語学学習 SNS の添削ログからの母語訳付き学習者コーパスの構築に向けて

水本 智也 (奈良先端科学技術大学院大学)†

Toward the Construction of a Learner Corpus with Native Language Translation: Using the Data of Language Learning SNS

Tomoya Mizumoto (Nara Institute of Science and Technology)

要旨

学習者の誤用発生の理由の分析や自動誤り訂正には、学習者コーパスが使用される。学習者の意図を考慮して誤用の理由を分析する、もしくは、学習者の意図を考慮して自動誤り訂正するためには、母語訳のついた学習者コーパスが有効であると考える。しかしながら、母語訳付き学習者コーパスの構築には多大な労力を要する。現在公開されている母語訳付き学習者コーパスには、国立国語研究所によって提供されている「作文対訳 DB」があるが、その作文数は限られている。そこで、本研究では語学学習 SNS に注目する。語学学習 SNS では、学習者の書いた作文とその作文に対して添削が行なわれている。この語学学習 SNS のエッセイ中には、学習者自身が母語によって訳を書いているものも存在する。そこで母語で訳が書かれているものを抽出し、学習者の作文、その添削、母語による対訳が付いたコーパスを自動で構築する。本稿では、語学学習 SNS から作られた Lang-8 Learner Corpora の簡単な概要と現在進行中の母語訳付き学習者コーパスの構築について述べる。

1　はじめに

自分の母語以外を学習する第二言語学習者は増加傾向にある。また、第二言語学習を支援するサービスも増加しており、第二言語学習支援に関する研究も盛んに行なわれている。第二言語学習を支援するサービスとしては、多言語対応日本語読解支援システム「あすなろ」[*1]や「語学学習 SNS Lang-8」[*2]がある。第二言語学習支援に関する研究として最も盛んに行なわれているのは、自動誤り訂正である。英語の文法誤り訂正は、共通のデータセットで訂正性能を競うコンペティションである Shared Task が 2011 年から 4 年連続で行なわれている [5, 4, 12, 11]。

† tomoya-m@is.naist.jp

*1 http://hinoki.ryu.titech.ac.jp/asunaro/main.php?lang=jp

*2 http://lang-8.com

また、中国語のスペルチェックの訂正のコンペティション [15] も行なわれており、自動文法誤り訂正が盛んであることがわかる。

自動誤り訂正や学習者の誤用発生の理由の分析には、学習者コーパスを使用する。学習者の意図を考慮して誤用の理由を分析する、もしくは、学習者の意図を考慮して自動誤り訂正するためには、母語訳のついた学習者コーパスが有効であると考える。学習者コーパスの開発が盛んに行なわれており、母語訳の付いていないコーパスは多く公開されている。一方、母語訳の付いた学習者コーパスの開発はほとんど行なわれていない。その理由の 1 つは母語訳付き学習者コーパスの構築には多大な労力を要するためである。現在公開されている母語訳付き学習者コーパスに、国立国語研究所によって提供されている「作文対訳 DB」があるが、その作文数は限られている。

そこで本研究では、母語訳付き学習者コーパスの構築を行なう。学習者コーパスを開発するにあたり、一から、学習者を募り、実際に作文とその対訳を書いてもらうことは非常に大変な作業である。そこで本研究では、Lang-8 Learner Corpora [9][*3]を用いて、そこから母語訳付き学習者コーパスの構築を試みる。Lang-8 Learner Corpora は、自動誤り訂正 [9, 8, 14]、学習者の書いた作文の母語推定 [1, 2]、問題自動生成 [13] に用いられており、自然言語処理による学習者支援に関する研究で効果が実証されている。これまでの自然言語処理による学習者支援の研究では、学習者が学習言語で書いた文とその添削文のみが用いられてきた。本研究では、学習言語の文とその添削文に加えて、母語訳が付いた 3 つ組で構成される母語訳付き学習者コーパスを自動で構築することを目標とする。

2　関連研究

現在、多くの学習者コーパスが存在している。英語の学習者コーパスは、Cambridge Learner Corpus (CLC) [*4]、NUS Corpus of Leaner English (NUCLE) [3]、Konan-JIEM Corpus (KJ) [10]、International Corpus of English (ICLE) [6]、NICT Japanese Learner English (NICT JLE) [7] など数多くある。誤りの訂正、タイプ付与が行なわれているものはあるが、これらのコーパスには母語訳が付いていない。

日本語のコーパスとしては、寺村誤用データ[*5]、大曽による日本語学習者の作文コーパス[*6]、東京外国語大学の日本語学習者言語コーパス[*7]国立国語研究所の作文対訳 DB[*8]などがある。この中で母語訳が付いているコーパスは作文対訳 DB のみである。しかしながら、その数は 1,754 作文と限られており、さらに添削がついているものはおよそ 250 作文だけである。

語学学習 SNS から作られた大規模な学習者コーパスとして、Lang-8 Learner Corpora がある。自然言語処理による学習者支援の研究で用いられているが、これまで使用されたのは学習者の文とその添削文のみであった。

