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

影浦峡
東京大学
kyo@p.u-tokyo.ac.jp

概要

翻訳産業や特許管理、専門文献管理など、単言語・多言語の技術言語実務において、専門用語の扱いは重要な位置を占めている。自然言語処理（NLP）の分野においても、単言語・多言語専門用語の自動抽出、シソーラス構築、定義抽出等、専門用語処理に関する研究は、持続的になされ、発展してきた。ところが、後者の技術は、前者において、あまり利用されていない。本稿では、技術言語実務における専門用語の扱いの基本的な性格を確認し、NLP における専門用語処理が技術言語実務に実際に使われるために考慮すべき課題を整理する。

1 はじめに

翻訳産業や特許管理、専門文献管理など、単言語・多言語の技術言語実務において、専門用語の扱いは重要な位置を占めている。例えば世界知的所有権機関（WIPO）には充実した専門用語管理部門が存在し、知財関係の多言語専門用語データベースを構築・維持・管理している[1)]。大規模な専門用語データベースとしては、米国医学図書館が管理している医学領域の SNOMED[2)] や欧州の IATE[3)] なども広く知られている。

翻訳実務において専門用語の扱いは重要な要素であり（Bowker 2015）、こうした多言語専門用語データが活用される他、翻訳サービス企業（TLP）は、社として及び／あるいはプロジェクトに対応して、専門用語データベースを蓄積し管理していることが多い（Warburton 2021）。翻訳の品質評価基準においても、しばしば専門用語は独立した評価の軸として定義されている（DFKI/QTLaunchPad 2005）。

自然言語処理（NLP）の分野においても、単言語や多言語での専門用語の自動抽出（Kageura & Umino 1996; Haylen & de Hertog 2015）、異形処理（Daille 2017）、シソーラス構築や定義抽出等、専門用語処理に関する研究は持続的になされており、技術的な発達が見られている。実験評価レベルでの専門用語処理のパフォーマンスは一般に向上している。そして、単言語あるいは多言語の用語抽出をはじめ専門用語処理応用に関する少なからぬ論文で、冒頭に、専門用語の急増に伴い人手での専門用語管理は限界に達しているといったことが言われ、言語実務における専門用語の扱いを補佐するあるいは代替するものであることが緩やかに示唆されている。それにも拘わらず、専門用語処理に関する NLP の技術は、言語実務における専門用語の扱いでは、パターンマッチと頻度のような基本的なもの以外、あまり導入されていない。これは、評価も含む現在の 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ことを示唆している。この背景を踏まえ、本稿では、技術言語実務における専門用語の扱いの基本的な性格を確認し、NLP における専門用語処理が技術言語実務に実際に使われるために考慮すべき課題を整理する。

2 専門用語の位置づけ

2.1 理論的位置づけ

「専門用語」という概念は、分野の専門概念を表現するという観点から機能的に析出されたものであり、喩えて言うと Brassica oleracea や Felis catus のようなカテゴリではなく、「青菜」や「ペット」のようなカテゴリである。個別の専門用語が専門用語として認定される前提条件として、「専門語彙」という概念が要請される。概念的には、専門語彙は専門用語に先行する（Kaguera 2015）。

言語単位としては語・句に相当し、多くの言語で多くが複合語であるが、形と概念の関係及び形の揺れの範囲において、化合物の表記のように完全に人

1） https://www.wipo.int/reference/en/wipopearl。なお、毎年、日中韓を含む世界から大学院レベルのインターンを募集している。
2） https://www.nlm.nih.gov/healthit/snomedct/index.html
3） https://iate.europa.eu/

工的な記号・命名体系と一般語との間に位置づけられる。専門用語としての機能を担うため、表現の同一性が人工的な記号・命名体系と近いかたちで求められる。また、分野の専門概念を表現することが第一義的な存在根拠であることに対応し、いわゆるテキストの文脈に（短期的には）依存しない。専門用語が機能的カテゴリとして認識され専門用語として独立した扱いを受けるのは、専門用語がテキストとは独立に分野に帰属するものとして存在するという認識を前提としている。

