学会記事

BERT-based Cohesion Analysis of Japanese Texts

植田 暢大†

1 はじめに

本稿では、COLING2020 に掲載された論文 "BERT-based Cohesion Analysis of Japanese Texts" (Ueda et al. 2020) について、採択までの過程を含め解説する。本論文では日本語テキストにおいて述語項構造や照応関係などの結実性を統合的に解析することを試みた。具体的には、用言に対する述語項構造解析（用言述語項構造解析）、体言に対する述語項構造解析（体言述語項構造解析）、橋渡し照応解析、共参照解析の 4 つのタスクを扱い、それぞれ既存研究を幅広く上回る結果が得られた。また、様々な組み合わせで同時解析を行うことで、共参照関係が他の関係と本質的に異なることを確認し、解析において特別に扱うべきことを示した。なお、本研究の解析システムはウェブ上でデモを公開しており、どちらも試していただけると幸いである。

論文の内容に詳しく触れる前に採択に至るまでの経緯を述べようと思う。この研究は、もっとも 4 つのタスクの中でも用言述語項構造解析を主な対象としていた。これは用言述語項構造解析の精度向上が、省略類が頻繁に発生する日本語において喫緊の課題だったためである。また、用言述語項構造解析は研究が盛んで、解析に有用な特性などの知見が蓄積されていたことも理由の一つである。研究に取りかかった 2018 年秋ごろは先行研究が小部、特性の追加やネットワークの工夫など考えていた。

しかし、BERT (Devlin et al. 2019) の登場で状況は一変した。BERT の fine-tuning により、様々なタスクが簡単かつ高精度で解けるようになった。述語項構造解析も例に漏れず、従来的 BiLSTM ベースのモデルを容易に上回ることができた。

このような背景から、用言述語項構造解析の枠を越え、広くテキスト全体の結実性の解析も視野に入れることが可能となった。用言述語項構造解析は用言・体言間の関係を対象としているが、テキスト全体の結実性には、体言における述語項構造、橋渡し照応関係、共参照関係など体言同士の関係も含まれる。テキスト全体の結実性を捉えるためには各々の解析を独立して実行するだけでは不十分である。テキスト中に存在するこれらの関係は複数の単語を共有し、互いに強く結びついている。したがって、これらの関係には相互作用があると考えられる。その相互作用を捉えるため、以後 4 つのタスクを同時に解析することにした。

† 京都大学
1 https://lotus.kuee.kyoto-u.ac.jp/cohesion-analysis/public/
その後実験を重ね、第26回年次大会（2020年）では「BERTとRefinementネットワークによる統合的照応・共参照解析」という題目で発表を行った。これは4つのタスクの統合的解析に加え、2段階の予測により精度を高める機構（Refinementネットワーク）を導入し、精度を向上させたものである。この論文は栄栄にも大会の若手奨励賞に選んで頂き、次は国際学会で発表しようという運びとなった。

しかし、国際学会への投稿は一筋縄ではいかなかった。RefinementネットワークはBERTベースのモデルを2つ使用しており、単純なアンサンブル学習に対する優位性がなかなか示せなかったのである。国際学会に投稿するにあたり、もう少し主張できるポイントが欲しかったため、その後様々なモデルで実験を行ったが、今ひとつ良い結果が得られなかった。そこで、同時解析を従理解り、その性質を調べてみるという方向に舵を切ることにした。まず、全てのタスクの組み合わせ（2^4 = 16通り）で同時解析の実験を行った。すると、共参照解析のみが同時解析に貢献しないことが明らかとなった。これは共参照関係がその他の関係とは本質的に異なることを意味するが、考えてみれば当然の結果である。共参照解析以外の3つのタスクを俯瞰すると、いずれも述語あるいは名詞に対してその情報を補う上で必須となる単語を予測するタスクである。一方、共参照解析は同一のエンティティを発見するタスクであり、かつ「それ」や「彼」などの照応的性質が強い単語を除けばそこに必須性は関係しない。これを主張の大きなポイントとすることで、無事本論文はCOLING2020に採択されることとなった。

以降、2章で本研究の手法の概略について、3章で実験とその結果について簡単に述べる。

2 手法の概略

本研究では種々の解析を統合的に行うため、それぞれのデータを一様に混合してシステムに入力し、最後にそれぞれの結果を同時に出力する。このような解析を本研究では同時解析とよぶ。

同様解析を行うためにはそれぞれのタスクを同じ枠組みで扱う必要がある。そのため、本研究では項目選択モデルを使用する。項目選択モデルとは、依存構造解析などでも使用されるモデルであり、テキスト中のある語を対象としてその語と係り受けなどの関係を持つ語をテキスト中から選択する枠組みである。この手続きを他の語も対象として行うことで、テキスト全体の関係を解析することが可能となる。

本研究における解析全体の手続きは、以下の3つのステップから構成される。

(1) 入力文書を基本句2単位に分割
(2) 基本句列から解析対象基本句を選択


2 結束性の解析における最小単位。核となる1つの自立語と、その前後に存在する接辞や助動詞などの付属語から構成される。
BERT-based Cohesion Analysis of Japanese Texts

図 1 モデルの概略. 述語「知っている」のガ格 (Nominative) の項を解析する場合を示す。

(3) 解析対象基本句に対して選択モデルを使用し、それぞれの関係を予測
(1), (2) に関しては既存の解析器を使用する. (2) の解析対象基本句とは、例えば用言述語項構造解析なら用言、共参照解析なら体言である.

