Vol.2017-NL-231 No.25
Vol.2017-SLP-116 No.25
2017/5/16

論述上重要な英語表現を専門分野に合わせて自動抽出する手法の試案

江原 遥[1,a)]

概要：英語で科学論文を書くためには、伝えたい内容を、その専門分野で受容される表現を用いて論述する技能が求められる。このような技能の習得には、個々の専門分野に合わせて作成された論述用の英語表現辞書がある事が望ましいと考えられる。 既存には、英語学習上重要な英語表現を集めた辞書については多くの研究があるが、ほとんどの研究が「一般的なコーパス上の単語頻度の高い表現が英語学習全般において重要である」という発想に基づいており、専門分野や論述上重要な英語表現に着目したものではない。 学術上重要な表現を人手で集めて収録した辞書も作成されてはいるものの、規模も小さいうえ、このような辞書を個別の専門分野に合わせて作成・更新し続けることは現実的ではない。 そこで、本稿では、生コーパスから論述上重要な英語表現を専門分野に合わせて自動抽出する手法を試案する。例えば、計算機科学分野の修士学生が英語論文を書くといった場面で、この手法が有用かどうかについて議論する。

1,a)

1. はじめに

国際的な科学技術論文は，現状，英語で論述する事が求められる．英語で科学論文を書くためには，伝えたい内容を，その専門分野で受容される英語表現を用いて論じる技能が求められる．このような技能を英語学習者が効率的に習得するために，「その専門分野の論文を書くために重要な表現」のリストを作り，学習者に提示することを考える．この時，このリストには，どのような表現が入れられるべきだろうか．

まずは，1）専門分野で受容される英語表現，また，そのうち特に重要なものを同定する必要がある．次に，2）それら表現のうち，学習者がまだ覚えていない表現のみをより分けて，まだ覚えていない表現から優先的に学習を進める方が効率的であると考えられる．英語で論文を書こうとする学習者は，ある程度の英語力を既に持っていると考えられるため，学習済みの表現ばかりリストに入っていては，そのリストを活用しようとは思わないだろう．

既存には，こうした，論述のための専門分野ごとの表現を構築する研究が少ない．例えば，英語学習上重要な英語表現を集めた辞書については多くの研究があるが [1], [2], ほとんどの研究が「一般的なコーパス上の単語頻度の高い表現が英語学習全般において重要である」という発想に基づいており，専門分野や論述上重要な英語表現に着目したものではない．学術上重要な表現を人手で集めて収録した辞書も作成されてはいるものの，このような辞書を個別の専門分野に合わせて作成・更新し続けることは現実的ではない．

本稿では，学習者が英語で論述する上で重要な表現を，専門分野に合わせて自動抽出する手法を試作する．まず，「表現」の単位としては，実際には n-gram やパターン，構文木の部分木などが考えられるが，本稿では，ごく単純に「単語」とした．次に，1）の「専門分野で受容される英語表現」については，実際に採択された論文の中で用いられている表現であれば，受容される表現とみなして良いと仮定した．抽出した表現の学習上の重要性については，論文中の頻度と一般コーパス（General corpus）との頻度の尤度比をみて，この値が高い語ーすなわち，通常より高頻度で用いられているとみなせる語ーを，学習上，重要な語とみなした．最後に，2）の覚えていない表現のよりわけについては，抽出した学習上重要な単語に対する学習者の語彙知識（単語を知っているかどうか）をどの程度の精度で識別可能かで評価した．

[1] 産業技術総合研究所 東京都江東区青海 2-4-7
[a)] y-ehara@aist.go.jp

2. 学習上重要な単語の抽出

前述の仮定のもとで，実際に，どのような単語が学習上重要な単語として抽出されてくるのかを示す．まず，対象とする専門分野としては，自然言語処理を選択し，自然言語処理の論文のコーパスとした．具体的には，ACL Anthology [*1] から，Annual Meeting of the Association of Computational Linguistics (ACL) の 2014 年, 2015 年, 2016 年の long paper および short paper を PDF 形式で取得し，pdftotext ツールを使ってテキスト化した．一論文を一文書だと考えて，**gensim** [*2] を用いて，トークン化および Bag of words 化した．このコーパスを，以後，「論文コーパス」と呼ぶ．一般コーパスには，British National Corpus (BNC) [3] を用いた．さらに，人名などが表示される事を防ぐため，12000 語の教育用語彙を集めた語彙集合である SVL12000 に登場する単語のみを実験の対象とした．

論文コーパスの単語の出現確率（単語頻度/述べ語数）と，BNC 中の単語の出現確率の比を r と定義する．この値が高いほど，一般コーパスに比べて論文コーパス中で単語が出現する確率が高い事を意味する．逆に，この値が 1.0 を下回り，低いほど，一般には使われるが，論文コーパスではほとんど使われないことを意味する．**表 1** に，各条件において抽出された 11,645 語のみを対象とする．