以上から、専門用語に関しては専門語彙が理論的に中核的な存在となる。専門語彙は、オントロジーや認識論と関連しつつもそれを構成する用語が言語的な単位としてあるという点でそれらとは独立であり、言語的な単位として現れるけれども概念を表すために存在するという点で言語学が捉える言語的対象ではない（Rey 1995）。

2.2 言語実務における位置づけと要請

言語実務においては、こうした専門語彙・専門用語の理論的位置づけに対応した扱いがなされる。主な点を以下にまとめておこう。具体性を持たせるため、言語実務として、技術文書や専門文書の翻訳を想定する。

1. 専門用語はテキストと独立なので、言語実務においては、分野やテーマに対応して、専門語彙そのものが管理の対象となる。持続的に専門翻訳を行う組織において、多言語専門語彙データの利用・構築・管理は必須であり、前提である。
   (a) 実際的観点からは、その分野のそれなりにまとまった専門語彙データがどの言語においてもまったく存在しないということはほぼないと考えてよい。ただし、ある言語において基本データが存在しないため、別言語での基本データが存在する専門語彙についてその言語で初期データから構築することはありうる。
   (b)「新たな専門用語」というとき、第一義的には、専門語彙データに登録されていない用語を意味する。未登録のものを登録するかどうかは、専門語彙データの管理基準に依存する。語彙データそのものの位置づけと性格、基幹言語におけるその用語のステータス、他言語における対応する用語のステータス等が考慮される。
   (c) 能動的に（翻訳フェーズで同定された未登録用語の報告を受けてではなく）語彙データを拡張するために参照する文書情報源は、それぞれの語彙データの位置づけと性格に応じて決められる。
2. 専門用語はテキストと独立なので、実際に文書を翻訳するフェーズでは、テキストの文脈を踏まえた意味の等価性の維持という翻訳の基本的基準ではなく、起点言語の用語に対応する目標言語の用語を原則としてそのまま用いることが求められる。
   (a) 起点言語文書に現れる専門用語はタイプ・トークンともにすべて同定しなくてはならない。
   (b) 専門用語の「翻訳」は、等価性を保った上で目標言語としてより自然な表現を求め定めるという行為ではなく、対応する用語の一貫した適用としてなされる。同定された専門用語に対して、妥当な専門語彙データに存在する目標言語の用語を利用することが専門用語の「翻訳」作業になる。
   (c) 最初に参照する専門語彙データに登録されていない用語の目標言語側用語を求めるためには、それ以外の、当該分野の語彙データを順に探索し決定する。それらにも用語が登録されていないときには専門文書を探索する。なお、翻訳の際に出現した未登録用語は、語彙データ構築管理部門に報告される。
   (d) 用語は生成されるのではなく選択され一貫して用いられるという前提を踏まえた上で、一般的な言語単位と比べて狭い範囲の特定のバリエーションが許容される（Rogers 1997）。

3 NLP における専門用語処理

ここでは、NLP における専門用語処理の現状をまず簡単に要約し、次に、言語実務の要請から考えられるタスクとそれに求められる要件を考え、NLP における専門用語処理の現状をそれとの対比で確認する。議論を拡散させないために、用語抽出を想定する。

3.1 専門用語処理の現状

NLP における専門用語処理応用として基本的なものは、テキストからの用語抽出である。ほとんどの場合、専門分野のテキスト・コーパス（と一般的な参照コーパス）から、何らかの手がかりで用語候補を抽出するもので、手がかりとしては、コーパス内の頻度、レファレンス・コーパスと分野コーパスの出現の偏り、複合要素の結束性、複合要素の組合せ可能性、出現文脈の偏りなどが用いられており（Kageura & Umino 1996; Heylen & de Hertog 2015）、文脈の偏り評価の流れで、最近では要素のベクトル表現等が活用されている（Terryn 2021）。

