専門用語自動獲得研究における評価手法の検討：実験環境評価から実務環境評価へ向けて

宮田玲 [1]　影浦峡 [1]
[1] 東京大学大学院教育学研究科

概要

専門用語を自動的に獲得する研究は 1990 年代から活発に行われ、実験環境でのパフォーマンスは着実に改善している。しかしながら言語実務で用いられている手法は品詞パターンと頻度を中心とする基本的なものにとどまっている。「一から作った専門用語候補リストを人手による正解リストと比較する」という業界標準的な評価の枠組みが、実用に求められる評価と乖離していることが一因と考えられる。この状況をふまえ、本稿では、専門用語自動獲得に関するタスクと既存の評価手法を整理し、現実に求められる評価手法のあり方を検討する。

1　はじめに

執筆、翻訳、文献管理といった言語・文書に関わる実務（以下「言語実務」と呼ぶ）において、専門用語を適切に処理することは極めて重要である。これまで、医療、法律、技術、特許など各分野で専門用語集が構築され、実務でも利用されている。自然言語処理 (NLP) の研究分野では、専門用語の自動獲得タスク[1)]は長年取り組まれ、近年は大規模言語モデル (LLM) を活用した手法上の発展も見られる [1]。しかし、言語実務の場面においては、最先端の用語獲得手法が直接取り入れられることは少なく、品詞パターンや頻度を利用した比較的シンプルな機構に基づく自動化技術を下処理として用いつつ、多くの作業は人手に任されている[2)]。

影浦 [3] は、この課題の背景に、NLP における専門用語処理研究が言語実務の要請を反映していない状況があることを指摘した上で、言語実務から定義した専門用語処理のタスクとして「専門語彙データを更新するフェーズ」と「翻訳において用語を同定するフェーズ」に注目し、そこで求められる要件を試論的に整理した。この論考が提示されてから約 4 年が経ち、関連するタスクへの取り組みは部分的に見られるものの、依然として、言語実務の要請と乖離した状況は続いている。筆者らは、「一から作った専門用語候補リストを人手による正解リストと比較する」という業界標準的な評価枠組みに偏り過ぎている点が背景にあると考える。実務をふまえたタスクを定義することに加えて、各タスクの結果を適切に評価するための手法・データセットのあり方を検討することが求められる。

そこで本稿では、まず先行研究 [3] の議論を出発点として、専門用語獲得に関するタスクを再整理する。続いて、これらのタスクに関連して、既存研究はどのような評価を行ってきたかを概観する。その上で、今後の評価手法のあり方の方針を検討する。なお、翻訳等の言語実務においては、対訳用語獲得も重要なタスクであるが、本稿ではベースとなる単言語での用語獲得の議論を中心とする。

2　専門用語獲得タスクの整理

2.1　専門用語、専門語彙、専門用語集

タスクの整理に先立ち、専門用語をめぐる各概念の理論的位置づけを述べる。専門用語は第一義的に分野の概念を表す単位である。形式的には言語表現として文書中に現れるものの、ある文字列が専門用語であるかの判定は、それがどのような言語的な属性を持つかではなく、それが当該分野に帰属するか、すなわち、当該分野において作られうる専門語彙に帰属するかによる[3)]。したがって、専門用語が

1）NLP 分野では「専門用語抽出」という用語が普及しているが、後述するように用語の抽出だけでなく生成を介する研究も広く含めるために、本稿では「獲得」の語を使うこととする。なお、テキスト中の用語の「同定」タスクも含める。

2）筆者らによる言語実務関係者とのコミュニケーションに基づく。なお、Wissik [2] も専門用語抽出の研究が、EU 等の機関における言語実務であまり活用されていないことをインタビュー結果をもとに明らかにしている。

3）詳しい議論は、文献 [4, 5] を参照されたい。

専門用語であることに関わる根本的な判断は、専門用語の理念的な集合である専門語彙という概念抜きには本来扱えない。ただし、この理念的な専門語彙は実際に観察できるものではないため、実務においては、専門語彙が具現化した個別の専門用語集が（既に存在していれば）起点として参照される。

