係り受け関係を一意に定めるために文脈、分野知識等を必要とするケース

上の例文（1）、（2）は（ii-1）のケースに相当する。一方、「受光部周辺の光透過層からの光入射によるスミアの防止を図る」といった文では、「光入射による」が「スミア」に係るのか「防止」に係るのかで文の意味が変わるが、どちらの解釈が正しいのかを判断するには分野特有の知識が必要であり、上のタイプ分けでは（ii-2）のケースに相当する。また、「熱疲労による障害の発生を検知する」という文では「熱疲労による」が「障害」に係っても「発生」に係っても文の実質的な意味には変化はないため、（i）に分類される。本稿では、「F ＋ A の B」型名詞句の係り受け解析のうち、文脈や分野知識に依存せずに係り受け関係が一意に定まるケース（ii-1）を取り上げる。理由は、現状の日本語解析器は未だに（ii-1）のケースにおいても精度良く曖昧性解消が出来るレベルにはないこと、そして現実的な問題として（i）や（ii-2）のケースを含んだ評価用データを作成することが容易ではないためである。

3 曖昧性解消の手法

日本語の連体修飾節、連体修飾句の係り先決定問題は、長い名詞句の構造解析の中心的課題である点や、意味的情報の役割が大きいという点で、英語における前置詞句の修飾先の決定問題（PP-attachment problem）と共通している。[6] は、動詞 v、名詞句 np_1、前置詞 p、名詞句 np_2 が順に並んでいる文において、前置詞句［p + np_2］が動詞 v と名詞句 np_1 のどちらを修飾するかという係り受け問題を解く際に、p が v に結びつく尤度と p が np_1 の主辞（head）n_1 に結びつく尤度を比較するために次式で表される LA（Lexical Association）score という尺度を導入している。

$$LA(v, n_1, p)$$
$$=\log_2[P(\text{v-attach } p \mid v,n_1)/P(\text{n-attach } p \mid v,n_1)]$$
$$=\log_2[P(p \mid v) * P(NULL \mid n_1)/P(p \mid n_1)] \quad ①$$

$\log_2$ の引数の分子は、v が出現する時に p を head とする前置詞句を伴って出現する確率と、名詞 n_1 を head とする名詞句が出現する時に前置詞句によって修飾されていない確率とを掛け合わせたものである。分母は、名詞 n_1 を head とする名詞句が出現する時に p を head とする前置詞句が名詞句に続いて出現する確率を表す。LA score が正のときに動詞修飾、負のときに名詞句修飾と判定される。

「F + A の B」型名詞句の係り受け解析においても PP-attachment 問題における LA score と同様の尺度を導入することが考えられる。ここであらためて F は複合辞を含む文節、A、B はそれぞれ名詞を表すことにする。

$$LA(F, A, B) =$$
$$\log_2[P(\text{A 係りの F} \mid A,B)/P(\text{B 係りの F} \mid A,B)] \quad ②$$

$$P(\text{A 係りの F} \mid A,B) = P(F \mid A) \quad ③$$
$$P(\text{B 係りの F} \mid A,B) = P(F \mid B) \quad ④$$

それぞれの確率は、コーパス中の共起頻度から推定することができる。

$$P(F \mid A) = freq(F, A)/freq(A) \quad ⑤$$
$$P(F \mid B) = freq(F, B)/freq(B) \quad ⑥$$

①〜⑥より

$$LA(F, A, B) =$$
$$\log_2[(freq(F,A)/freq(A))/(freq\ (F,B)/freq\ (B))] \quad ⑦$$