評価は、個別に評価用の参照リストを用いる場合と、評価用タグ付きコーパスを利用する場合がある。評価用タグ付きコーパスも複数作られている（Hätty et al. 2017; Terryn et al. 2018; Zadeh & Schumann 2014）。そして、こうした評価用コーパスを利用したシェアドタスクが提供され、異なる手法が評価されることもなされている（Terryn et al. 2020）。評価に当たっては F1-measure 等が用いられることが一般的である。

3.2 言語実務から定義した専門用語処理

NLP としてなされている専門用語処理を離れ、2.2 で述べた、言語実務における専門用語の扱いの位置づけと要請を考慮して、テキストからの用語抽出タスクそしてそれに対する要請を定義してみると、包括的ではないが、以下のようなものが考えられよう。

1. テキストからの用語抽出は、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で、有効でありうる。
2. 専門語彙データの更新を想定した用語抽出を考える場合
   (a) 更新対象となる専門語彙データの存在は前提となり、「新たな用語」の抽出が評価の対象となる場合、語彙データに存在しないことが基準となる。抽出タスクとしても、更新対象となる語彙データを資源として用いることは可能であるだけでなく当然求められることになる。従って、特定の語彙データを想定しその属性を考慮した上での抽出タスクが定義される。
   (b) 抽出の元になる文書データは、更新対象となる語彙データの性質に対応したものを選択する必要がある。「分野コーパス」一般では適切でないことがある。語彙データに対応した分野コーパスの構築タスクが単に外から与えた分野コーパス構築と異なるものとして定義することができる。
   (c) 候補の選択は文書・テキストの出現に基づいてなされるとしても、登録の決定は語彙データの性格との関係でなされる。従って、抽出と選択のフェーズは明確に区別される。抽出の結果は、選択の候補列挙としてなされる必要があり（*n*-best だけでなくその際にどのような情報を提供するかも評価対象となりうるだろう）、選択フェーズのタスクがそれを引き継いで定義される。
   (d) 評価は、更新対象となる専門語彙データの属性の観点から、新規性だけでなく、一貫性や新規追加用語の活用可能性、体系性等を含め、なされる必要がある。
3. 翻訳時の起点言語文書における用語同定を想定した用語抽出を考える場合
   (a) その文書に現れる用語をすべて抽出するタスクとして定義される。従って、F1-score といった指標は不適切であり、漏れは 0 であることを前提とするか、あるいは漏れる要素の一貫性がどこまで明示されるかも評価される必要がある。
   (b) 専門語彙データを想定し、マッチングを含めた抽出として定義することが自然である。
   (c) その際に、専門用語において許容されるバリエーションも含めた名寄せ同定タスクが有効である。
   (d) その文書が扱う主題が属する分野の用語とそれ以外の分野の用語を区別するタスクを考えることができる。

3.3 専門用語処理の状況

3.1 で概観した専門用語処理の現状を、タスク及びその要請として言語実務の観点から定義した専門用語処理の範囲と比較すると、基本的に、オーソドックスな NLP における専門用語の処理は、言語実務で想定されるタスクや要件を考慮していない、それらに対応していないことがわかる。

いくつかのタスクについて例外はあり（Sasaki et

al. 2005; Iwai et al. 2016)、またフィルタリングのタスクなどは注目されてはじめているが、専門用語処理の研究は、それが言語実務を想定しそれへの応用を本気で想定するならば、まだ本格的に始まってもいないと言うことができる。

4 おわりに

本稿では、言語実務の観点から専門用語の扱いの実際とそこにおける要請の主なものを導入し、現在主流の専門用語抽出を概観した上で、それが言語実務の観点から要請される処理のタスク定義や要件とどの程度合致しているかを評価した。一部を除いて、とりわけ標準的なアプローチは、言語実務から要請される要件とかなり乖離していることが確認された。

言語実務の観点から定義された専門用語処理のタスクは、これまでなされてきた用語処理と比べて具体的な文書集合や専門語彙データへの依存性が高い。それゆえ、それらは個別対応であって科学的ではない、ということではない。そもそも、言語学的な抽象の方向で捉えられた言語表現とその処理は、世界に存在する認識上有意味な内容を担った言語表現の蓄積への対応という、言語実務が取り組んでいる問題と、最初から乖離している。そして、理論言語学の展開は、存在との対応における有意味性を担う言語表現一般の性質の特徴付けに関するメタ科学的な検討とは別のものであり、後者が、論理実証主義の失敗移行、総合的蘇生に成功していないことは、特に後者の場当たり性や科学性を意味するものではまったくない。

