異なる言語間における専門語彙の体系性の対応の分析

岩井　美樹 †　竹内　孔一 ‡　　石橋　和也 ‡‡　　影浦　峡 ††

† †† 東京大学大学院　学際情報学府

‡ ‡‡ 岡山大学大学院　自然科学研究科

†mikii@g.ecc.u-tokyo.ac.jp　‡koichi@cl.cs.okayama-u.ac.jp

‡‡ploi5t2g@s.okayama-u.ac.jp　††kyo@p.u-tokyo.ac.jp

1　背景

専門語彙は体系的であると言われているが、実証的にその体系性を、とりわけ異なる言語間での体系も含めて、実証的に分析した研究は限られている。語彙の体系性とその多言語間対応を高い解像度で把握することは、技術文書の多言語展開や学習、多言語専門用語処理、機械翻訳などに直接・間接に貢献する。

本稿では英語と日本語の対訳専門用語集を用いて、異なる 2 言語間での専門語彙の体系性の対応について調査する。これまでにも異なる言語間における専門語彙の体系性の対応について分析 [2] を行ったが、今回は新たに 2 分野を追加して分析を行う。具体的には、対訳専門用語集に含まれる全ての専門用語を用いて、専門語彙ネットワークを生成する。次に生成した専門語彙ネットワークにコミュニティ検出アルゴリズム (クラスタリングアルゴリズム) を適用し、各専門語彙に存在している関連する概念の集合 (クラスタ) 単位に分割する。分割されたクラスタの形と各クラスタに含まれる用語を観察することで、異なる言語間における体系性の対応度を分析する。

2　専門語彙の体系性

専門用語は独立して存在しているのではなく、専門用語の集合である専門語彙の中で体系的に存在している [3]。専門用語は複合語であることが多く、用語を構成するそれぞれの語がその用語がもつ概念（意味）を構成的に表現しており、共通の語を通して互いに結びつくことで概念の関連性を示す傾向にある。その専門用語の集合である専門語彙は、この特徴により専門語彙に含まれている概念を構成するネットワークを体系的に形作る。例えば図 1 は計算機科学に含まれている専門用語を用いて生成された専門語彙ネットワーク

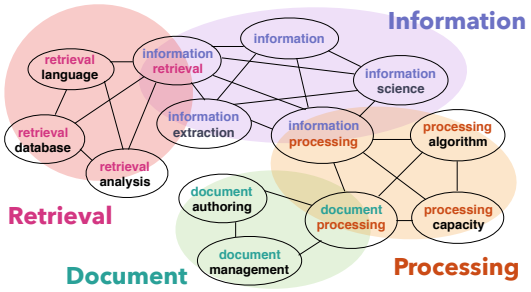

図 1: 計算機科学分野に含まれる専門用語が形作るネットワークの例

を示している。これらの専門語彙ネットワークは次の式で定義される。

$$\begin{aligned} \text{degree}(v_i) &\simeq \sum_j \text{frequency}(c_{ij}) \\ \text{weight}(e_{ik}) &= |\{c_{ij}\} \cap \{c_{kl}\}| \end{aligned}$$

v_i は i 番目の頂点（用語）を示しており、c_{ij} は用語 v_i を構成する j 番目の語を示している。2 つの用語 v_i と v_k について、共通する語により定義される e_{ik} は v_i と v_k 間の辺を示している。関連する概念（意味）をもつ専門用語同士は共通の語を通して結びつくことで、その関連した概念をもつ用語集合を作ることができる。例えば図 1 より、計算機科学分野には “Information”, “Retrieval”, “Document”, “Processing” の 4 つの概念の集合が体系的に存在することがわかる。多くの専門用語は、この既存の概念集合の中に存在する語を用いて新しく作り出される。この概念集合は生成された専門語彙ネットワークに適切なコミュニティ検出アルゴリズムを適用することで、専門語彙に体系的に存在する関連する概念の分析することができる [4]。

専門用語の体系性は専門用語が作り出されるための重要な要素の一つであるにも関わらず、専門語彙の体

表 1: 生成された専門語彙ネットワークの分析

分野	言語の種類	用語数	語構成要素数	機能語	頂点数	辺数	孤立点の数
Com.	英語	16259	5542	636	14186	992425	1102
	日本語	16259	4885	6049	13710	997277	1042
Ecn.	英語	9120	5374	2150	8922	279088	746
	日本語	9120	4684	4442	8672	267344	786
Phy.	英語	11081	5281	642	9966	243968	1066
	日本語	11081	4274	5378	9357	294504	841
Psy.	英語	7026	4183	473	5900	96363	1316
	日本語	7026	3281	3605	5645	125144	752