(3) に関してモデルの概略を図 1 に示す. 図は、述語「知ってる」のガ格の項を解析する場合を表している. モデルは文中の他の全ての基本句について述語「知っている」のガ格の項らしきのスコア \( s_{\text{NOM}}(t_i, t_j) \) を計算し、その内で最もスコアが高かったものを「知ってる」のガ格の項として採用する. 図 1 では「知ってる」のガ格の解析のみを示したが、この計算をガ格など他のガ格について、さらに文中の全ての述語に対してそれぞれ行う. 共参照解析や橋渡し照応解析もこの枠組みで解くことができる.

選択モデルでは系列中から必ず一つの基本句が選択されるが、述語照応文中には項や先行詞が存在しないのも存在する. また、外界照応の場合、文章中対象は出現しない. そこで、入力系列に特殊トークンを加えることでこれらの対象に対応する. 本研究では、特殊トークンとして [著者]・[読者]・[不特定:人]・[NULL]・[NA] の 5 つを使用する. [著者]・[読者]・[不特定:人] は外界照応解析におけるターゲットとして使用される. [NULL] は述語項構造解析・橋渡し照応解析において項や先行詞が存在しない場合の, [NA] は共参照解析において共参照関係にあるメンションが存在しない場合のターゲットである. 詳細についてはソースコードを公開しているのでそちらを参照してください.

3 実験

実験では、BERT を用いることで述語項構造解析の中でも特に難しいゼロ照応解析や、共参照解析のスコアが大きく改善することを示した. さらに、橋渡し照応解析においても F1 スコア

---

3 BERT を使用する都合上、実際の処理はサブワード単位で行われる。
4 https://github.com/nobu-g/cohesion-analysis
で、0.5〜0.6と、ある程度の精度で解析が可能であることが分かった。しかし、意外なことに格解析のスコアはBERTを使用していない既存研究（Shibata and Kurohashi 2018; Kurita et al. 2018）を下回る結果となった。これは、本研究が既存研究とは異なり、係り受けや選択選好などの素性を使用していなかったためと考えられる。論文には記載していないが、実際モデルに正解の係り受け情報を与えることで格解析を中心にスコアが向上することも確認されている。したがって、今後の展望として係り受け解析もシステムに組み込むことが効果的かもしれない。

同時解析の効果に関して、用言述語項構造解析と体言述語項構造解析は関係セットが同じであることもあり、相互に利得があることが確認された。一方、共参照解析については、他のタスクに加えてもそのタスクのスコアを向上させることはなかった。また、共参照解析のスコアは、共参照解析を単独で学習する場合が最も良く、他のタスクを加えることで低下した。したがって、共参照解析は単純な同時解析ではシステムの性能に寄与せず、特別に扱う必要があると考えられる。本論文では共参照解析を特別に扱うモデルを作成し、評価も行ったが、十分な性能は得られなかった。共参照の情報をより効果的に他のタスクに利用できるようなモデルの考案が今後の課題である。

システムの実際の解析例を図2に載せる。システムは「中国が援助」や「基地を建設」など、比較的表層的な特徴から予測が可能なものについては正しく解析できている。一方で、「中国の海軍」や「ミャンマーに建設計画」などは予測していなかった。これらは文書の内容に関するより深い理解が必要となるような例であるが、他の関係を考慮すれば正しく解析できる可能性がある。例えば「中国の海軍」は、正しく解析できている「海軍の基地を中国が建設」という関係から推論が可能である。したがって、本論文の手法では未だ十分に結論性を捉えられているとは言えず、他の関係の情報をさらに解析に活用していくことが今後の課題である。

図2 システムの解析例。下側の矢印はシステムの予測を表す。破線はシステムが予測を誤った例であり、上側にシステムが正しく予測できなかった正解を示す。
4 おわりに

用語句に対する述語項構造解析は基礎解析の中でも特に難しいタスクであり、形態素解析や構文解析とは異なりこれまで実用化は難しかった。しかし、本研究により実用化に耐えうる程度の解析が可能になったと考えている。実際に本研究の成果は、情報集約システムである因果関係グラフ（清丸 他 2020）において利用されている。

その一方、未だ体言に対する述語項構造解析や橋渡し照応解析には大きな改善の余地が存在する。これらタスクについても十分な精度での解析を可能にし、ゆくゆくはテキスト中のあらゆる結び性を捉え、高度に構造化されたテキストを得ることが目標である。

参考文献

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” In Proceedings of NAACL, pp. 4171–4186.

清丸寛一、植田暢大、児玉貴志、田中佑、岸本裕大、田中リベカ、河原大輔、黒橋禎夫（2020）。因果関係グラフ：構造的言語処理に基づくイベントの原因・結果・解決策の集約。言語処理学会第26回年次大会発表論文集，pp. 1125–1128，茨城。[H. Kiyomaru et al. (2020). Inga Kankei Gurafu: Kozo-teki Gengo Shori ni Motozuku Ibento no Genin · Kekka · Kaiketsusaku no Shuyaku. Proceedings of the 26th Annual Meeting of the Association for Natural Language Processing, pp. 1125–1128, Ibaraki].

Kurita, S., Kawahara, D., and Kurohashi, S. (2018). “Neural Adversarial Training for Semi-supervised Japanese Predicate-argument Structure Analysis.” In Proceedings of ACL, pp. 474–484.

Shibata, T. and Kurohashi, S. (2018). “Entity-Centric Joint Modeling of Japanese Coreference Resolution and Predicate Argument Structure Analysis.” In Proceedings of ACL, pp. 579–589.

Ueda, N., Kawahara, D., and Kurohashi, S. (2020). “BERT-based Cohesion Analysis of Japanese Texts.” In Proceedings of the 28th International Conference on Computational Linguistics (COLING 2020), pp. 1323–1333, Barcelona, Spain (Online).

略歴

植田 慶大：2019年京都大学工学部電気電子工学科卒業。2021年京都大学大学院情報学研究科修士課程修了。同大学院博士後期課程に進学。現在に至る。