専門用語は言語の側からは機能的バリエーションとして捉えられ、概念とその体系という自律的な世界を構成するため、命題的内容や知識の構成まで総合的に含めたメタ科学的な言語表現の有意味性の範囲を扱う文書の処理の観点からは、その比較的扱いやすい構成要素として存在する。その点で、言語処理において、何でもいいからあり得た表現が入ってくれば処理するというあり方に、存在の認識として有意味な言語表現を扱うというあり方を付加する問題を検討するためにも有用な応用であろう。

謝辞

本研究は JSPS 科研費補助金基盤 (S) 19H05660 の助成を受けたものです.

参考文献

Bowker, L. 2015. “Terminology and translation,” In Kockaert, H. J. and Steurs, F. (eds.) *Handbook of Terminology*. Vol. 1. Amsterdam: John Benjamins. pp. 304–323.

Daille, B. 2017. *Term Variation in Specialised Corpora: Characterisation, Automatic Discovery and Applications*. Amsterdam: John Benjamins.

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH and QTLaunchPad. 2005. Multidimensional Quality Metrics.

Hätty, A., Tannert, S. and Heid, U. 2017. “Creating a gold standard corpus for terminological annotation from online forum data,” *Language, Ontology, Terminology and Knowledge Structures Workshop 2017*, 8 pp.

Heylen, K. and de Hertog, D. 2015. “Automatic term extraction,” In Kockaert, H. J. and Steurs, F. (eds.) *Handbook of Terminology*. Vol. 1. Amsterdam: John Benjamins. pp. 203–221.

Iwai, M., Takeuchi, K., Ishibashi, K. and Kageura, K. 2016. “A method of augmenting bilingual terminology by taking advantage of the conceptual systematicity of terminologies,” *Computerm 2016*, pp. 30–40.

Kageura, K. 2015. ‘Terminology and lexicography,” In Kockaert, H. J. and Steurs, F. (eds.) *Handbook of Terminology*. Vol. 1. Amsterdam: John Benjamins. pp. 45–59.

Kageura, K. and Umino, B. 1996. “Methods of automatic term recognition: A review," *Terminology* 3(2), pp. 259–289.

Rey, A. 1995. *Essays on Terminology*. Amsterdam: John Benjamins. [Sager, J. C. (trans.)]

Rogers, M. 1997. “Synonymy and equivalence in special-language texts: A case study in German and English texts on genetic engineering,” In Trosborg, A. (ed.) *Text Typology and Translation*. Amsterdam: John Benjamins, pp. 217–245.

Sasaki, Y., Sato, S. and Utsuro, T. 2005. “Automatic compilation of terminological lexicon using the Web,” *Proceedings of the 11th Annual Meeting of the Association for Natural Language Processing*, pp. 895–898. (in Japanese)

Terryn, A. R., Hoste, V. and Lefever, E. 2018. “A gold standard for multilingual automatic term extraction from

comparable corpora: Term structure and translation equivalents," *LREC 2018*, pp. 1803–1808.

Terryn, A. R., Hoste, V., Drouin, P. and Lefever, E. 2020. "TermEval 2020: Shared task on automatic term extraction using the annotated corpora for term extraction research (ACTER) dataset," *Computemm 2020*, pp. 85–94.

Terryn, A. R., Hoste, V. and Lefever, E. 2021. "HAMLET: Hybrid adaptable machine learning approach to extract terminology," *Terminology* 27(2), pp. 254-293.

Warburton, K. 2021. *The Corporate Terminologist*. Amsterdam: John Benjamins.

Zadeh, B. Q. and Schumann, A.-K. 2014. "The ACL RD-TEC: A dataset for benchmarking terminology extraction and classification in computational linguistics," *Computerm 2014*, pp. 1862–1868.