日本語の論文コーパスにおける「問題」の語義アノテーション

平林　照雄
茨城大学理工学研究科
20nd303t@vc.ibaraki.ac.jp

河野　慎司
茨城大学理工学研究科
20nm709n@vc.ibaraki.ac.jp

古宮　嘉那子
茨城大学理工学研究科
kanako.komiya.nlp@vc.ibaraki.ac.jp

新納　浩幸
茨城大学理工学研究科
hiroyuki.shinnou.0828@vc.ibaraki.ac.jp

1　はじめに

学術論文において、主題は問題解決プロセスであることが多い。また、問題を明示するため、「問題」をはじめとする特徴語が、同一文または、周辺文に存在すると予想される。従って、「問題」のような特徴語から、その問題内容を情報抽出するタスクを考えることができる。Heffernan ら [1] は英語科学論文からパターン抽出を利用して “problem” に注目した情報抽出を行った。しかし、日本語においては、「問題は〜である。」と明記されていることは稀であり、パターン抽出による情報抽出手法をとることが難しい。また、「問題」は多義語であるから、語義曖昧性解消を行う必要がある。

本研究では、科学論文の問題解決プロセスにおける問題内容を特徴語から情報抽出を行う準備として、「問題」の語義タグ付きコーパスの作成を行う。さらに、作成したコーパスを用いて自己学習によるコーパスの増補を行う。

2　関連研究

Heffernan ら [1] は、“problem” とその類義語を利用した情報抽出により、英語科学論文から問題提起箇所を含む「問題内容」と「解決法」コーパスを作成し、さらに、文中に「問題内容」または「解決法」を含むか否かという分類を行った。同論文で Heffernan らは、“problem” は “problematic” と “task” の二つの意味を持ち、問題解決プロセスにおいては、“problematic” の意味の “problem” のみが対象になるとしており、この意味の “problem” を抽出するにあたり、機械的にパターン抽出を行った後で、問題提起を行わない文を人手で除外している。

本研究では、Heffernan ら [1] の研究に倣い、日本語の「問題内容」と「解決法」を抽出することを目的とし、その準備として、日本語論文コーパス中の「問題」の語義曖昧性解消を行う。

日本語のコーパスに語義タグを付与する研究は多く、Shirai ら [2] や Okumura ら [3] の研究がある。Shirai ら [2] は、『岩波国語辞典』に基づいた語義タグを新聞コーパスに付与している。さらに、Okumura ら [3] は Shirai ら [2] と同様の語義タグを『現代日本語書き言葉均衡コーパス』[4] に付与した。また、語義曖昧性解消のタスクにおいて、対象単語の周辺単語を用いることは一般的である。例えば、Komiya ら [5] は、多義語の周辺に現れる語義の分布を利用する周辺語義モデルを提案している。自己学習を語義曖昧性解消に用いる手法も一般的である。Yarowsky ら [6] らや鈴木ら [7] は、教師なし学習による語義曖昧性解消の精度向上のため、ブートストラップ法による自己学習を行った。

3　提案手法

本研究では、初めに、「問題」の語義のアノテーションルールを定め、人手によるタグ付けを行った。次に、作成した語義タグ付きコーパスをもとに分類器を学習し、平文コーパスに適用することで語義タグ付きコーパスの増補を行った。

3.1　「問題」の語義

Heffernan ら [1] は “problem” の多義性について指摘したが、筆者らは日本語の「問題」も同様の多義性を持ち、さらに同様の語義を含むと考えた。『岩波国語辞典』によると、「問題」は表 1 のように定義されている。問題解決プロセスの際に使用される「問題」の意味は<1><イ>の意味をとる。この意味を先行研究に合わせ、“problematic” の意味と

表 1 『岩波国語辞典』における「問題」

	もんだい【問題】
<1>	答を求めて他が出しまたは自分で設けた、問い。
<ア>	実力をためしたり練習したりするための問い。「算数の―」
<イ>	研究・議論により、または策を講じて、解決すべき事柄。「―提起」「この計画にはまだ解決すべき―（点）が多い」「死活の―」「―にならない」（取り上げる価値がない）「時間の―」（↓じかん（時間））
<あ a>	▽難点の意にも使う。「いささか―がある」
<2>	問題 (1) に似たあり方のもの。
<ア>	扱いが面倒な事件。「また女の事で―を起こした」
<イ>	人人の注目を集めている、また集めてしかるべきこと。「これが―の人物です」「最近の―作」

呼ぶ。また先行研究で主張された “task” の意味は「問題」<1><ア>の意味に相当する。同様に先行研究に合わせ “task” の意味と呼ぶ。 そこで、「問題」の語義のアノテーションルールを作成する際には、“problematic” の意味と “task” の意味の意味決定に重点を置いた。