2.2 中核タスク

以上のように、専門用語という個々の単位を扱う場合でも、理念的・現実的な専門用語の集合を扱うことは理論上の要請であり、専門用語獲得タスクを定義する上で考慮する必要がある。本稿では、先行研究 [3] をベースに、言語実務で想定される中核タスクとして、大きく以下のタスク A, B を扱う。いずれのタスクにおいても、所与の用語集の使用を前提とするか否かでタスクを細分化できる。

タスク A　文書群を情報源として用語集を構築する
- **A1**　用語集を一から構築する
- **A2**　既存の用語集を更新・拡張する

タスク B　個別の文書を対象に用語を同定する
- **B1**　用語集を使わずに用語を同定する
- **B2**　用語集と照合しながら用語を同定する

それぞれのタスクの特徴や相違点を述べる。まず、タスク A の成果物は用語集であり、獲得される用語は一定の利用範囲は想定しつつも将来にわたって様々な用途・文書で参照されうる「タイプ」として蓄積される。一方、タスク B は特定文書に対する執筆や翻訳といった個々の実践に関わるものであり、各用語は当該文書における固有の文脈で言及された「トークン」として処理される。また原則として、タスク A では 1 つの専門分野を対象にその分野の用語が同定される一方、タスク B では複数の専門分野の用語がどの分野の用語か分かる形で同定される[4)]。タスク A と B にはこのような大きな違いはあるが、タスク A を遂行するために文書を参照する作業はタスク B と一部重なる手続きとなる[5)]など、作業工程上で両者は関係している。

明示的に既存の用語集を使うタスク A2 では、新しい用語の獲得と用語集への追加だけでなく、用語集内の廃れた用語の除外や既存用語の修正といったタスクも含まれうる。また、タスク B2 は、用語集に存在しない用語の同定、すなわちタスク B1 も含むことが想定される。

なお、タスク A, B のいずれにおいても、用途に応じて、専門用語だけでなく固有表現が扱われることがある。言語実務においても、「同一の表記を一貫して用いる」という点で両者は似ているため、しばしばまとめて扱われるためであるが、専門用語を規定する属性が分野の語彙に帰属するという点にあるのに対して、固有表現はその表すものの性質によるという点で理論的な位置付けは異なる。

以上のように整理したとき、これまで NLP 分野で取り組まれてきたタスクは、A1 が中心であるといえる。正確には、研究において自動で獲得した用語の集合は、そのままの形で分野の専門用語集として認められることはないため、「専門用語候補リスト」の作成までが扱われる。タスク A2 は言語実務においてしばしば要請されるものであるが [3]、研究としてはほとんど取り組まれていない。例外的な試みとしては、既存の用語集を元に「ありうる」用語候補を生成し、現実の文書群によりそれらの存在可能性の検証することで、用語を獲得する枠組みの研究がある [6, 7]。また、一定規模の用語集をシードとして、その用語集を漸進的に拡張する手法を提案する研究もある [8]。タスク B1 に関しては、近年関連した取り組みが見られるが [9, 10]、コーパスレベルの手法を個々の文書に適用した段階に留まり、タスクとして独立して扱われていない。とはいえ、系列ラベリングに基づく用語獲得手法 [1, 11, 12] は、このタスクに比較的容易に適用できるだろう。タスク B2 に関しては、明示的に取り組んだ研究は見当たらないが、関連タスクが存在する（2.3 節で述べる）。

2.3 関連・派生タスク

専門用語獲得そのものではないが、タスク A, B に共通する重要なタスクとして、用語のバリエーションの同定・まとめ上げ [13] がある[6)]。特に、タスク B2 においては、バリエーションを考慮した用語集との照合は中心的なタスクとなる[7)]。同一概念を指す異なる表現を適切に同定・処理することは、執筆、翻訳、索引語付与などの各種言語実務において重要であるだけでなく、専門用語が分野の「概念」を指すことに対応した理論的にも重要なタスクである。

4)　例えば、「映画産業における生成 AI 活用の法的問題点」に関する文書を翻訳する場合、翻訳者は、自身の専門分野によらず、そこに登場する人工知能分野や映画分野、法律分野の各用語を網羅的に認識し、適切に処理する必要がある。