[*3] http://cl.naist.jp/nldata/lang-8/
[*4] http://ilexir.co.uk/applications/clc-fce-dataset/
[*5] http://teramuradb.ninjal.ac.jp
[*6] http://kaken.nii.ac.jp/d/p/08558020.ja.html
[*7] http://cblle.tufs.ac.jp/llc/ja/index.php?menulang=ja
[*8] http://jpforlife.jp/taiyakudb

表 1　Lang-8 に含まれる学習言語ごとのエッセイ数

学習言語	エッセイ数	学習言語	エッセイ数
English	237,843	French	12,392
Japanese	185,991	German	11,111
Mandarin	28,154	Russian	4,069
Korean	21,779	Traditional Chinese	4,052
Spanish	12,606	Italian	3,339

3　Lang-8 Learner Corpora

Lang-8 Learner Corpora は語学学習 SNS Lang-8 から作られた学習者コーパスであり、現在、奈良先端科学技術大学院大学自然言語処理学研究室 (NAIST) で公開されている。Lang-8 は学習者が学習している言語で作文を書くと、その学習言語を母語とするユーザが添削してくれる。また反対に添削された学習者自身も、自分の母語で書かれた他のユーザの作文を添削できる。Lang-8 では、2011 年 10 月時点で 80 言語をサポートしており、317,307 人のユーザが登録している。

NAIST で公開している Lang-8 Learner Corpora は、2011 年までの作文データが収録されている[*9]。Lang-8 Learner Corpora は、580,549 エッセイからなり、様々な言語から構成されている。表 1 に Lang-8 Learner Corpora のページで挙げられているエッセイ数の多いトップ 10 の言語とそのエッセイ数を示す。1 番エッセイ数が多い言語は英語であり、2 番目が日本語、3 番目が中国語となっている。

現在、公開されている Lang-8 Learner Corpora は、JSON 形式で保存されている。図 1 に Lang-8 Learner Corpora の保存形式の例を示す。破線より上がデータの構造を示しており、破線より下が具体例を示している。保存されている情報は、学習者の作文とその添削に加えて、エッセイ ID、ユーザ ID、学習言語、母語である。本研究で構築する母語訳付き学習者コーパスで必要となる、学習者の文（図中の青字下線部分）、その添削文（図中の赤字破線部分）はこの構造から簡単に抽出することができる。一方、母語訳がどの部分であるかは Lang-8 Learner Corpora の JSON 形式では明示的に示されていない。母語訳が書かれているエッセイもあるが、その場合は学習者の書いた文（図中の青字下線部分）に母語訳が書かれている。そのため、母語訳付き学習者コーパスを作成するためには、学習者の書いた文から学習言語の文と母語訳の文を判別して抽出する必要がある。

4　母語訳付き学習者コーパスの構築

本節では、Lang-8 Learner Corpora から母語訳付き学習者コーパスを構築する方法について述べる。母語訳付き学習者コーパスを構築するための処理は、大きく分けると以下の 2 つに分類される。

[*9] 2012 年以降のデータを使いたい場合は、Lang-8 から買うことで使用可能である

["エッセイID", "ユーザID", "学習言語", "母語",
["学習者文1", "学習者文2",...],
[["学習者文1に対する添削文1", "学習者文1に対する添削文2",...],
["学習者文2に対する添削文1", "学習者文2に対する添削文2",...], ...],]

["772869", "227504", "English", "Spanish",
["My prefer color", "Hello people,", "Today I didn't know to tell us.",
"My prefer color is red.", "Because is funny and diferent."],
[[],[],["Today I didn't know how to say it this:"],["My favourite color is
red."],["Because it is funny and different."]]]]

図 1　Lang-8 Learner Corpora の JSON 形式で保存されている情報の例

表 2　対訳候補として抽出されたエッセイ数。「—」は言語を限定せず、全ての言語を表す。

学習言語	母語	エッセイ数
Japanese	—	28,978
Japanese	English	19,885
Japanese	Mandarin	5,586
English	—	33,533
English	Japanese	28,753
—	—	81,560

1. Lang-8 Learner Corpora から、学習言語と母語訳が含まれているエッセイを対訳候補エッセイとして抽出する
2. (1) で抽出したエッセイから学習者の文と母語訳が対訳になっているものを抽出する

現在、作業が済んでいるのは上記の (1) までであり、(2) は現在も進行中である。そのため本稿では、(1) についてのみ述べる。

Lang-8 Learner Corpora から、学習言語と母語訳が含まれているエッセイを抽出する手順は以下の通りである。

1. JSON 形式のファイルから、各エッセイごとに学習者の文とその添削文を取り出す
2. エッセイから取り出された学習者の文に対して、言語判定を行なう
3. (2) で判定された言語と、各エッセイに含まれている学習言語情報、母語情報を比べて同じであればそれぞれ数を数える
4. (3) で得た学習言語で書かれた文と、母語で書かれた文が一定の割合以上のものを対訳候補エッセイとして抽出する

以下、実際の作業について述べる。(2) の言語判定には、language-detection[*10] ツールを使用した。このツールは 53 言語の判定をすることができる。今回は (4) の学習言語と母語の割合

[*10] https://code.google.com/p/language-detection/

表 3　抽出してきた対訳候補の例（対訳になっている例）

Japanese	いま、だいがっくとてもいそがしです。
English	Right now, School is very busy.
Japanese	たくさんテストーがあります。
English	We have many tests.

