パーサ適応と格フレームによる特許文構文解析精度の向上

吉田 節行† 横山 晶一‡
山形大学大学院理工学研究科

1．研究概要

特許文は 200 文字を超える長大な文からなることがあり、係り受け構造が複雑なことで知られている。文章の構造の複雑さから係り受け解析に誤りが生じ、その結果が機械翻訳の精度などに影響を及ぼしている。近年、特許などの知的財産が社会における貴重な存在として認められてきている。これに伴い、特許申請数が増加している。また、国際化により国際特許も増加傾向である。特許文を機械処理することは、翻訳コストの削減、翻訳時間短縮による知的生産時間の生成、即時性のある対応が可能になるなどの観点から重要な研究要素である。

本研究では、特許領域へのパーサ適応による特許文構文解析精度の向上を目的として研究を行った。パーサ適応の手法の 1 つとして、2 つの異なるパーサによる Ensemble という手法を導入した。その結果、MST+KNP の組み合わせにおいて、精度の向上が見られた。

2．背景

2.1 研究背景と資料

昔から日本語の構文解析に関しては、詳しい研究がなされているが、特許文に限定した研究や文献は言語学的にも機械的にも少なく、長文に対する安定した構文解析もいまだに困難な状況下にある。

また、本研究では資料として AAMT/Japio 研究会特許情報データベース DVD[1]を用いる。これは特許広報をコンピュータデータベース化したものである。特許文は、特許庁 HP 中の特許電子図書館として、WEB 上でも無料で一般公開されている。

2.2 過去の研究

(1) 特許文係り受け解析誤りの分類・修正

特許文の係り受けを解析し、係り受けが誤る場合、分類可能なものを使用例ごとに分類することを目指した研究[2]がある。さらに、この研究をもとにそれぞれの誤りの分類に対する修正についての研究[3]が行われた。その結果、「並列構造」、「名詞修飾」、「接続詞」、「特許特有表現」の４分類に関して、効果のある修正案が提示できた。この研究をもとに自動的に係り受け解析誤りを修正するシステムの構築を行った研究[4]がある。この研究では、係り受け解析誤り分類のうち、「並列構造」の誤り、「特許特有表現」の誤りについて、自動的に修正を行うシステムの作成を行っている。特許特有表現については修正に成功しているが、並列構造については一部問題が残る結果となった。その原因として、修正手段の一つとして用いた日本語語彙大系[5]の構造などが挙げられている。今後の展望として辞書に類語大辞典[6]を用いるなどの方法が提案されている。

(2) Ensemble によるパーサ適応

特許文の構文解析に関する過去の研究では、パーサによる解析結果を外部で修正する形で行われてきた。それに対し、本研究では、パーサ自体を特許文に特化したものチューニングすることを考える。その為に本研究では Ensemble という手法を用いてパーサの適応を行う。この手法は Tsujii らによって提案されたもので[7]、2 つの異なるパーサの解析結果を用いて、あるドメインに対してパーサを適応させる。本研究では、KNP[8]、MST Parser[9]、malt Parser[10]を用いて Ensemble を行う。KNP は日本語の構文解析器で、並列構造の解析に強いとされる。また、MST Parser、malt Parser は多言語パーサであり、対応するフォーマットのデータを与えることで日本語にも対応できる。

3．研究内容

3.1 係り受け解析正解データ

評価用のデータとして人手により修正を行った係り受け解析データを用いる。この係り受けの正解データの作成は京都大学テキストコーパス[11]のアノテーションツールを用いて行う。2004 年公開分の特許より G06F（電気的デジタルデータ処理）、C12N(生物)、F01(機械)の 3 分野について抽出したものからランダムに 1000 文抜き出し、これらににについてアノテーションを行った。

3.2 実験

KNP, MST Parser, malt Parser を用いて Ensemble を行う。使用データは AAMT/Japio 研究会特許情報データベースのうち 2004 年度追加分の F01(機械)より、適切でないものを削除した約 25 万文である。以下に手順を述べる。

(1) 京大コーパス約 4 万文を用いて MST Parser、malt Parser を日本語に適応する。これをデフォルトのパーサとする。

(2) F01 分野約 25 万文を(1)の 2 つのデフォルト

Improvement of Parsing Accuracy of Patent Sentences using Parser Adjustment and Case Frames

† YOSHIDA Takayuki Yamagata University
‡ YOKOYAMA Shoichi Yamagata University

パーサと KNP で解析。
(3) MSTParser+KNP、MSTParser+maltParser の 2 通りの組み合わせで一致する結果を蓄積。
(4) (3)の結果を用いて MSTParser を学習。
(5) (4)の結果、作成された解析モデルを用いて、G06F (電気的デジタルデータ処理) 、C12N(生物)、F01(機械)の 3 分野 3, 000 文について評価する。