5)　一方、タスク B を遂行する上でも、一度タスク A を経由して仮の用語集を作成する場合もある。

6)　規範的用語処理の観点からは、集約したバリエーションに対して承認形・非承認形を定義するタスクもある [14, 15]。

7)　このタスクは用語正規化の研究 [16] と密接に関連する。

この他にも、派生タスクとして、獲得した用語に対して、分野の下位カテゴリ、専門度合い、難易度、用語間関係（上位語、下位語、対義語等）、定義文など各種の属性・追加情報を付与するタスクもあり、NLP の各分野で取り組まれている。

3　評価手法に関する既存研究の概観

NLP のシステムやタスクの評価は、大きく外的 (extrinsic) 評価と内的 (intrinsic) 評価に分かれる。

用語獲得の研究では、外的評価は、構築した用語候補リストや用語同定の結果が、下流のタスクにおいてどれだけ有用であるかを評価するものである。これまで機械翻訳 [17] や翻訳品質推定 [18]、通訳の準備 [19] といったタスクで評価がなされてきたが、研究の蓄積は多くない[8)]。

内的評価は、対象となるタスクの結果それ自体の品質（妥当性）を評価するものである。まず、前節で整理したタスク A, B ともに、獲得した個々の用語の妥当性は、分野の概念を表しているかという用語性の点で評価される。これを前提としてタスク A では、成果物である用語候補リストが専門用語集として妥当か判定されうる。その基準は体系化されていないが、例えば、以下の観点がありえる。

- 分野全体の専門語彙に対する網羅性 [15]
- 用語集の目的に応じた適合性 [20]
- 用語集内の用語間関係の体系性 [7]
- （特にタスク A2 では）元の用語集との一貫性

一方、タスク B の場合は、文書中の用語を網羅的に同定できているかも問われる。

タスク A1 に関連した研究では、獲得した用語候補を対象に、適宜上位 K 語に限定しつつ、専門家が個々の用語の妥当性（すなわち適合率）を判定する評価がなされてきたが [21]、試行毎に人手評価のコストがかかることや、全体集合が定められず網羅性の評価が原理的に難しいという課題があった。それに対して、あらかじめ対象コーパスから人手で構築した用語集を正解データ (gold standard) として適合率・再現率・F 値等の指標を測定する方法が現在主流である [1, 22]。生命科学、計算言語学、風力エネルギーなど様々な分野で、比較的規模の大きいベンチマークデータセットが構築され [23, 24, 25]、広く研究に使われている。このような評価手法は、対象コーパスにおける再現率（すなわち網羅性）を測定できることや、人手評価を介さず繰り返し適用でき開発・評価のサイクルを回しやすいという利点があるが、対象コーパスに限定された網羅性であることや用語集としての体系性、目的適合性、一貫性といった観点を測定するものではない。NTCIR の TMREC Task では、提出された複数の結果を正解データとして評価することに加えて相互に比較し特徴づける試みがなされているが、定量的な評価にはなっていない [26]。

タスク A2 に関連した研究 [6, 7, 8] では、最終的に獲得した用語の全部または一部を対象に、用語として妥当か否かを人手により評価している。Sato ら [6] および Iwai ら [7] は、用語候補生成の段階で用語集の体系性を考慮しているが、最終的に得られた結果に対して、改めて体系性や一貫性を直接評価しているわけではない。Jerdhaf ら [8] は、獲得した用語の大部分がシードの用語集と重複するようになったタイミングで、用語集拡張の繰り返し処理を止める方針を示しているが、その妥当性は実際のデータで検証されていない。

タスク B1 に関連して、個々の文書を単位とした研究 [9, 10] は、原則として A1 のベンチマーク評価を文書単位で適用しており、用語をトークンではなくタイプとして扱い、適合率・再現率を計測している。これに対して、系列ラベリング手法で文書中の用語を同定する研究 [11, 12] では、用語のトークンを対象に、コーパス中に付与された BIO タグに類する正解ラベルを基準として適合率・再現率が計測されている。これらの研究では、文書中に用語箇所がアノテーションされたデータセット（GENIA コーパス [23]、ACL RD-TEC 2.0 コーパス [24]、ACTER コーパス [25] など）を使用して、系列ラベリング問題を解いている。しかし、これらの既存コーパスは原則として特定の分野に限定されている。分野外 (out-of-domain) の用語もアノテーションする試み [27] はあるが、言語実務を想定すると、どの分野の用語であるかの情報も必要である。