3.2 アノテーションルール

「問題」という単語を含む一文内の情報によって「問題」が “task”,“problematic”,“non-problematic”(後述の第一ルールで定義を示す),「それ以外」のいずれかを判定するアノテーションルールを作成した。また、各ルールは原則第一ルールから順番に適用を行うが、第十ルールのみ、アノテーションルールの策定後に追加で作成されたルールである。そのため、作業時には、例外的にルールの適用の優先度を高くしてほしいと作業者に説明を行った。作業者は自然言語処理を専門とする学生である。

第一ルール

「～は問題でない」「～は問題にならない」 等を含む文の「問題」は “non-problematic” 判定を行う。

これらを含む文の周辺文に問題に対する解決法が期待されないため設定した。

第二ルール

固有表現を含む「問題」に対しては、固有表現を含む内容であることを明記し、可能であるならば “problematic”、“task” の判定を行う。

第二ルールが適用される文では、専門的な内容が多く、しばしば作業者の知識を必要とするため固有表現が適用されることを明示した後、判断不可とすることを許した。例としては、「ゼロ頻度【問題】」などがあげられる。

第三ルール

「問題点」中の「問題」、または「問題」を「問題点」に置換可能である際、“problematic” 判定を行う。

タグ付けの際に本ルールが適用された文を以下に示す。

> 文書連想検索を実現する際の【問題】点は，類似文書の検索に時間がかかることである．

第四ルール

研究課題の「問題」である際、“task” 判定を行う。

このルールにより付与された “task” は、後述する第八ルールでの “problematic task” が頻出することに留意する。

タグ付けの際に本ルールが適用された文を以下に示す。

> 本研究は，一つの用語から，それに関連する用語集合を収集するという【問題】を扱っている．

第五ルール

実験・研究中に発生した又は考えられる「問題」である際、“problematic” 判定を行う。

タグ付けの際に本ルールが適用された文を以下に示す。

> また、ある文書空間内での共起情報を用いれば関連度を計算可能と思われるが，どのような文書空間を用いるべきかが【問題】となる．

第六ルール

解答を求める問いや、試験などの問い、「問い」と置換できる「問題」である際、“task” 判定を行う。

例としては、「数学の【問題】を解く。」や、「入試【問題】」があげられる。

第七ルール

「疑問点」と置換可能である場合、“task” 判定を行う。「困難」と置換可能である場合、“problematic” 判定を行う。

第八ルール

これまでのルール内で “task” 判定を行ったもののうち、“problematic” の意味を持つ “problematic task” の場合、“problematic” 判定を行う。

“problematic task” は作業者の判断が最も揺れる例である。

タグ付けの際に本ルールが適用された文を以下に示す。

> また，英語は他の欧州諸言語と比較して，性・数・格に応じた活用などが簡略化された言語として有名であり，語形から統語情報が失われることで発生する曖昧性の【問題】もある．

第九ルール

一文以上の文脈を見ないと決められないものは「それ以外」判定を行う。

第十ルール

「対処する」、「対処される」又は「解消する」、「解消される」等の「問題」は “problematic” 判定を行う。

タグ付けの際に本ルールが適用された文を以下に示す。

> この【問題】に対処するための手法がいくつか提案されている。

3.3 アノテーション実施方法

作成したアノテーションガイドラインを作業者に周知し、人手でタグ付けを行う。この時、作業者 2 名一組で判定を比較する。判定が一致する場合、正解データとしてタグ付けを行う。一致しない場合、どちらの判定が正しいかの最終判断は、筆者が実際にタグ付けを行う文を確認し行った。

4 実験

タグ付けを行う対象コーパスは、『言語処理学会論文誌 LaTeX コーパス』[1)] を用いた。人手によるタグの付与数及び、κ 値を表 2 に示す。ただし、11 人の作業者にアノテーションを依頼し、2 人 1 組ずつ評価を行ったため、ランダムに作業者を疑似的な作業者 A と B に割り振り、疑似的な作業者 A と B 間の κ 値を求めている。コーパス増補後のタグの付与数を表 3 に示す。

表 2 人手による「問題」コーパス

人手によるアノテーション数	239
problematic	143
task	85
non-problematic	6
それ以外	5
κ 値	0.449

表 3 増補後の「問題」コーパス

増補後のアノテーション数	5321
problematic	3398
task	1796
non-problematic	59
それ以外	68

コーパスの増補に用いた分類器は線形の SVM を用いる。素性として、周辺 3 単語に対象の「問題」を含む、7 単語を新納ら [8] が作成した短単位分散表現辞書 NWJC2vec を用いて分散表現化し、連結したものを使用する。

5 コーパスの評価

人手により作成された語義タグ付きコーパスから学習した分類器と、コーパス増補後の語義タグ付きコーパスから学習した分類器を、Okumura ら [3] が『岩波国語辞典』の語義タグを『現代日本語書き言葉均衡コーパス』[4] に付与したコーパスに適用することで、精度評価を行う。

5.1 実験設定

評価に用いる文は、『岩波国語辞典』の語義タグが付与された「問題」のうち、対応が取れている “problematic” と “task” に対応する語義のみを使用した。また、評価用コーパスが “problematic” と “task” の正解データのみを持つため、“problematic” と “task” のタグが付与された文のみを用いて分類器の学習を行った。分類器の素性及びアルゴリズムはコーパスの増補と同様のものを使用する。実際に評価に使用した「問題」のアノテーション数を表 4 に示す。

