大規模情報抽出のための頑健な述語項構造解析の研究

Robust Predicate Argument Structure Analysis for Large Scale Information Extraction

小町　守（Mamoru KOMACHI, Ph. D.）
首都大学東京 システムデザイン学部 准教授
(Associate Professor, Tokyo Metropolitan University, Faculty of System Design)
言語処理学会　情報処理学会　人工知能学会　電子情報通信学会
Association for Computational Linguistics
受賞：言語処理学会 20 周年記念論文賞（2014 年度）　人工知能学会論文賞（2010 年度）　情報処理学会山下記念研究賞（2010 年度）　言語処理学会年次大会最優秀発表賞（2009 年度）　言語処理学会年次大会優秀発表賞（2008 年度）
研究専門分野：自然言語処理

１．概要

ウェブの発展に伴い、Twitter や Facebook のようなユーザが内容を書き込むサイトや、楽天や Amazon のような巨大なオンラインショッピングサイトが登場した。これらのサイトは規模が大きく、かつ更新が速いため、有用な知識が多く含まれている。このようなウェブからの知識獲得には、誰が何をどうした（娘がインフルエンザにかかった）、あるいは何がどうである（iPhone の画面がきれい）、といった情報（述語項構造）を解析することが必要であるが、ウェブテキストは砕けた表現や新しい用語が多く含まれるため、既存の手法で解析することは困難である。

そこで、本研究では、日本語の述語項構造解析の高精度化と頑健化に関する研究を行い、前者では先行する手法を上回る精度を達成し、後者では Twitter テキストと楽天市場のリビューテキストを含むウェブテキストを対象にした述語項構造解析器を作成した。

２．研究目的

文書から情報抽出を行うに当たって、誰が何をどうした、あるいは何がどうである、といった情報（述語項構造）を解析することは、意味を理解した知識獲得のための基礎的な技術である。しかしながら、現在の日本語述語項構造解析の精度は新聞記事を対象として 7 割程度と、いまだ実用的な精度には達していない。また、近年大規模なテキストデータとしてウェブから得られるテキストがあり、有益な情報が含まれるが、既存の新聞記事を対象とした述語項構造解析器では頑健に解析を行うことができない。そこで本研究は、①日本語の述語項構造解析器の高精度化、そして②ウェブテキストを対象とした日本語述語項構造器の作成、という 2 つを目的とする。

３．研究背景

日本語の述語項構造解析は、述語と項の相対的な関係により、項の自動推定の難しさが異なるが、それぞれの述語に対して項をどのように決めるかによって様々な手法が存在する。たとえば、Iida et al. (2007) [1] や Taira et al. (2008) [2] は、述語と項の関係ごとに解析システムを分け、経験的なルールを用いて簡単なものから項を同定する手法を提案した。一方、笹野ら (2011)[3] や吉川ら (2013)[4] は、明示的にシステムを分けず、機械学習を用いて同時に項を同定する手法を提案した。しかしながら、順番に項候補を見る手法では全ての候補を比較できないという問題がある一方、全ての項候補を比較検討する手法は計算量が多いという問題があった。

また、ほとんどの研究は、京都大学テキストコーパス[*1] あるいは NAIST テキストコーパス[*2] という新聞記事を対象としたデータを用いており、数少ない例外は京大・NTT 解析済みブログコーパス[*3] であったが、収録されているデータが少なく、機械学習を行うには不十分である、という問題点があった。

日本語以外では、PropBank[*4] や OntoNotes[*5] といった英語・中国語・アラビア語を対象とした述語項構造解析（意味役割付与）のデータを用いた研究が盛んであるが、英語や中国語、アラビア語は日本語のよう

に省略が多くないため、これらの言語における解析のアルゴリズムをそのまま日本語に適用することはできない。

*1 http://nlp.ist.i.kyoto-u.ac.jp/index.php?
京都大学テキストコーパス

*2 https://sites.google.com/site/naisttextcorpus/

*3 http://nlp.ist.i.kyoto-u.ac.jp/kuntt/

*4 http://verbs.colorado.edu/~mpalmer/projects/ace.html

*5 https://catalog.ldc.upenn.edu/LDC2013T19

４．研究方法

従来の日本語述語項構造解析手法で問題となっていた計算量の問題を抑えつつ、全ての項候補を対象とした解析を行うために、述語と項の位置関係ごとに解析システムを分け、それぞれのシステムの中で最ももっともらしい候補を選出し、最後に選抜された候補のみでの比較を行うことで、効率のよい解析を行う手法を提案した[5]。