系性について調査した研究はほとんどない [1, 4]。専門用語は専門語彙に存在する語を用いて作り出されるため、同じ意味を持つ用語でも言語が異なると表現される概念（意味）の構成が異なり、専門語彙の体系性も異なる可能性が考えられる。そこで本稿では日英の対訳専門用語集を用いて、異なる言語の専門語彙の体系性はどの程度対応しているのかを観察する。

3 専門語彙ネットワークの構築

3.1 対訳専門用語集

今回は計算機科学（Com.）[1]、経済学（Ecn.）[2]、物理学（Phy.）[3]、心理学（Psy.）[4]の 4 分野の対訳専門用語集を用いた。

専門語彙ネットワークの生成は英語と日本語それぞれ別々に行った。英語の専門用語を用いて専門語彙ネットワークを生成する場合、次の手順で前処理を行った。

1. スペース区切りで用語を分割し、lemmatiser[5]によりレンマ化（見出し語化）する
2. 用語が表現する概念とは直接関係のない機能語（数字、記号、冠詞など）を除去する

日本語の用語も英語と同様に、次の手順で前処理を行った。

1. 日本語形態素解析システム MeCab[6]により、用語を語構成要素単位に分割する
2. 機能語（数字、記号など）を除去する
3. 助詞、助動詞を除去する

今回 MeCab で用いる辞書は UniDic[7]を用いた。

前処理の後、python igraph library[8]を用いて専門語彙ネットワークを生成した。

3.2 生成されたネットワークの分析

表 1 は生成された専門語彙ネットワークの詳細を示している。表 1 より、まず日本語の専門用語は英語の専門用語よりも機能語が多く含まれていることがわかる。どの分野でも英語の機能語の総数よりも日本語の機能語の総数の方が多い。用語を構成している語（語構成要素）の総数は英日の間であまり差が見られない。孤立点の数も Psy. 以外はほぼ同じような数値である。

生成された専門語彙ネットワークを観察すると、ネットワークは中心に存在する一つの大きな集合（最大サブネットワーク）とその周囲に存在する孤立点を含めた小さな集合から構成されていることがわかった。そのため、多くの用語が密集していると考えられる最大サブネットワークを抽出し、それにクラスタリングアルゴリズムを適用することとした。

4 ネットワークのクラスタ分割

4.1 形式的な分析

抽出した最大サブネットワークにクラスタリングアルゴリズムを適用し、ネットワークをクラスタ単位に分割した。Reichardt and Bornholdt により提案されたアルゴリズム [5] を適用し、分割数は 25 分割と 10 分割で試した。表 2 はその結果を示している。“max” は最大クラスタに含まれる頂点（用語）の数、“min”

[1]相磯秀夫 (1993)「情報処理用語辞典」
[2]金森久雄 他 (1986)「有斐閣経済辞典」
[3]日本物理学会 (1990)「学術用語集: 物理学編」
[4]日本心理学会 (1986)「学術用語集: 心理学編」
[5]http://www.nltk.org/api/nltk.stem.html
[6]http://taku910.github.io/mecab/
[7]http://pj.ninjal.ac.jp/corpus_center/unidic/
[8]http://igraph.org/

表 2: クラスタリングの結果

分野名	言語の種類	分割数	max	min	標準偏差	分割数	max	min	標準偏差
Com.	英語	24	1510	50	341.53	10	2838	480	720.76
	日本語	24	1538	38	338.53	10	2589	552	717.57
Ecn.	英語	25	720	31	159.36	10	1120	360	203.97
	日本語	25	769	5	214.98	10	1585	324	354.71
Phy.	英語	25	806	169	157.74	10	674	218	162.53
	日本語	25	615	55	149.53	10	980	229	210.85
Psy.	英語	25	415	18	85.88	10	1151	349	231.55
	日本語	25	307	12	75.09	10	1599	401	338.68

