様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２７ 年　 ６ 月　 １ 日現在

機関番号：１２６０１
研究種目：挑戦的萌芽研究
研究期間：2012～2014
課題番号：２４６５０１２２

研究課題名（和文）Ｗｅｂからの能動的候補獲得による専門用語対訳辞書の自動拡張

研究課題名（英文）Augmenting Terminologies through Proactive Extraction of Term Translation Pairs from the Web

研究代表者
影浦　峡（Kageura, Kyo）

東京大学・大学院情報学環・教授

研究者番号：００２１１１５２

交付決定額（研究期間全体）：（直接経費）　　3,000,000円

研究成果の概要（和文）：専門用語において、漢語・外来語の語構成要素はどのような役割を担っているか、用語が拡大していくときに、外来語・漢語の語構成要素はどのように使われまたどのような新要素が出現するか、専門語彙の集合としての一貫性はどのように評価されるかをモデル化するとともに、それを基盤として、与えられた用語集には存在しないけれども存在可能性の高い用語を対訳関係を保持したまま生成し、その用語候補をWebからクロールした分野コーパスを用いて、二言語で、及び各言語ごとに検証することで、最初に与えられた用語集合を拡張する手法を開発した。また、本研究の辞書学上の位置づけも明らかにした。

研究成果の概要（英文）：How native and borrowed constituent elements contribute to the construction of technical terminology, how these elements are used when the terminology glows. By defining terminological network (with terms as vertices and shared constituents as edges) and constituent network (with constituent elements as vertices and co-occurrence in terms as edges), indices to evaluate consistency and coherency of terminology were defined. By using these observations, we developed a method of producing bilingual new term pair candidates from existing terminologies and validating them through monolingual and comparable domain corpora obtained from the web. Experiments have shown that the performance of bilingual term crawling is at least comparable with existing corpus-based extraction method, and complementary in the sense that they extract different types of pairs, which are more relevant to existing terminologies. Theoretical implications of this work was clarified in terms of lexicograpic issues.

研究分野：言語・メディア処理

キーワード：専門語彙　Webクローリング　対訳抽出　語彙成長　語彙ネットワーク

１．研究開始当初の背景

自動専門用語抽出（Automatic Term Extraction）は 1990 年代以降、コーパスベースの自然言語処理の流れとともに活発に研究され、技術的に発展してきた。二言語の対訳用語抽出も、パラレルコーパスからの抽出及びコンパラブルコーパスからの抽出の研究が進み、2014 年に終了した EU の大規模プロジェクト（TTC）で現時点での集大成とも言える技術的な集積がなされている。しかしながら、これらの技術的展開には目を見張るものがあるものの、現実的な応用の観点からは、以下の２点が大きな問題として残されていた。

(1) 一般に自動的な用語抽出では新語あるいは新分野への対応、あるいはこれまで扱われてこなかった言語対への対応が課題として掲げられるが、それらは実のところ sparseness の問題から、先端の技術でも十分にカバーされていない。対応して、実験室的な評価では、評価用の用語集合を用いるが、現実の観点からは、そもそも用語集合があるなら抽出する必要がない。

(2) 用語集合は大きければ大きいほどよい、というのは実際には利用の実情にあっていない。ある見出し語集合は一定の一貫性を持っており、そこから逸脱した用語が含まれていることは、逆に包括性及び／あるいは信頼性に対して不信感を生み、現実的な利用を阻害する。例えば、apple, orange があれば banana も見出し語集合に含まれていた方がよいが、apple, orange, pawpaw, star fruits があって banana がない、という見出し語集合はあまり良いものとは言えない．そして、人間の利用においては、こうした一貫した見出し語の集合を有する語彙資源が依然として極めて重要な役割を果たしている。これに対してコーパスに基づく専門用語抽出は解を与えていない。

これらの問題はまた、コーパスから語彙を抽出するという作業が、テキスト空間を特徴付ける属性としての用語（キーワード）ではなく、一貫した語彙空間を構成する見出し語を抽出しその語彙集合を構成して行くというレベルを扱えていないことを意味する。

以上の背景から、我々は、問題をコーパスからの自動用語抽出ではなく、既往の専門用語辞書の拡張として捉え、対訳専門用語集の自動拡張手法の開発を行った。言語対としては日本語と英語の対を対象とした。