式①では、動詞修飾の確率を計算する際に、前置詞句が動詞にかかっていることだけではなく、前置詞句が名詞句にはかかっていないことをも条件に入れるために P（NULL | n_1）を掛け合わせている。式③，④においてはそのような項を導入してはいないが、F と名詞の共起頻度の定義において、実際に F が名詞にかかっている場合のみを「共起」と定義する必要がある。すなわち、F と名詞 X（A あるいは B）との共起頻度 freq（F, X）は、コーパス中で F が X を修飾している頻度を指す。しかしながら、実際に F が X にかかっているかどうかは不明なため、freq（F, X）を厳密に求めることはできない。そこでここでは、F が X にかかっていると解釈するのが最も確からしいケースに絞って式⑦の値を推定することにする。たとえば、freq（品質管理における，エンドユーザ）を計算する際に、「品質管理におけるエンドユーザからみた改善点」もカウントに入れてしまうと余分なカウントを招くことになる。同様に「品質管理におけるエンドユーザの役割」、「品質管理におけるエンドユーザに対する配慮」なども F と X が係り受け関係にないケースである。一般的に言って、文中で F に続いて X が現れたとき、X に助詞「を」が続いた場合、他の「の」、「に」、「と」、「から」などの助詞が続いた場合に比べて、F を含む名詞句が X で終わっている可能性が高いと考えられる。X で終わっていれば F が X にかかっているのは確実である。そこで今回は freq（F, X）の代わりに freq（F, "X を "）を用いることとし、同時に freq（X）

の代わりに freq（“X を”）を用いることとする。なお、コーパスからの最尤推定の際にはいつも問題になることであるが、頻度が低い場合の推定値の信頼度は低い。特に式⑦中に現れる頻度が 0 の時は式自体がなりたたない。そこで、最も簡便な処理ではあるがここでは頻度 freq（.）にすべて 0.5 を加えることとする。以下に具体的な手順を示す。

1）コーパスを既存の形態素解析器 juman［7］および構文解析器 knp［8］を用いて解析する。但し解析は、文を文節単位に区切った上で名詞句の境界を判定する目的のみに用いる。名詞句内の構造は考慮せずフラットな構造（文節の列）として名詞句を抽出する。但し、名詞句が入れ子構造になっている場合、すべての名詞句を抽出する。
2）抽出された名詞句の集合の中から、「X を」で終わる名詞句の出現頻度をカウントし、freq（X）の値とする。
3）「X を」で終わる名詞句の内、F で始まる名詞句の出現頻度をカウントし、freq（F, X）の値とする。
4）式⑦より LA（F, A, B）の値を求め、値が正ならば A 係り、負あるいは 0 ならば B 係りと判定する。

4 文節クラスの共起情報を用いた汎化手法

前節に記した Lexical Association score を用いる手法は、文節同士の共起頻度が高い場合は有効であると期待されるが、実際にオープンテストで評価した場合は頻度 0 の出現を避けることはできない。そこで、クラスタリングにより文節を意味的クラスに分類し、文節の代わりに文節の属するクラスの頻度を用いることでデータスパースネス問題を回避する手法も同時に検討する。具体的には以下の手順で行なう。

1）Juman および knp を用いてコーパスを文節単位に区切る。
2）文節を単位として、相互情報量に基づく凝集クラスタリングアルゴリズム［9,10］により文節のクラス分けを行なう。具体的には［9］に記載の MI-Clustering により初期クラスタを生成した後、Reshuffling を適用してより精緻化されたクラスタを生成する。
3）式⑦において、文節の出現頻度に替わり文節クラスの出現頻度を用いて LA score を求めて係り先の判定を行なう。

5 評価実験

本稿では小規模なテストセットを用いた予備評価の結果を報告する。国内出願特許における抄録の課題文を実験対象とした。1993 年から 2007 年までの国内出願特許から抽出した課題文 4,801,350 文を、2006 年までの出願分 4,481,178 文と 2007 年の出願分 320,172 文に分け、前者を data-1、後者を data-2 とした。data-1 を用いて文節のクラスタリングおよび頻度情報の抽出を行ない、data-2 を用いて提案手法の評価を行なった。Data-2 は文単位でランダムにシャフルしたものを用いた。

クラスタリング

Data-1 より出現頻度 2 以上の文節 2,334,045 個を抽出し 1000 個のクラスに分類した。

複合辞の選択と評価用データ

特許文に頻出する 3 つの複合辞「による」、「における」、「での」を対象に評価を行なった。それぞれの複合辞について、data-2 より作業者 1 名による人手で、「F ＋ A の B」型名詞句を含む文を抽出し、A 係り／B 係りの判定結果合計 50 件ずつを作成した。