は最小クラスタに含まれる頂点数（用語数）を示している。

全体的に、英語と日本語という言語の違いはあっても、同じ分野であれば各クラスタの用語数の分布が似ている。25 分割の Ecn. の min, 標準偏差、Phy. の min を除いて max, min, 標準偏差の値が英日間で近い。異なる言語であっても、専門語彙の体系性は分野が同じであればある程度類似している傾向にあるといえる。

4.2 用語レベルでの分析

次に、各クラスタに含まれている用語がどの程度対応していたかについて観察する。ここではもともと対訳関係にあった専門用語同士がどの程度の割合でそれぞれのクラスタに含まれていたかを計算した結果をもとにヒートマップ図を作成し、英日それぞれの分割が用語レベルでどの程度一致していたのかを観察する。

図 2 と図 3 はそれぞれのクラスタに含まれる英日の専門用語の一致率を示している。高い対応度を示すほどセルが濃い青色で表現される。英語のクラスタを基準に計算した対応率と、日本語を基準に計算した対応率をそれぞれ別々に計算したため、2 枚の図で日英間の用語の一致率を図示している。

まず図 2 の 25 分割のときについて考察する。全体的に、英語を基準とした一致率の図では高い対応度を示すクラスタが左下に、日本語を基準とした一致率の図では高い対応度を示すクラスタが右上の方に集中している。特に、Com. と Ecn. ではその傾向がわかりやすい。Phy. と Psy. の中間サイズのクラスタは、どちらの言語を基準にしても高い対応率を示したクラスタが一致していることが多く、ヒートマップ図がよく似ている。また、Com. と Phy. は Ecn., Psy. と比べると比較的高い一致率を示したクラスタが多く見られる。これは、Ecn. や Psy. には「一般的交換手段 (a medium of exchange)」、「一般的価値形態 (general form of value)」のように日本語は同じ「一般的」という語を用いていても英語は異なる語（“medium”, “general”）を用いていることから、関連する概念集合が異なった可能性が考えられる。

次に図 3 の 10 分割の結果を見ると、25 分割のときと同様、Com. や Phy. の方が Ecn. や Psy. に比べると高い対応率だったクラスタが多い。特に、Com. は他の 3 分野と比較すると、高い対応率を示したクラスタが多い。また Ecn. は 25 分割のときと同様、英語を基準にしたときは左下に比較的高い対応率を示すクラスタが、日本語を基準にしたときは右上に比較的高い対応率を示すクラスタが集まっている。Phy. と Psy. は 10 分割のときも英語を基準にしたとき、日本語を基準にしたときのどちらのヒートマップ図も類似している。

5 結論と今後の課題

本稿では英語と日本語の対訳専門用語集を用いて、異なる言語間における専門語彙の体系性の対応度について調査した。頂点を用語とし、各専門用語に含まれる共通の語を結ぶことで生成した専門語彙ネットワークにクラスタリングアルゴリズムを適用し、クラスタ単位に分割した結果を用いて、各専門語彙の体系性について観察し、各クラスタに含まれている用語の数や対訳関係にある用語を含む数を観察することで、異なる言語間における体系性の対応度について調査した。

今後の課題として次の 2 点が挙げられる。

英語と日本語の専門用語の語の単位を揃える

今回は英語はスペースによる分割、日本語は UniDic を用いた形態素解析を行うことで日英それぞれの用語を構成している語の単位を揃えた。UniDic の結果を見ていると、英語よりも細かく分割されていたり、語によっては機能語に分類される