２．研究の目的

(1) 理論的な目標は、与えられた専門語彙集合に基づき、(a) そこから語彙の成長を予測する枠組みを導入し実際に専門語彙の成長をモデル化すること、(b) 語彙集合の内的構造や一貫性を評価する視点と指標を導入すること、(c) (a)と(b)を統合し、語彙集合の一貫性を保ちながら、潜在的に存在しうる用語を導入することで語彙を拡張するモデルを構築すること、である。

(2) 応用面の目的は、既存の対訳専門用語集から、「専門用語対訳候補」を生成し、ウェブ等でそれらの存在を検証することで、対訳対を収集し専門用語集を拡張するシステムを構築することである。

３．研究の方法

以下の分野の用語集を、出発点として存在する対訳専門用語集として用いた：

理論分析：農学、植物学、計算機科学、心理学、物理学、化学

応用評価：それらに加えて経済学、法学

理論研究においては、専門用語の 70～80 パーセントが複合語であることを利用し、語構成要素を基本単位として、その成長と複合語における組み合わせをモデル化する。ベキ分布に従うような対象において標本未出現事象を予測する枠組みを語彙成長の予測問題に適用した。このとき、パラメトリックなモデルを用いることも検討したが、基本的に我々の課題では、現在存在する用語よりも少しだけ大きい用語集合の範囲を扱うため、与えられた標本量の 1.5 倍程度までは有効な Good-Toulmin の補外法を利用した。また用語集合の特徴付けと記述には、用語を頂点、共有する語構成要素を辺とする用語グラフ及び語構成要素を頂点、用語中での共起を辺とする語構成要素グラフを定義し、グラフ及び複雑ネットワークに対して用いられている諸指標を介して用語集合の一貫性を測定した。これら両者の組み合わせは、理論的には可能であるが、とりわけ未出現事象をめぐる連接確率の推定が仮にモデルとして妥当だとしても現実的な利用可能性から見たときにそれほど有用でないと判断し応用とつなぐかたちで検討はしなかった。

応用では、基本的に語構成要素の対訳対応を抽出すること、連接を生成すること、ウェブにおいてそれらの存在を検証すること、の課題に対して用いることができる手法は複数存在するので、それらの妥当性を検討することとなる。

４．研究成果

語彙成長に関しては、とりわけ日本語を扱う場合、外来語語構成要素と漢語語構成要素（専門用語では和語はほとんどない）の振舞いの違いが問題となる。これについては、計算機科学・化学・物理学では用語が拡大したときに、異なり数では外来語語構成要素が漢語語構成要素よりも多くなる可能性が明らかになった。一方、心理学、農学、植物学では、ある程度以降の新語構成要素利用率はほぼ外来語と漢語で同様となることが予測される。既往の語構成要素を組み合わせて用語候補を生成する場合には問題とならないが、未知の語構成要素を想定した場合、語種の出

現傾向は重要な手がかりとなる。

見出し語集合における外来語語構成要素と漢語語構成要素の振舞いの違いもまた分野によって特徴があることが明らかとなった。語構成要素グラフに基づき assortative coefficient（同種のもののみが結合する：1，ランダム：0，異種のもののみが結合する-1）を計算すると、物理学：0.134、心理学：0.283、計算機科学：0.285、化学：0.305、農学：0.337、植物学：0.481 であり、物理学では外来語語構成要素と漢語語構成要素はあまり区別されていないことがわかる。興味深いのは計算機科学で、外来語語構成要素の依存度は高いが、物理学程交合は進んでいないことがわかる。これらの詳細については Kageura (2012)を参照。

対訳用語集の拡張に関しては、大きく二つの課題がある。第一は、既存の語構成要素を組み合わせる際にどのように組み合わせ爆発を抑えるか（用語集合は数千とはいえ、組み合わせを考えると計算論的にではなく用語の規模として非現実的になる）、第二は未出現の語構成要素をどう考えるか、である。

既存の語構成要素を組み合わせるときに、可能な連接の組み合わせをどのように絞り込むかについては、2 項係り関係に基づいて 2 部グラフを作成し、欠落している辺をつないで候補を生成する前に組み合わせの爆発を押さえるために 2 部グラフを分割する方法（Kernighan-Lin）を最初の選択肢として検討した。それに基づき候補対訳を生成し、評価実験を行ったところ、数千の見出し語用語集から、潜在的な候補は 100 万前後となるものを、2 万から 3 万の候補語に抑えることができる（Sato, Takeuchi and Kageura 2013 を参照）。それ以外に、用語ネットワークの段階で partitive clustering を適用することが考えられ、現在、その検証を行っているところである。