タスク B2 を直接解く試みは見当たらないが、評価手法としては、タスク B1 と基本的には共通である。ただし、同定した用語を用語集の特定の用語と照合させる工程があるため、用語バリエーション同定 [13] や用語正規化 [16] の研究で用いられる評価手法が関係する。

8）この要因としては、生の自動獲得結果が外的評価に耐える水準にないことや、そもそも下流タスクを考慮した自動獲得がなされていないことが挙げられる。

4　実務環境での評価に向けた検討

実務環境の要請をふまえ [2, 14, 28]、実験環境とのギャップを埋めるための評価の枠組みと方向性を以下で提案する。

用語集としての妥当性の評価　特にタスク A を考えた場合、抽出された用語の集合を用語集としてみたときの、全体としての妥当性を評価する必要がある。3 節で述べたような用語集としての妥当性を評価する手法とデータセットの構築が求められる。

実験環境と異なり、実務環境では用語集の真の全体集合は入手不可能であり、網羅性は測定できない。そこで、既存の専門用語集のデータから潜在的な専門語彙サイズを推定し、それとの比較により網羅性を推定する手法 [15] が使える可能性がある。

用語集の目的適合性を評価するには、目的ごとに用語認定ガイドラインを作り、同一分野で複数の異なる用語集を、参照用に作成する方法がありえる。既存のベンチマークデータセットをベースに使う場合は、（もし付与されていれば）用語の専門性の度合いや、専門用語と固有表現の区別の情報を活用することも有効だろう。

用語集の体系性や一貫性は用語集全体に対しても、部分的な集合に対応しても評価される。前者としては、例えば極めて特定的な用語が非体系的に存在しているといった評価である。後者は特にタスク A2 に関わり、元の用語集でこの用語があるのだから、拡張する際にこの用語を入れるのが妥当、といった局所的な体系性や一貫性に関する評価である。タスク A2 を繰り返し可能な形で評価するには、既存の用語集を「拡張済の用語集」とみなした上で、時系列情報を考慮して用語集を意図的に縮小させたものをベースのデータとして、そこから外れた用語を獲得すべき評価データとする方法がありえる。

多分野の用語の網羅的なアノテーション　タスク B の研究においては、文書中に現れる多分野の用語を、その分野情報含めて網羅的にアノテーションしたデータセットが有用である。既存のアノテーション済コーパスを拡張する形でも構築可能であろう。さらに、タスク B2 を想定すると、特定の用語集を用意し、そのエントリーと文書中の用語を事前に対応付けておくことで正解データを作ることができる。このとき、用語のバリエーションも考慮して、柔軟にエントリーとマッチさせておくことで評価手法の幅が広がる。

有用性に関わる要素の評価　自動獲得の結果を人間の作業の補助に使うことを想定するならば、実務における有用性に関わる要素の評価も重要となる。例えば、タスク A では、最もスコアの高い用語候補だけでなく、他の代替候補も評価対象となる。また、単にスコアだけでなく、その用語を用語集に含める上での判断材料となる情報（用語獲得の情報源である文書の規範性や作成日時等）も評価されうる。タスク B では、用語同定の一貫性（後から一括で修正できるか等）も有用性に関わるだろう。

現実の言語実務タスクでの評価　最後に最も実務に近い形での実用評価について述べる。タスク A に関しては、一から、もしくは既存の用語集を起点に実務で使えるレベルの用語集を作る作業を通じて、自動用語獲得手法をどのプロセスで取り入れたか、獲得結果はどの程度採用されたか、どのような課題があったかを検証・報告する方法である。タスク B に関しては、3 節の序盤に述べたような外的評価の延長で、現実の執筆や翻訳といった作業工程における用語同定の有用性を実務環境で評価する方法である。これらの個々の評価実践は再現可能なものではないが、これらの実践の記録が蓄積・共有されることを通じて、言語実務において用語獲得手法に求められる要件が徐々に明らかになるだろう。そして、このような形で構築された用語集や研究で用いられた文書群は、次なる研究資源として活用できる。