表１にそれぞれの手法の実験結果を掲げる。実験はNAISTテキストコーパスを用いたもので、「誰が何を何にどうした」に相当する、ガ格・ヲ格・ニ格をそれぞれ文章中から推定するというタスクである。表から分かるように、日本語でもっとも数の多いガ格の推定精度は、既存の手法と比べて最も高くなった。また、ヲ格に関しても Iida et al. (2007) と比較して統計的に有意（$p < 0.05$）に精度が向上しており、提案手法の有効性が示された。

表１　日本語述語項構造解析の比較
精度は適合率と再現率の調和平均（F 値）

	ガ格	ヲ格	ニ格
Iida et al. (2007)	76.85	88.65	74.86
Taira et al. (2008)	57.4	79.5	83.15
Imamura et al. (2009)	73.2	89.2	76.6
提案手法	77.59	88.96	75.01

高い精度で解析が行えるヲ格に比べ、ガ格とニ格は難しい事例が多く、複数の項を同時に推定する、あるいは簡単な事例から順番に推定していくことで、同じ名詞が同時に１つの述語の複数の項にならない、といったような制約を考慮する、といった方向性が考えられる。また、人間においても項の同定は複雑な制約や選好によって実現されているため、単純な機械学習手法では適切に解析できない。そこで、近年脚光を浴びている深層学習を適応することにより、複雑なシステムを作ることなく高精度・高速な解析を達成する、といった可能性も有望であると考える。

一方、新聞記事以外のデータを対象とした日本語述語項構造解析の性能調査として、ウェブテキスト（Yahoo! 知恵袋）に対する述語項構造解析の比較[7]と、新たに本研究で楽天データセットのレビューに対して属性が省略されている場合に属性情報を付与し、属性の自動推定を行う研究[9] に取り組んだ。

表２は、日本語書き言葉均衡コーパスの一部として作成された Yahoo! 知恵袋コーパスと新聞記事コーパス[6] を対象として、それぞれ述語項構造解析器を機械学習によって訓練し、それぞれのデータを対象として比較した結果である。表から分かるように、基本的には訓練データと評価データを同じジャンルにした方が精度は高いが、評価データが知恵袋の場合、新聞記事と合わせて訓練したものが最も精度が高かった。これは、知恵袋データに短い文が多く、データが少ないためであると考えられる。

表２　Yahoo! 知恵袋と新聞記事を用いた日本語述語項構造解析の比較（文内ガ格）
精度はＦ値

		評価データ	
		知恵袋	新聞
訓練データ	知恵袋	78.1 / 49.5	79.2 / 47.0
	新聞	76.7 / 43.7	84.1 / 59.8
	知恵袋＋新聞	78.1 / 51.8	77.4 / 48.0

左側の数字は述語と項が係り受け関係にある場合で、
右側は述語と項が係り受け関係にない場合の精度

述語と項の間に係り受けのある場合と比較して、係り受けのない場合（文内における主語の省略）の精度は低く、5 割程度であるが、これはデータの不足に起

因するものと、項同定に必要な情報をシステムがうまく利用できていない、という 2 つの原因が考えられる。前者に関しては、近年 word2vec [8] のような、分散表現と呼ばれる大規模テキストから単語の意味表現を学習する手法を適用する方向で解決することができると考えられるが、後者に関しては、文内において複数の述語が存在するときどのような項の共有関係があるか、といったような知識や制約をシステムで活用していくことが必要であろう。

一方、表 3 は、楽天データセットのノート PC に関するリビュー文書に対して評価値（主語）とその対象となる属性値情報を付与し、機械学習によって評価値・属性値を推定するシステムを訓練して推定する手法と、最頻出のラベルを常に出力するベースラインとを比較した実験結果を示している。表から分かるように、分野を絞って情報を付与することで、評価値の推定は高い精度で行うことが可能である。しかしながら、属性値の推定は難しく、どのような属性が項になりやすいか、といったような知識や、筆者らが参考文献[5]で作成したような、より高度な機械学習手法を適用する必要があることが分かった。

表３　楽天データセットを用いた省略のある評価値・属性値推定の比較

精度はＦ値

	評価値	属性値
最頻出ラベル	0.69	0.18 / 0.42
提案手法	0.89	0.24 / 0.55

属性値の左側の数値は文書内に属性値が存在する場合で、右側は文書内に属性値が存在しない場合の精度

また、今回作成したリビューのデータは、500 件と規模が小さく、より規模の大きいデータを活用する分野適応の手法や、人手を極力かけることなく機械学習を行う半教師あり手法の適応が課題である。後者に関しては、間接教師あり学習（distant supervision）[10]と呼ばれる手法が近年注目されており、人手で作成した辞書的な知識を用いて、大規模なレビュー文に対して自動で情報を付与し、訓練データとして活用する手法が効果を見込める。