$r > 100.0$ においては，“antonym” や “semantic” など，明らかに自然言語処理に関係する単語が上位に来ている事が確認できる．一方, $5.0 < r < 5.5$ においては，“sarcastic” や “intrinsic” などの高難易度語と，“test” や “speech” といった低難易度語が共存ししている事が分かる．実際に，単語の難易度が記載されている weblio 辞書 [*3] では，前者はそれぞれレベル 9,8 であるのに対し，後者は，双方レベル 1 であった．このことから，実際の論文では，難易度が高い単語であっても理解が求められる事がある事が分かる．一方，$r < 0.005$ においては，“door” や “town” など，頻度が高く英語学習の初期に習得されると思われる語が来ているが，これらの語は論文中ではほとんど使われないため，論文を書く上では重要ではないと考えられる．

条件	語
$r > 100$	twitter, neighbor, dropout, neighborhood, antonym, neighboring, compute, neural, sentiment, predicate, corpus, behavioral, dependency, semantic, translation, polyglot, vector, alignment, behavior, extraction
$5.0 < r < 5.5$	uphill, anytime, composition, zodiac, meaningful, combination, count, congressional, distribution, renew, streetcar, sarcastic, character, correctly, efficient, figurative, test, correct, marker, filter, learn, triple, queue, employ, reconstruct, neuron, intrinsic, violation, namely, artificial, toward, vanilla, portion, inclusive, grid, tool, resource, plural, recognition, minimal, den, dictation, speech, derive, median, boundary, relevance, incomplete, train, overestimate, poster
$r < 0.005$	estate, discipline, shit, thick, sit, justice, wonder, door, shortly, deny, oak, quietly, giant, town, guard, deeply, worried, dad, pale, threat, provision, owner, foot, waste, chairman, firm, gone

表 1 各条件において抽出された語

3. 語彙知識の識別精度

次に，語彙知識の識別精度について述べる．評価用データセットについては，15 人について，11,999 語に対し，5 段階で語を知っているかどうかを訪ねた ESL Vocabulary Dataset[*4] を用いた．5 段階のうちの 5，すなわち，最も語を知っている度合いが高いと回答した場合のみを正例，それ以外を負例とみなすことで，2 値判別問題に帰着させた．識別手法としては，[4] と同様，ロジスティック回帰を用いた．素性としては，前述の BNC に対して，トピック数 150 の Latent Dirichlet Allocation[5] を適用した場合の，各トピック所与の単語出現確率の対数値を素性に用いた．

対象とする 11,645 語のうち，前述の方法で，$r > 5.0$ の語を抽出した所，548 語あった．この 548 語を論文を書く上で重要な語と捉え，テストセットとして用いた．比較のため，11,645 語から，この 548 語を除いた 11,097 語から，さらに 548 語をランダムに抽出し，もう一つのテストセットとした．残りの 10,549 語を訓練データ・開発データに用いた．

全 15 学習者に対する語彙知識の識別精度 (accuracy) を計測した．**表 2** に訓練語数に対する識別精度を示す．現実的に短時間（２０分以下）でテストできるのは，数百語程度であるため，まずは，この範囲の精度を計測した．全体として，重要語セットでの精度と，ランダムセットでの精度に，大きな違いは見られなかった．訓練語数を非現実的な $10,000$ まで増やした場合でも，精度に大きな違いは見られなかった．

以上のことから，重要語についても，ランダムセットと同様の精度で語彙知識を識別することが可能であると考えられる．

[*1] http://aclweb.org/anthology/
[*2] https://radimrehurek.com/gensim/
[*3] http://ejje.weblio.jp/
[*4] http://yoehara.com/esl-vocabulary-dataset/

訓練語数	重要語セットでの精度	ランダムセットでの精度
100	0.709	0.724
200	0.736	0.731
250	0.744	0.739
10,000	0.767	0.756

表 2 語彙識別の精度

4. おわりに

本稿では，論文を書く上で重要となる語を抽出し，そのうち，学習者が知らないことが予想され学習上特に重要な語を抽出する手法を試案した．今後の課題は多々ある．例えば，対象を語ではなく句に広げること，「論述上重要」の定義をdiscourseや知識データベースなどの情報も用いてより厳密に定義すること，対象分野を自然言語処理以外に広げること，実際に抽出された重要語のリストを学習者に提供するサービスを立ち上げることが挙げられる．本研究については，今後も，主に `http://vocabularyprediction.com/` にて，資源・サービスなどを提供する予定である．

5. 謝辞

研究はJSPS科研費15K16059の助成を受けた．

参考文献

[1] Nation, I. and Beglar, D.: A vocabulary size test, *The Language Teacher*, Vol. 31, No. 7, pp. 9–13 (2007).
[2] Meara, P. M. and Alcoy, J. C. O.: Words as species: An alternative approach to estimating productive vocabulary size, *Reading in a Foreign Language*, Vol. 22, No. 1, pp. 222–236 (2010).
[3] The BNC Consortium: The British National Corpus, version 3 (BNC XML Edition) (2007). Distributed by Oxford University Computing Services on behalf of the BNC Consortium. URL: `http://www.natcorp.ox.ac.uk/` (Retrieved on October 26, 2012).
[4] Ehara, Y., Sato, I., Oiwa, H. and Nakagawa, H.: Mining words in the minds of second language learners: learner-specific word difficulty, *Proceedings of the 24th International Conference on Computational Linguistics (COLING)* (2012).
[5] Blei, D. M., Ng, A. Y. and Jordan, M. I.: Latent dirichlet allocation, *Journal of machine Learning research*, Vol. 3, No. Jan, pp. 993–1022 (2003).