連体形複合辞に修飾された名詞句の係り受け解析に関して比較可能な過去の報告がないため、今回は参考値として、日本語の係り受け解析において現在標準的に使われている 2 つの係り受け解析器 knp および cabocha［11］による「F ＋ A の B」型名詞句の係り受け解析の精度を記すことにした。Cabocha は形態素解析エンジン mecab［12］と組み合わせて用いた。

評価結果

表 1 に評価結果を示す。Knp および cabocha においては、「F ＋ A の B」型名詞句の係り受け解析では隣接文節間の係り受けを優先して過剰に A 係りが選択される傾向があり、精度は低い。少数サンプルによる予備評価結果ではあるが、これまで国内で標準的に用いられてきている構文解析器の精度であり、本課題の難しさを表していると言える。文節ベースの提案手法では、knp および cabocha の精度は上回っているものの、すべての係り受け関係を対象とした場合の標準的係り受け解析精度よりは遥かに低い。一方文節のクラスにより汎化を行った文節クラスベースの提案手法においては 90%以上の精度が得られており、汎化の効果が十分高いことが確認できた。

表 1 「F ＋ A の B」型名詞句の係り受け解析精度（%）

複合辞	knp	cabocha	文節ベースの提案手法	文節クラスベースの提案手法
による	14.0	10.0	30.0	92.0
における	4.0	4.0	32.0	90.0
での	16.0	16.0	46.0	92.0

6 まとめ

長い名詞句表現の代表例の 1 つである「〜による A の B」、「〜における A の B」と言った、連体形複合辞に修飾された名詞句の係り受け解析問題を取り上げ、大量のテキストコーパスから自動抽出した共起情報をもとに係り受けの曖昧性を解消する一手法について考察し、実験によりその有効性を検証した。更に共起情報のスパース性問題に対処するために、大量のテキストコーパスから自動構築した文節クラスを用い、名詞句表現の汎化を行った。少数サンプルを用いた予備評価実験において、汎化の効果が十分高いという結果が得られ、また人手によるトレーニングデータを一切使わずに解析精度の向上が図れることが分った。今後はより多くの複合辞に関して本手法の有効性を検証するとともに、他の種類の係り受け解析に対する適用を検討していく。

参考資料

[1] 潮田明. Japio 2011 Year Book (2011) pp.284-285.

[2] 安井敏，徳久雅人，村上仁一，池原悟.「A の B」型名詞句に対する連体修飾節の係り先の決定. 言語処理学会第 12 回年次大会発表論文集，pp.125-128，2006.

[3] 美野秀弥，橋本泰一，徳永健伸，田中穂積. 日本語の連体修飾関係に関する研究. 言語処理学会第 10 回年次大会発表論文集，pp.600-603，2004.

[4] 中井慎司，池原悟. 白井諭：「の」型名詞句における名詞間の係り受け規則の自動生成法，信学技報，Vol.98, No.53, pp.15-22 (1998).

[5] 潮田明. 連体形複合辞に修飾された名詞句の係り受け解析. 言語処理学会第 18 回年次大会発表論文集，pp.967-970，2012.

[6] Hindle, D., Rooth, M. (1993) “Structural ambiguity and lexical relations”. Computational Linguistics, v.19 n.1.

[7] http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman/juman-5.1.tar.gz

[8] http://nlp.kuee.kyoto-u.ac.jp/nl-resource/knp/knp-2.0.tar.gz

[9] Ushioda, A. (1996) "Hierarchical Clustering of Words". Proceedings of the 16th International Conference on Computational Linguistics, August 05-09, 1996, Copenhagen, Denmark.

[10] Ushioda, A. (1996) “Hierarchical clustering of words and application to NLP tasks”. Proceedings of the 4th Workshop on Very Large Corpora, pp. 28-41.

[11] http://code.google.com/p/cabocha/

[12] http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html