E-005

日本語文中における意味的に重複した語句の検出手法の検討
Study of Semantic Pleonasm Detection in Japanese Sentences

衣川 和尭† 美野 秀弥† 河合 吉彦†
Kazutaka Kinugawa Hideya Mino Yoshihiko Kawai

1. はじめに

本稿では、日英ニュースの対訳コーパスから抽出したデータセットで日本語文中の意味的に重複した語句の検出に取り組む。一般に、文中にこうした語句が含まれている場合、これを適切に削減することで、文章の簡潔性を向上させ、原稿の限られた紙面を有効活用できる可能性がある。特に英語ニュースではこうした簡潔性が重視されている [1] 。一方、こうした削減は文意を曖昧にする恐れがあり、逆に同一の情報を繰り返すことで読み手にとって正確な文章を提供できるという側面もある。実際に記事を読むと、日本語ニュースは英語とは対照的に後者の性質を重視していることが観察される。日・英ニュース間のこのようなスタイルの違いを考慮した機械翻訳器を実装するためには、日本語文中から意味的に重複した語句を検出する必要がある。さらに、書き手が意図していない重複も検知できれば、機械翻訳に限らず日本語単言語のシナリオでも有益である。

上述の語句は意味論的な冗語と呼ばれており、よく認知された事象であるが、これを取り扱う研究は限定的である。これは、取り扱う問題が文法よりもスタイルに関するものであるため、冗長か否かを明確に定義することが難しいことが理由の一つとして挙げられる。また、テキスト中で頻繁に観察される事例ではないため、分量を確保することが容易ではないことも挙げられる。こうした理由から、コーパスが最も豊富な言語である英語においてすら、この問題を取り扱うデータセットは限られている [2] 。

この２つの課題に対し、本稿ではまず対訳文を手がかりに事例を収集する。対訳コーパスにおいては文意が等価な日本語文と英語文をペア付けすることができるが、上述したスタイルの違いから日本語側で意味的に重複した語句が英語側では削減して訳出されていることがある。これは翻訳時にある種のリライトが行われていると解釈できる。単言語データのみから事例を収集する場合は先に述べた簡潔性と非曖昧性のトレードオフの問題から明確なアノテーション規定の策定が難しくなるが、本稿ではその研究の動機から収集すべき正例の要件が自然に定義されている。次に、収集した事例に対して大規模言語モデル (Large Language Model: LLM) を用いて意味的に重複した語句の検出を試みる。LLM は few-shot learning [3] と呼ばれる、複数の事例をプロンプトに含めて推論時の性能を向上させる技術が知られており、訓練事例の分量が限られている状況下で性能を向上させる効果が期待できる。本稿では研究用のオープンな LLM を用いて、文中の意味的な重複の有無およびそのスパンの特定を目的としたプロンプトを設計し、性能を評価した。実験結果から few-shot learning により意味的重複の有無の検出で最大 4% 程度の精度向上が認められた一方、スパンの特定に課題があることがわかった。

† NHK 放送技術研究所 NHK Science & Research Laboratories

2. データセットの構築

2.1 設定

まず、収集する事例の定義と問題設定について述べる。本稿では、意味的に重複している語句の中でも表層が一致しているものに注目し、対訳文を手掛かりに適当な事例を収集する。例として簡単な対訳を以下に示す。

彼がデータを解析して解析結果をメールで送るだろう。
He would analyze the data and email the results.

ここでは“解析”という単語が２回現れるが、後者の“解析結果”は対訳文では単に“results”（“結果”）と訳出されている。これは、“結果”が“**解析の**結果”であることが文脈的に推測可能であるためである。つまり、２つ目の“解析”は削減可能であると考えられ、これを根拠に文中の“解析”同士が意味的に重複していると判断できる。意味論的な冗語は一般に“頭痛が痛い”のように明らかに意味が冗長なものを指すことが多いが、本稿では、意味が冗長とまでは言えないが削減可能なものを扱う。

事例の具体的な定義は次の通りである。ある日本語文 x が、同一の表層 w を持つ単語を n ($n > 1$) 個含むとする。それらの単語を $(w_1, \ldots, w_n)$ としたとき、このうちの１つを含むスパンを x から取り除いてもその意味が文脈から推測できる（つまりスパンを削減しても文意が変わらない）場合に、そのラベル y を正例と定義する。ここで言及しているスパンは“解析の”のように表層 w だけでなく隣接する助詞なども含み得ることを想定している。削減可能なスパンが含む w のインデックスを i （上記の例であれば $i = 2$ ）、このスパンを削除した結果の文を x^* として、これらをまとめた (x, w, n, x^*, i, y) を１つの事例として扱う。同一の表層 w が複数回表出する文であっても、次の条件を満たすもののラベルは負例と定義する。w を１つ含むスパンをどのように取っても、それを削減することで (1) 文法的に破綻する、あるいは、(2) 文の意味が変わってしまう。例えば、“***男性からの支持が女性からの支持に繋がる。***”という文では、同一の表層 （“支持”）が複数あってもこれらは互いに異なる項を持ち、意味が重複しているわけではない。本タスクでは (x, w, n) を入力として (y, i) を正しく出力することを目的とする。（負例の場合は x^* および i は定義せず、予測するのは y のみとする。）x^* は 3 節で述べる few-shot learning に用いる。