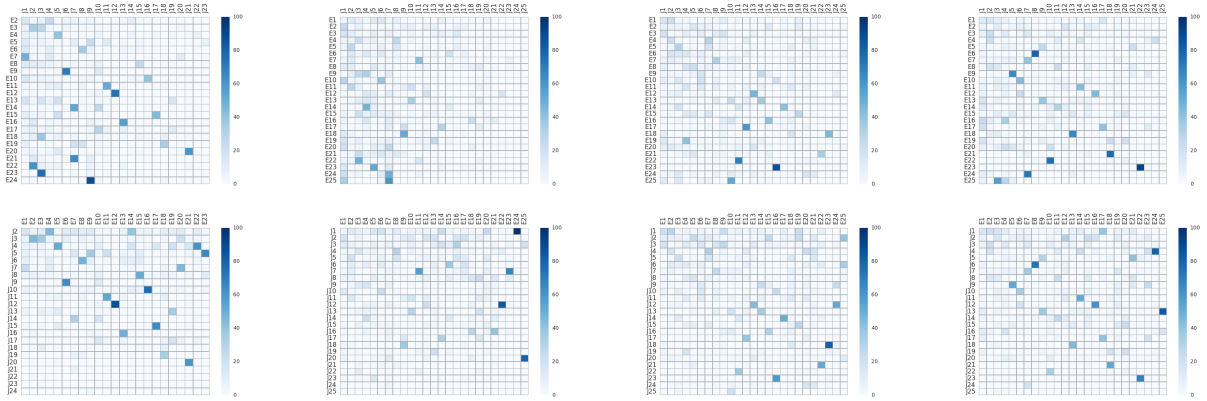

図 2: 25 分割のときの各クラスタごとの日英間の用語の一致率 (上: 英語を基準にしたとき; 下: 日本語を基準にしたとき; 左から: Com., Ecn., Phy., Psy.)

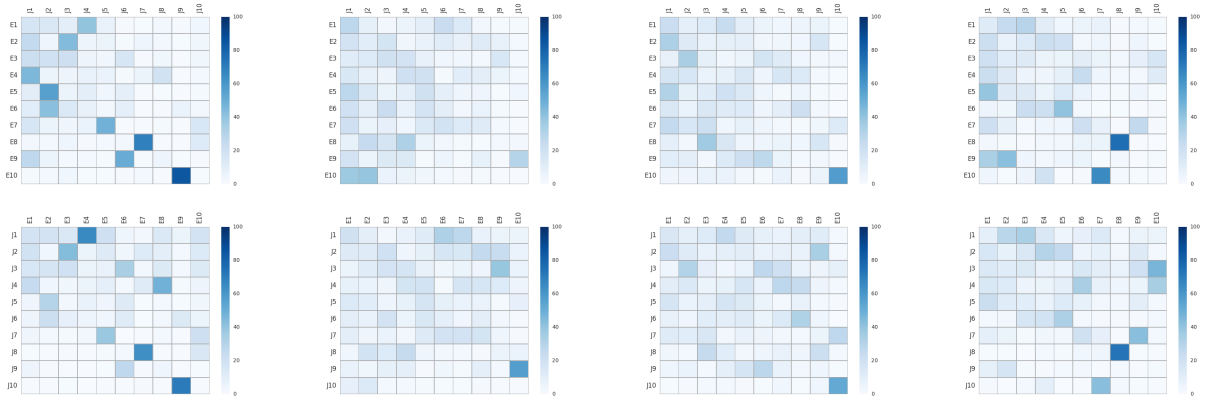

図 3: 10 分割のときの各クラスタごとの日英間の用語の一致率 (上: 英語を基準にしたとき; 下: 日本語を基準にしたとき; 左から: Com., Ecn., Phy., Psy.)

べき語が名詞として処理されてしまっていることがある。どのように英日の異なる言語の語基を揃えるかは今後見直すべき課題である。

機能語の基準を見直す

それぞれの対訳専門用語集に含まれている用語を構成する全ての語の中で最頻出の語を確認したところ、「率（rate)」、「法（method)」のようば内容に直接的な関係のない語が最頻出となっていたことがわかった。従って、どこまでの語を内容に関する概念を含まない機能語と定義するかを再定義する必要がある。

参考文献

[1] Takuma Asaishi and Kyo kageura. Comparative analysis of the motivatedness structure of Japanese and English terminologies. In *Proc. 9th TAI*, pp. 38–44, 2011.

[2] Miki Iwai, Koichi Takeuchi, and Kyo Kageura. Cross-lingual structural correspondence between terminoogies: The case of English and Japanese. In *Proc. 12th TKE*, pp. 14–23, 2016.

[3] Kyo Kageura. *The quantitative analysis of the structure and dynamics of terminologies*. Amsterdam: John Benjamins, 2012.

[4] Kyo Kageura and Takeshi Abekawa. Modelling and exploring the network structure of terminology using the potts spin glass model. In *Proc. 10th PACLING*, pp. 236–245, 2007.

[5] Joerg Reichardt and Stefan Bornholdt. Statistical mechanics of community detection. *Physical Review E*, Vol. 74, No. 1, pp. 16–110, 2006.