5　おわりに

専門用語は特定の分野（の専門語彙）に帰属する、専門用語は具体的な専門用語集に帰属する、専門用語は概念を表す、という専門用語の存立条件に立ち戻ると、専門用語自動獲得の研究と言語実務をつなぐ道筋が見えてくる。研究で用いられる既存のベンチマーク手法は、人手で認定した専門用語データを使うことよりこれらの存立条件を一旦確保した上で、実験環境でその再現度合いを評価するものである。しかし、言語実務においては、文書中で目の前に現れた文字列をその時点において専門用語として認定すべきか、という形で専門用語の存立条件が常に問われる。これは原理的には再現不可能な営みではあるが、それでも、専門用語の存立条件を部分的にでも問えるような評価のあり方を追求することは、専門用語自動獲得技術の実用化に向けた大きな一歩となるだろう。そしてこれは、専門用語の専門用語性を扱うために必要な理論的な要請でもある。

謝辞

本研究は JSPS 科研費 JP24H00736 および JP23K28378 の助成を受けたものです。本稿の内容を検討する上で、有益なコメント・示唆をくださった Amir Hazem 氏、矢田竣太郎氏、藤井俊英氏に感謝いたします。

参考文献

[1] Kang Xu, Yifan Feng, Qiandi Li, Zhenjiang Dong, and Jianxiang Wei. Survey on terminology extraction from texts. **Journal of Big Data**, Vol. 12, Article 29, 2025.

[2] Tanja Wissik. Impact of automatic term extraction on terminology work: A qualitative interview study in institutional settings. **Terminology**, Vol. 31, No. 1, pp. 110–135, 2025.

[3] 影浦峡. 言語実務における専門用語の扱いと NLP における専門用語処理. 言語処理学会第 28 回年次大会発表論文集, pp. 1907–1911, 2022.

[4] Kyo Kageura. **The Quantitative Analysis of the Dynamics and Structure of Terminologies**. John Benjamins, 2012.

[5] 影浦峡. 言語と文書の間 : メタ科学としての図書館情報学と専門用語研究. 生涯学習基盤経営研究, Vol. 49, pp. 38–48, 2025.

[6] Kaoru Sato, Koichi Takeuchi, and Kyo Kageura. Terminology-driven augmentation of bilingual terminologies. In **Proceedings of MT Summit XIV**, pp. 3–10, 2013.

[7] Miki Iwai, Koichi Takeuchi, Kazuya Ishibashi, and Kyo Kageura. A method of augmenting bilingual terminology by taking advantage of the conceptual systematicity of terminologies. In **Proceedings of the 5th International Workshop on Computational Terminology**, pp. 30–40, 2016.

[8] Oskar Jerdhaf, Marina Santini, Peter Lundberg, Tomas Bjerner, Yosef Al-Abasse, Arne Jönsson, and Thomas Vakili. Evaluating pre-trained language models for focused terminology extraction from Swedish medical records. In **Proceedings of the Workshop on Terminology in the 21st century: many faces, many places**, pp. 30–32, 2022.

[9] Antonio Šajatović, Maja Buljan, Jan Šnajder, and Bojana Dalbelo Bašić. Evaluating automatic term extraction methods on individual documents. In **Proceedings of the Joint Workshop on Multiword Expressions and WordNet**, pp. 149–154, 2019.

[10] Elena Senger, Yuri Campbell, Rob van der Goot, and Barbara Plank. Crossing domains without labels: Distant supervision for term extraction. In **Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, Industry Track**, pp. 1366–1378, 2025.

[11] Maren Kucza, Jan Niehues, Thomas Zenkel, Alex Waibel, and Sebastian Stüker. Term extraction via neural sequence labeling a comparative evaluation of strategies using recurrent neural networks. In **Proceedings of the 19th Annual Conference of the International Speech Communication Association**, pp. 2072–2076, 2018.

