Contextualized and Generalized Sentence Representations by Contrastive Self-Supervised Learning: A Case Study on Discourse Relation Analysis

清丸 寛一[†]

1 はじめに

本稿では NAACL2021 に採択された筆者の表題の論文 (Kiyomaru and Kurohashi 2021) を紹介する。採択論文は、複数文からなる入力テキストに対して各文の表現を計算するモデルを考え、それを対照学習 (contrastive learning) の枠組みで学習することを提案するものである。

採択論文の紹介に入る前に、まずは着想に至った経緯について述べたい。2020 年の中頃、筆者は日本語談話関係解析 (岸本 他 2020) の解析精度の向上に取り組んでいた。ここ数年の談話関係解析に対する主流なアプローチはニューラルネットワークにもとづくものであり、解析精度の向上を目指して様々な構成のモデルが提案されていた。そこに汎用言語モデルの BERT (Devlin et al. 2019) が登場し、BERT の上でごく単純な分類器を学習するだけで既存手法を大きく上回る性能が得られることが分かった。これは、あれこれと凝らした解き方の工夫を補って余りある程に BERT が作り出す談話単位の表現が優れていることを意味していた。これを受け、談話関係解析が“解けてしまう”表現を大規模コーパスから自己教師あり学習する方向性に更なる精度改善の可能性を感じていた。

一方で、第二著者である黒橋先生と議論する中で、典型的な話題遷移にもとづくトップダウンな意味理解が談話関係解析の精度改善に寄与するのではないかという構想が膨らんでいた。たとえば、ある問題に対する解決方法をアピールする文章にはしばしば「問題 → 問題の影響 → 解決方法」という話題遷移が見て取れる。こうした典型的な話題遷移があることを念頭に文章を読めば、文や段落の大意がよく分かり、結果として談話関係を上手く読み取れるのではないかという構想である。

しかし、解析対象はウェブテキストコーパス (岸本 他 2020) である。大半のテキストは好きなことを好きな順番で書いていて、文の意味はボトムアップに理解せざるを得ないケースばかりかもしれない。そうだとしたら、典型的な話題遷移を考慮した手法を考えても精度改善には繋がりにくい。

[†] 京都大学大学院情報学研究科

この構想に突き詰めるだけの価値があるのか、コーパスに収録されている文章を眺め、典型的な話題遷移を探すことを通して検討した。典型的な話題遷移を探すといっても、どんなものがどれだけの種類あるかも分からないので、まずは似たようなドメインの文章を集めて眺めることから始めた。一見して典型的な話題遷移が見て取れたのは広告ドメインの文章だった。ほとんどが製品・サービスの導入に始まり、その（ポジティブな）特徴の説明が続く。こうしたわかりやすい例を足がかりに、コーパス中の文書に談話単位レベルで意味ラベル—たとえば「サービスに関する特徴（ポジティブ）」といったラベル—を付与することを試みた。意味ラベルは、話題遷移の観点で類似した文章をひとまとめにできる程度に汎化された、しかしなるべく具体的なものを考えるようにした。付与した意味ラベルをもとに、類似の話題遷移をしている文章を見つけ、そうした文章をまとめられるような意味ラベルを再付与する、という作業を何度か繰り返した。

検討を続ける中で、ウェブ上の多様なテキストの中にもいくつかの典型的な話題遷移が認められそうなのことが分かり、それを談話関係解析に利用することにそれなりの手応えを感じ始めた。しかし、意味ラベルの粒度は終始悩ましく、一貫性のあるアノテーション基準を設けて教師データを作成しそれを学習するという方向性は茨の道に思えた。

そこで、意味ラベルを陽に与えて学習する代わりに、意味ラベルを付与するためにわれわれが取った手続き—前後の文脈が/から分かる程度に汎化された文の意味を考える手続き—そのものをタスク化し、それを最適化することで典型的な話題遷移を考慮するモデルを構築しようと考えた。そこに自己教師あり学習が有望そうだという昨今の知見が合わさり、採択論文のアイデアが生まれた。

2 提案手法

提案手法では、複数文からなる入力テキストに対して各文の表現を計算するモデルを考え、それを対照学習によって最適化する。手法の概要を図 1 に示す。

2.1 エンコーダ

エンコーダには BERT と同じ構成の Transformer を利用する。入力テキスト中の各文に対する表現を計算する方法として、本研究では Liu and Lapata (2019) の方法を採用する。この方法では、各文の直前に特殊トークンである〈CLS〉を挿入し、その最終層の出力を文表現として利用する。

図 1 提案手法の概要.

2.2 対照学習にもとづく文表現学習

文表現は対照学習の枠組みで学習する．まず入力テキスト中の一文をターゲット文として無作為に選ぶ．図 1 では， $k(1 \leq k \leq T)$ 番目の文がターゲット文である．target sentence の文表現を s_{pos} と呼ぶ．次に，ターゲット文を特殊トークンである $\langle \text{SENT-MASK} \rangle$ に置き換えた入力テキストを用意する．その $\langle \text{SENT-MASK} \rangle$ の文表現を s_{anc} と呼ぶ．最後に，ターゲット文を無作為な文で置き換えた入力テキストを用意する．その無作為な文の文表現を s_{neg} と呼ぶ．

提案手法では s_{pos} と s_{anc} の類似度を最大化しつつ s_{neg} と s_{anc} の類似度を最小化することを学習する．

$$\mathcal{L} = -\log \frac{\exp(\langle s_{pos}, s_{anc} \rangle)}{\sum_{s \in \mathcal{S}} \exp(\langle s, s_{anc} \rangle)}, \quad (1)$$