3.4 実験結果

Accuracy :

$$\frac{\text{正解データと一致した形態素数}}{\text{全形態素数}}$$

Complete Correctness(C.C) :

$$\frac{\text{正解データと文全体において一致した形態素数}}{\text{全形態素数}}$$

表 1. 評価結果

		F01 (機械)	C12N (生物)	G06F (情報)
デフォルト	Accuracy	0.9039	0.9176	0.9084
	C.C	0.1131	0.1781	0.1381
MST+KNP	Accuracy	**0.9236**	**0.9273**	**0.9219**
	C.C	0.1931	0.2342	0.2162
MST+malt	Accuracy	0.9062	0.9142	0.9086
	C.C	0.1361	0.1891	0.1431

今回の実験では F01(機械)分野の特許文を用い、KNP と MST Parser、MST Parser と malt Parser の 2 つの組み合わせについて Ensemble を行った。その結果、MST+KNP において、精度の向上が見られた。MST+malt においては、デフォルトに対し精度の向上があまり見られなかったが完全一致に関しては若干の向上が見られ、Ensemble という手法が、特許文領域へのパーサ適応の手段として有効であるということが言える。

4. まとめ

実験により、Ensemble による特許領域へのパーサ適応が有効であることがわかった。しかし、今回評価に用いた正解データは、KNP の解析結果をもとに人手で修正を行ったものであったため、KNP の解析結果が反映されている MST+KNP の組み合わせで最も良い結果が得られた可能性がある。さらに正確なデータを得るためには正解データの見直しが必要である。

また、さらに精度を向上させるため、パーサ適応を行ったパーサによる解析結果を修正について考える。次の方法で簡単な実験を行ったところ、いくつかの例について修正が可能であることがわかった。

複数の動詞をまたいで係っているものを探し、その間に含まれる動詞について格フレームを検索する。図 1 の例では、「完了」「通知」について調べる。「通知」についてヲ格として「番号」という体言をとることがわかる。この例では、「通知される」と受身なので、目的格ヲは主格になり得る。このことから、係り受け修正を行う。

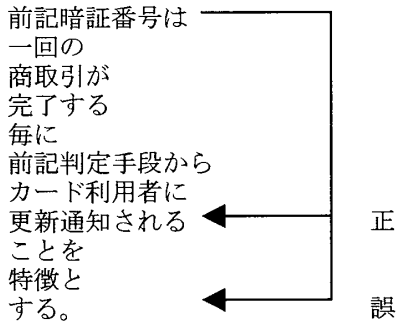

図 1. 例文の一部

このような修正案に基づき自動的に修正を行い、さらに修正後のデータを用いて学習することにより、結果をパーサに反映させることで精度の向上が望める。

謝辞

データを提供して頂いた AAMT/Japio 研究会に感謝致します。

本研究の一部は科学技術研究費 (基盤研究(C)) 「動的シソーラスを用いた特許文の解析システム」 (課題番号 18500102) に基づき行われた。

参考文献

[1] AAMT/Japio 研究会特許情報データベース

[2] YOKOYAMA Shoichi and KANEDA Yuya: Classification of Modified Relationships in Japanese Patent Sentences, Proceedings of Workshop on Patent Translation in the 10th Machine Translation Summit (2005) pp.16-20

[3] 佐原 洋輔：特許文の係り受け解析と修正、平成１７年度山形大学工学部卒業論文(2006)

[4] KENNEBDAI Shigehiro and YOKOYAMA Shoichi : A System Correcting Modification Error in Patent Sentences (in Japanese), Proceeding of the 69th Meeting of the Information Processing Society Japan (IPSJ)(2007)6Q-3, pp.2-427-8

[5] 池原悟、宮崎正弘、白井諭、横尾昭男、中岩浩巳、小倉健太郎、大山芳史、林良彦：日本語語彙大系、岩波書店(1997)

[6] 柴田武 山田進：類語大辞典、講談社(2002)

[7] Tsujii, Sagae : Dependency Parsing and Domain Adaptation with LR Models and Parser Ensembles, EMNLP-CoNLL 2007, pp.1044-1050

[8] KNP : 京都大学黒橋研究室(http://www-lab25.kuee.kyoto-u.ac.jp/nl-resource/knp.html)

[9] R. McDonald, F. Pereira, K. Ribarov, and J.Hajic : Non-projective Dependency Parsing using Spanning Tree Algorithms, Proceedings of HLT/EMNLP (2005)

[10] maltParser(http://w3.msi.vxu.se/~jha/maltparser/)

[11]京都大学テキストコーパス(http://www-lab25.kuee.kyoto-u.ac.jp/nl-resource/corpus.html)