[12] Hanh Thi Hong Tran, Matej Martinc, Andraz Repar, Nikola Ljubešić, Antoine Doucet, and Senja Pollak. Can cross-domain term extraction benefit from cross-lingual transfer and nested term labeling? **Machine Learning**, Vol. 113, No. 7, pp. 4285–4314, 2024.

[13] Béatrice Daille. **Term Variation in Specialised Corpora: Characterisation, Automatic Discovery and Applications**. John Benjamins, 2017.

[14] Kara Warburton. **The Corporate Terminologist**. John Benjamins, 2021.

[15] Rei Miyata and Kyo Kageura. Building controlled bilingual terminologies for the municipal domain and evaluating them using a coverage estimation approach. **Terminology**, Vol. 24, No. 2, pp. 149–180, 2018.

[16] Yongqi Fan, Kui Xue, Zelin Li, Xiaofan Zhang, and Tong Ruan. An LLM-based framework for biomedical terminology normalization in social media via multi-agent collaboration. In **Proceedings of the 31st International Conference on Computational Linguistics**, pp. 10712–10726, 2025.

[17] Mihael Arcan, Marco Turchi, Sara Topelli, and Paul Buitelaar. Enhancing statistical machine translation with bilingual terminology in a CAT environment. In **Proceedings of the 11th Conference of the Association for Machine Translation in the Americas**, pp. 54–68, 2014.

[18] Yu Yuan, Yuze Gao, Yue Zhang, and Serge Sharoff. Cross-lingual terminology extraction for translation quality estimation. In **Proceedings of the 11th International Conference on Language Resources and Evaluation**, pp. 3774–3780, 2018.

[19] Ran Xu and Serge Sharoff. Evaluating term extraction methods for interpreters. In **Proceedings of the 4th International Workshop on Computational Terminology**, pp. 86–93, 2014.

[20] Gregor Thurmair. Making term extraction tools usable. In **Proceedings of the Joint Conference of the 8th Workshop of the European Association for Machine Translation and the 4th Controlled Language Applications Workshop**, 2003.

[21] Ziqi Zhang, José Iria, Christopher Brewster, and Fabio Ciravegna. A comparative evaluation of term recognition algorithms. In **Proceedings of the 6th Language Resources and Evaluation Conference**, pp. 2108–2113, 2008.

[22] Hanh Thi Hong Tran, Matej Martinc, Jaya Caporusso, Antoine Doucet, and Senja Pollak. The recent advances in automatic term extraction: A survey. **arXiv:2301.06767**, 2023.

[23] Jin-Dong Kim, Tomoko Ohta, Yuka Tateisi, and Jun'ichi Tsujii. GENIA corpus—A semantically annotated corpus for bio-textmining. **Bioinformatics**, Vol. 19, No. Suppl. 1, pp. i180–i182, 2003.

[24] Behrang QasemiZadeh and Anne-Kathrin Schumann. The ACL RD-TEC 2.0: A language resource for evaluating term extraction and entity recognition methods. In **Proceedings of the Tenth International Conference on Language Resources and Evaluation**, pp. 1862–1868, 2016.

[25] Ayla Rigouts Terryn, Veronique Hoste, Patrick Drouin, and Els Lefever. TermEval 2020: Shared task on automatic term extraction using the Annotated Corpora for Term Extraction Research (ACTER) dataset. In **Proceedings of the 6th International Workshop on Computational Terminology**, pp. 85–94, 2020.

[26] Kyo Kageura, Masaharu Yoshioka, Keita Tsuji, Fuyuki Yoshikane, Koichi Takeuchi, and Teruo Koyama. Evaluation of term recognition task. In **Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition**, pp. 417–434, 1999.

[27] Ayla Rigouts Terryn, Véronique Hoste, and Els Lefever. A gold standard for multilingual automatic term extraction from comparable corpora: Term structure and translation equivalents. In **Proceedings of the Eleventh International Conference on Language Resources and Evaluation**, pp. 1803–1808, 2018.

[28] Kyo Kageura and Elizabeth Marshman. Terminology extraction and management. In Minako O'Hagan, editor, **The Routledge Handbook of Translation and Technology**, pp. 61–77. Routledge, 2019.