５．将来展望

本研究で開発した日本語述語項構造解析手法は、オープンソースソフトウェアとして一般公開予定である。述語項構造解析技術は精度の高精度化もさることながら、ソフトウェアのインストールの容易さや速度、メモリの使用量といった使い勝手の面で、大いに改善の余地がある。これらのソフトウェア開発的な課題を一つ一つ解決していくことで、広く使われる技術になると考えている。

また、述語項構造解析の応用先の一つとして、疾病や疾患の流行検出がある[11,12]。これは、Twitter のようなリアルタイムで更新されるマイクロブログサービスからつぶやきを解析することで、インフルエンザの流行をいち早く検出する、というものであるが、このようなアプリケーションにおいては、最終的な目的（流行予測）が達成される限りにおいて、必ずしも用いている個々の技術の精度が 100%である必要はない。実際、筆者らは参考文献[11] においては、風邪の症状がある主体の推定精度が高くなくても、流行検出には効果があることを示すことができた。

要素技術としての蓄積が長いにも関わらず、必ずしも精度が高くない技術においては、現在の水準の技術でできることを積み重ね、応用タスクで有用性を広く示す、ということが、自然言語処理という分野全体で必要となってきている意識の一つであろう。その一方で、辞書作成やコーパス（テキストデータ）構築のように、時間がかかり独自のノウハウが必要な基盤技術について地道に研究を積み重ねていくことが、自然言語処理研究者に求められている。

参考文献

[1] Ryu Iida, Kentaro Inui and Yuji Matsumoto. Zero-anaphora resolution by learning rich syntactic pattern features. ACM Transactions on Asian Language Information Processing (TALIP). Vol 6, Issue 4, Article 12, 2007.

[2] Hirotoshi Taira, Sanae Fujita and Masaaki Nagata. A Japanese Predicate Argument Structure Analysis using Decision Lists. In Proceedings of the EMNLP, pp.523-532. 2008.

[3] 笹野遼平, 黒橋禎夫. 大規模格フレームを用いた識別モデルに基づく日本語ゼロ照応解析. 情報処理学会論文誌, Vol.52, No.12, pp.3328-3337. 2011.

[4] 吉川克正, 浅原正幸, 松本裕治. Markov Logic による日本語述語項構造解析. 自然言語処理, Vol.20, No.2, pp.251-271. 2013.

[5] 林部祐太, 小町守, 松本裕治. 述語と項の位置関係ごとの候補比較による日本語述語項構造解析, 自然言語処理, Vol.21, No.1, pp.3-25, 2014.

[6] 小町守, 飯田龍. BCCWJ に対する述語項構造と照応関係のアノテーション. 日本語コーパス平成 22 年度公開ワークショップ, pp.325-330. 2011.

[7] 吉本暁文, 小町守, 松本裕治. 複数の分野のコーパスを用いた述語項構造解析の比較 --『現代日本語書き言葉均衡コーパス』を用いて--. 第 3 回コーパス日本語学ワークショップ. 2013.

[8] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality. In Proceedings of NIPS, pp.3111-3119. 2013.

[9] 柏木潔, 小町守, 松本裕治. レビュー文書からの省略された属性の推定を含めた意見情報抽出. 言語処理学会第 19 回年次大会論文集, pp.528-531, 2013.

[10] Mike Minz, Steven Bills, Rion Snow and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In Proceedings of the ACL-IJCNLP, pp.1003-1011. 2009.

[11] Shin Kanouchi, Mamoru Komachi, Naoaki Okazaki, Eiji ARAMAKI and Hiroshi Ishikawa. Who caught a cold? - Identifying the subject of a symptom. In Proceedings of the ACL-IJCNLP, pp.1660-1670. 2015.

[12] Yoshiaki Kitagawa, Mamoru Komachi, Eiji ARAMAKI, Naoaki Okazaki, Hiroshi Ishikawa. Disease Event Detection based on Deep Modality Analysis. In Proceedings of the ACL-IJCNLP Student Research Workshop, pp.28-34. 2015.

この研究は、平成２２年度ＳＣＡＴ研究助成の対象として採用され、平成２３～２５年度に実施されたものです。