1） https://www.anlp.jp/resource/journal_latex/index.html

表 4　『岩波国語辞典』の語義が付与された「問題」の数

評価用アノテーション数	171
problematic	157
task	14

5.2　実験結果

学習した分類器のそれぞれの精度を表 5 に示す。

表 5　各分類器の精度

人手コーパスから学習した分類器	0.696
増補後のコーパスから学習した分類器	0.713

6　考察

表 5 の結果から、精度が高いが少量のコーパスである人手コーパスから学習した分類器が、自己学習の結果によって、精度の高い分類器を作成できた。このことから、自己学習が効果的であることがわかった。

作成したアノテーションルールは、できる限り作業者による差異が少ないルールの作成に努めたが、第八ルールの “problematic task” の判断については作業者がその “task” を “problematic”、つまり「解決すべきこと」と考えるか、に依存しており、作業者による判断の差異が生まれやすい。

コーパスの評価のため用意した『岩波国語辞典』の語義タグが付与された「問題」は、“problematic” の語義が多く、「問題」の語義を単純に “problematic” と分類する分類器が強力である。そこで、評価に用いた「問題」のうち、“problematic” と “task” の語義をそれぞれランダムに 10 文ずつ選択した合計 20 文に対して、評価で作成した分類器を適用する追加実験を行った。この時の分類器の精度を表 6 に示す。分類器の精度は実験毎に変化するため、2000 回試行し、平均をとった。表 6 より、増補後のコーパスから作成された分類器はランダムよりわずかに良いことがわかった。

表 6　追加実験における各分類器の精度

人手コーパスから学習した分類器	0.476
増補後のコーパスから学習した分類器	0.519

コーパスの増補によって分類器の判定が改善された一例を挙げる。

> サーバーまたはネットワークに【問題】があるか、またはアイドル時間が長すぎた可能性があります。＞

この用例の「問題」の正解は “problematic” で、増補前の分類器は “task” の判定だったが、増補後の分類器は “problematic” の判定を行った。

展望として、作成した「問題」の語義のアノテーションルールを拡張し、「問題」の類義語の語義タグ付きコーパスの作成を目指す。さらに、作成したコーパスから、「問題」や「問題」の類義語を特徴語として、「問題内容」及び「解決法」の情報抽出を行うことで、先行研究の日本語版「問題内容」と「解決法」コーパスの作成を目指す。

7　おわりに

本研究は、日本語の科学論文から問題解決プロセスの情報抽出を行うための準備として、人手によるアノテーションによって、「問題」の語義タグ付きコーパスを作成し、自己学習によるコーパスの増補を行った。さらに研究を進め、最終的には日本語版「問題内容」と「解決法」コーパスの作成を目指していきたい。

謝辞

本研究の先行研究の著者である Kevin Heffernan さんと Simone Teufel さんには、論文内で使用した詳細なコーパスアノテーションルールについてご教授いただける等研究に多大なるご助言をいただきました。ここに感謝の意を示します。本研究は、茨城大学の特色研究加速イニシアティブ個人研究支援型「自然言語処理、データマイニングに関する研究」に対する研究支援　および JSPS 科研費 17KK0002 の助成を受けたものです。

参考文献

[1] Kevin Heffernan and Simone Teufel. Identifying problems and solutions in scientific text. *Scientometrics*, Vol. 116, No. 2, pp. 1367–1382, 2018.

[2] Kiyoaki Shirai. Senseval-2 japanese dictionary task. In *Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pp. 33–36, 2001.

[3] Manabu Okumura, Kiyoaki Shirai, Kanako Komiya, and Hikaru Yokono. On semeval-2010 japanese wsd task. *Information and Media Technologies*, Vol. 6, No. 3, pp. 730–744, 2011.

[4] Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako

Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den. Balanced corpus of contemporary written japanese. *Language resources and evaluation*, Vol. 48, No. 2, pp. 345–371, 2014.

[5] Kanako Komiya, Yuto Sasaki, Hajime Morita, Minoru Sasaki, Hiroyuki Shinnou, and Yoshiyuki Kotani. Surrounding word sense model for japanese all-words word sense disambiguation. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, pp. 35–43, 2015.

[6] David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *33rd annual meeting of the association for computational linguistics*, pp. 189–196, 1995.

[7] 鈴木類, 古宮嘉那子, 浅原正幸, 佐々木稔, 新納浩幸. 概念辞書の類義語と分散表現を利用した教師なし all-words wsd. 自然言語処理, Vol. 26, No. 2, pp. 361–379, 2019.

[8] 新納浩幸, 浅原正幸, 古宮嘉那子, 佐々木稔. nwjc2vec: 国語研日本語ウェブコーパスから構築した単語の分散表現データ. 自然言語処理, Vol. 24, No. 5, pp. 705–720, 2017.