ここで $\langle \cdot, \cdot \rangle$ はベクトルの内積， N は target sentence を置き換えるのに利用する無作為な文の数， $\mathcal{S} = \{s_{pos}, s_{neg}^1, \dots, s_{neg}^N\}$ である．

s_{anc} は s_{pos} に近い表現であることを求められるわけだが，文脈のみからターゲット文の内容を完全に予測することはできないため， s_{anc} は文脈的に妥当な文が共通して持つ意味—すなわちターゲット文の汎化された意味—を捉えるようになると期待される． s_{pos} もまた s_{anc} に近い表現であることを求められるため， s_{pos} はターゲット文の内容を観測した上でその意味を汎化して捉えるようになると期待される．

提案手法は，文分割の規則さえ定めてしまえば生コーパスから学習事例を無尽蔵に作成できるため，自己教師あり学習の枠組みで広範な言語表現を学習することが可能である．

文脈	エンコーダ	正解率	適合率	再現率	F 値
利用しない	BERT _{BASE}	80.68 $\pm$ 1.59	45.90 $\pm$ 4.06	41.42 $\pm$ 8.35	43.37 $\pm$ 6.28
利用する	BERT _{BASE}	84.36 $\pm$ 2.05	62.55 $\pm$ 10.26	39.13 $\pm$ 7.38	47.67 $\pm$ 6.68
	Ours	85.02 $\pm$ 1.85	63.51 $\pm$ 5.90	41.04 $\pm$ 4.24	49.74 $\pm$ 4.11

表 1 KWDLC における正解率, 適合率, 再現率, F 値.

3 実験：談話関係解析

提案手法の有効性を談話関係解析の性能を通じて評価した. 実験は英語と日本語で行ったが, ここでは日本語の結果に絞って紹介したい.

日本語の談話関係解析のデータセットとして KWDLC (岸本 他 2020) を利用した. KWDLC はウェブ文書の冒頭三文を集めたコーパスであり, 三文内にある談話単位の全ての組み合わせについて談話関係がアノテーションされている.

談話関係解析のモデルとして, BERT (Devlin et al. 2019) を土台に, 談話単位の文脈を利用するモデル, 利用しないモデルの二つを考えた. 文脈を利用するモデルには三文すべてが入力として与えられる. モデルは 2.1 節で説明した方法で各談話単位に対する表現を計算する. そして, 談話関係を認識したい二つの談話単位の表現を結合したものを三層パーセプトロンに渡し, 談話関係を予測する. 文脈を利用しないモデルには, 談話関係を認識したい談話単位のペアのみが与えられる. モデルは Devlin et al. (2019) が提案する sentence pair classification の方法で談話関係を予測する.

表 1 に結果を示す. 文脈を利用するモデルは, 利用しないモデルと比べて性能が大きく向上しており, 提案手法を用いることで更にその性能が向上することが分かった.

提案手法によって改善した一例を紹介したい.

- (1) $\langle$ Arg1 以前から計画していたホームページを開設することができ、 $\rangle$ $\langle$ Arg2 嬉しいかぎりである。 $\rangle$

談話関係: 原因・理由

談話単位 Arg1 と Arg2 の間には「原因・理由」の談話関係がある. 提案手法を利用したモデル以外が「談話関係なし」の談話関係を予測する中, 提案手法を利用したモデルは正しく「原因・理由」の談話関係を予測することができた. 体験・経験の話の後にそれに対する感想・心情が続くのはウェブコーパスを観察する中で認められた典型的な話題遷移の一つであり, その間にはしばしば「原因・理由」の談話関係がある. 提案手法は, 例文の談話単位の意味をそのレベルの抽象度で汎化して捉えることができたために, 談話関係の認識に成功したのではないかと考えている.

4 おわりに

採択論文では、文脈化文表現を対照学習によって学習する手法を提案した。実験では、提案手法を英語・日本語の談話関係解析に適用し、提案手法によって性能が改善することを示した。提案手法によって学習される文表現の汎化の程度を調べることは今後の課題である。そのための足がかりとして、文表現の類似度にもとづく文検索の実験を行っており、採択論文にはその結果も掲載している。興味を持たれた方は参照されたい。

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 the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- 岸本裕大, 村脇有吾, 河原大輔, 黒橋禎夫 (2020). 日本語談話関係解析: タスク設計・談話標識の自動認識・コーパスアノテーション. 自然言語処理, **27** (4), pp. 889–931. [Y. Kishimoto et al. (2020). Japanese Discourse Relation Analysis: Task Definition, Connective Detection, and Corpus Annotation. *Journal of Natural Language Processing*, **27** (4), pp. 899–931.].
- Kiyomaru, H. and Kurohashi, S. (2021). “Contextualized and Generalized Sentence Representations by Contrastive Self-Supervised Learning: A Case Study on Discourse Relation Analysis.” In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5578–5584, Online. Association for Computational Linguistics.
- Liu, Y. and Lapata, M. (2019). “Text Summarization with Pretrained Encoders.” In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3730–3740, Hong Kong, China. Association for Computational Linguistics.

略歴

清丸 寛一：2017年同志社大学理工学部情報システムデザイン学科卒業。2019年京都大学大学院情報学研究科博士前期課程修了。現在、京都大学大学院情報学研究科博士後期課程に在籍中。