2.2 事例の収集

時事通信社の日英ニュース記事集合より事例を収集する。まず日本語記事と英語記事を記事 ID で照合し、同一のニュースを報じている記事同士をペア付けする。次に、内山・井佐原の手法 [4] を用いて、各記事対において日本語文 x と英語文 x' で、表層的な類似度をもとに内容的に等価とみ

なせる文同士をペア付けする。ここで、類似度が 0.5 未満のペアは候補から除外する。残ったペアについて、日本語文 x 側で GiNZA[‡] を用いて形態素解析を行い、同一の表層 w が複数回現れる文を抽出する。最後に、対訳文を読んで、2.1 節で示した例のような対訳側で削減が起きているものを正例として人手で選ぶ。同様に、対訳側で w の削減が起きていないペアの中から、 2.1 節で定義した条件を満たすものを負例として人手で選ぶ。これらの作業を通じて正例・負例それぞれ 48 事例ずつ収集した。正例・負例ともに 8 事例ずつを開発データとして選び、残りをテストデータとする。

3. 実験

3.1 設定

モデルは日本語版 LLM である Swallow-{7b,13b}-instruct-v0.1 [5] を用いる。プロンプトの指示文のテンプレートは次の通りである。“***『x』という文には『w』という単語が n 個含まれています。このうちいずれか 1 つを含むスパンを削除し、かつ、それ以外の部分を書き換えることなく、同じ意味の文を作成することはできますか？「はい」か「いいえ」で回答してください。「はい」の場合は、何番目の『w 』を抜き取ったかも回答してください。***”ここでは、「はい」、「いいえ」の回答がラベル y の予測、抜き取った w の番号がインデックス i の予測にあたり、これらの精度を評価する。指示文に n を含めているのは、モデルの予測が $1 \leq i \leq n$ の範囲に収まるよう促すことを意図している。y, i が出力されなかった場合は、これらが得られるまで同一のプロンプトを入力し続ける。

few-shot learning のデモンストレーションの数を k ($k \in \{0, 1, 2, 4\}$) とする。正例・負例を k 個ずつ開発データからランダムサンプリングし、その入出力を上記の指示文の後に挿入する。さらに、Chain of Thought (COT) [6] も導入する。具体的には、削減後の文を提示させたあとでラベルとスパンを回答させることでこれらの正確な予測を促進する。few-shot learning では次の文をプロンプトに挿入する。正例の場合は“***理由: 与えられた文は『x^*』と書いても意味が変わらないため。***”、負例の場合は“***理由: いずれの『w』も、抜き取ると文法が破綻したり、文の意味が変わってしまうため。***”とする。

3.2 結果

結果を表 1 に示す。y の予測精度 Acc (label) と (y,i) の予測精度 Acc (label+span) を載せている。表 2 より、ラベルの予測には few-shot learning で最大 4% 程度の性能向上が認められる一方で、スパンの認識精度には課題がある。この傾向は Kashefi らの実験結果 [2] と同様である。エラー分析として COT 有りのモデルが出力した“*理由：…*”を観察すると、スパン削減の結果として提示している文が実際には削減されていないパターンが多い。また、指示に反してスパン以外の書き換えを行っているなど、指示追従能力の不足も確認された。この観察から、性能の改善のためには、スパン削除や文同士の意味の等価性判定など本タスクに必要なサブタスクの訓練や、サブタスクに分解しての処理などの指示形式の工夫が必要となると考えられる。

‡ https://megagonlabs.github.io/ginza/

表 1　実験結果

Size	# Shots	Acc (label) (%)	Acc (label+span) (%)
7b	k=0	50.0	20.0
	k=1	50.0	21.3
	k=2	52.5	16.3
	k=4	51.3	13.8
	k=0 w/ COT	50.0	22.5
	k=1 w/ COT	47.5	8.8
	k=2 w/ COT	51.3	12.5
	k=4 w/ COT	**53.8**	8.8
13b	k=0	50.0	**23.8**
	k=1	47.5	17.5
	k=2	50.0	12.5
	k=4	52.5	16.3
	k=0 w/ COT	50.0	22.5
	k=1 w/ COT	42.5	11.3
	k=2 w/ COT	**53.8**	17.5
	k=4 w/ COT	47.5	11.3

4. おわりに

本稿では、日英対訳コーパスを用いて日本語文中で意味的な重複を含む語句の検出タスクのデータを構築し、 LLM による検出実験を行った。今後はスパン削減などサブタスクも考慮したプロンプト設計に取り組む。

謝辞

貴重なデータをご提供いただいた時事通信社の朝賀英裕氏と川上貴之氏に感謝を申し上げます。本研究成果の一部は、国立研究開発法人情報通信研究機構の委託研究 (課題 225) により得られたものです。

参考文献

[1] Rene J. Cappon, “The Associated Press Guide to News Writing”, Peterson’s, fourth edition, (2019).
[2] Omid Kashefi, Andrew T. Lucas, and Rebecca Hwa, “Semantic pleonasm detection”, Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), (2018).
[3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever and Dario Amodei, “Language Models are Few-Shot Learners”, arXiv, (2020).
[4] Masao Utiyama and Hitoshi Isahara, “A Japanese—English patent parallel corpus”, Proceedings of Machine Translation Summit XI:Papers, (2007).
[5] Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota and Naoaki Okazaki, “Continual Pre-Training for Cross-Lingual LLM Adaptation: Enhancing Japanese Language Capabilities”, arXiv, (2024).
[6] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou, “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models”, Advances in Neural Information Processing Systems, vol 35, (2022).