一方、未出現の語構成要素については、基本的にワイルドカード扱いで処理できるが、この有効性は検証コーパスをどこまで関連するものに絞れるかに大きく依存する。

対訳関係をどのように保証するかについては、(a) 語構成要素の対訳対応を抽出し候補語対訳を生成する、(b) 同一テキストに出現するかどうかに応じて重みを考える、という二段階の方法を採用した。精度の観点からは、(b)を利用することが有効であるが、(a)のみでも部分一致的な用語対訳は生成検証できており、これについて、コーパス内共起に依存せずにどれだけ精度を挙げられるかは現在継続的に検討しているところである。

＊ ＊ ＊

潜在的に無限の言語表現を処理できればよいとする言語処理の研究者にはあまり知られていないが、見出し語集合をどのように確定するかは、辞書学における最大の未解決問題であり、コーパスの頻度などは参考にされてはいるものの、これまでのところ、現実には辞書編集者が経験により決めてきたところが大きい。本研究では、いわゆる自然言語処理応用として見たときに、対訳抽出において既往のコーパスベースの手法以上のパフォーマンスを出していることが一つの成果であるが、それ以上に、問題設定を概念的に変更し、辞書学の未解決課題に対して、専門用語は複合語が多いという特徴を利用したため専門用語辞書にしか今のところ適用は保証されないとはいえ、一定の道筋を示したことが理論的には極めて大きな成果である。本研究における業績の中に、基調講演が多いのも、その点を反映しているものであると言える。技術的な成果については、現在、評価実験を行い論文を執筆中であり、それについてもまとまったかたちで公表する予定である。

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 0 件）

〔学会発表〕（計 5 件）

① Kyo Kageura (2014) “The sphere of terminology: between ontological system and textual corpora,” *Terminology and Knowledge Engineering*. 19-21 June 2014, Berlin, Germany. (基調講演)

②Koichi Takeuchi and Kyo Kageura (2014) “Terminology-driven terminology augmentation,” *The 14th China-Japan Natural Language Processing Joint Research Promotion Conference*. 12-14 October 2014, Chengdu, China. (査読無)

③ Koichi Sato, Koichi Takeuchi and Kyo Kageura (2013) “Terminology-driven augmentation of bilingual terminologies,” *MT Summit XIV*. 2-6 September 2013, Nice, France. (査読付)

④Kyo Kageura (2013) “On some issues of technical terms in translation: focusing on the gap between CL technologies and human translation activity,” *3rd International Conference on Law, Language and Culture*. 31 May-2 June 2013, Hangzhou, China. (基調講演)

⑤Kyo Kageura (2012) “The status of ‘new terms’ from the point of view of language practitioners, and the crawling of new translation pairs from the Web,” *Neology in Specialised Languages: Detection, Implantation and Circulation of New Terms*. 2-3 July 2012, Lyon, France. (基調講演)

〔図書〕（計 4 件）

① Kyo Kageura (2015) “Augmenting terminology by crawling new term translation pairs from textual corpora,” Dury, P. et al. eds. *La Néologie en Langue de Spécialité*. Lyon: CRTT,

pp. 37-50.
② Kyo Kageura (2015) “Terminology and lexicography,” Kockaert, H. J. and Steurs, F. eds. *Handbook of Terminology*. Amsterdam: John Benjamins, pp. 45-59.
③Kyo Kageura and Takeshi Abekawa (2013) “The place of comparable corpora in providing terminological reference information to online translators: a strategic framework,” Shaloff, S., Zweigenbaum, P. and Rapp, R. eds. Building and Using Comparable Corpora. Berlin: Springer, pp. 285-301.
④Kyo Kageura (2012) *The Quantitative Analysi of the Dynamics and Structure of Terminologies*. Amsterdam: John Benjamins. 243pp.

〔産業財産権〕
○出願状況（計 0 件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計 0 件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
取得年月日：
国内外の別：

〔その他〕
ホームページ等

６．研究組織
(1)研究代表者
影浦峡（KAGEURA, Kyo）
東京大学・大学院教育学研究科／大学院情報学環・教授
研究者番号：００２１１１５２

(2)研究分担者
竹内孔一（TAKEUCHI, Koichi）
岡山大学・大学院自然科学研究科・講師
研究者番号：８０３１１１７４

(3)連携研究者
（　　）

研究者番号：