表 4　抽出してきた対訳候補の例（対訳になっていない例）

English	I have my final Japanese oral exam in a few days.
English	I hope everything goes well on the exam!
Japanese	十一年間ぐらいバイオリンをひいているから、...
Japanese	そこで、夢をかなえるために来年大学で音楽を ...

が 10:3 以上となっているものを対訳候補エッセイとして抽出した。

表 2 に抽出してきた対訳候補エッセイの数を示す。対訳候補エッセイの総数は、81,560 であった。学習言語が日本語である対訳候補エッセイ数は 28,978 で、学習言語が英語の対訳候補エッセイ数は 33,533 であった。表 1 で示したように日本語で書かれたエッセイは 185,991 であるため、およそ 15.6% のエッセイが対訳候補として抽出されている。同様に英語の方も約 14.0% のエッセイが対訳候補として抽出されている。また、英語が母語で学習言語が日本語であるエッセイは 19,885 であった。

表 3 と表 4 に対訳候補として抽出してきたエッセイの一部を例として示す。表 3 は学習言語（日本語）で書かれた文と母語（英語）で書かれた文が対訳になっているような例である。一方、表 4 は学習言語で書かれた文と母語で書かれた文が対訳になっていない例である。今後は、表 4 のような対訳になっていないエッセイを取り除き、対訳になっているエッセイを取り出し、文同士の対応を自動で取る作業を行なう予定である。

5　おわりに

現在進行中である語学学習 SNS からの母語訳付き学習者コーパス構築について述べた。Lang-8 Learner Corpora の中には、学習者が母語訳を書いているエッセイがある。本稿では、学習者の書いた文に対して言語判定を自動で行ない、学習言語で書かれた文と母語で書かれている文が含まれているエッセイの抽出を行なった。その結果、学習言語が日本語であるエッセイでは、約 15.6% のエッセイが対訳候補エッセイとして抽出された。その中には、対訳となっていないエッセイも含まれているため今後は、そのようなエッセイを取り除いていく予定である。

謝辞

Lang-8 のデータ使用に関して，快諾してくださった喜洋洋さんに感謝いたします．本研究は JSPS 特別研究員奨励費の助成を受けたものです。

参考文献

[1] Brooke, J. and Hirst, G.: Native Language Detection with 'Cheap' Learner Corpora, *Proceedings of LCR 2011* (2011).

[2] Brooke, J. and Hirst, G.: Robust, Lexicalized Native Language Identification, *Proceedings of COLING 2012*, pp. 391–408 (2012).

[3] Dahlmeier, D., Ng, H. T. and Wu, S. M.: Building a Large Annotated Corpus of Learner English: The NUS Corpus of Learner English, *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 22–31 (2013).

[4] Dale, R., Anisimoff, I. and Narroway, G.: HOO 2012: A Report on the Preposition and Determiner Error Correction Shared Task, *Proceedings of BEA*, pp. 54–62 (2012).

[5] Dale, R. and Kilgarriff, A.: Helping Our Own: The HOO 2011 Pilot Shared Task, *Proceedings of ENLG*, pp. 242–249 (2011).

[6] Granger, S., Dagneaux, E., Meunier, F. and Paquot, M.: *International Corpus of Learner English v2*, Presses universitaires de Louvain (2009).

[7] Izumi, E., Uchimoto, K. and Isahara, H.: Error Annotation for Corpus of Japanese Learner English, *Proceedings of LINC-05*, pp. 71–80 (2005).

[8] Mizumoto, T., Hayashibe, Y., Komachi, M., Nagata, M. and Matsumoto, Y.: The Effect of Learner Corpus Size in Grammatical Error Correction of ESL Writings, *Proceedings of COLING*, pp. 863–872 (2012).

[9] Mizumoto, T., Komachi, M., Nagata, M. and Matsumoto, Y.: Mining Revision Log of Language Learning SNS for Automated Japanese Error Correction of Second Language Learners, *Proceedings of IJCNLP*, pp. 147–155 (2011).

[10] Nagata, R., Whittaker, E. and Sheinman, V.: Creating a Manually Error-tagged and Shallow-parsed Learner Corpus, *Proceedings of ACL-HLT*, pp. 1210–1219 (2011).

[11] Ng, H. T., Wu, S. M., Briscoe, T., Hadiwinoto, C., Susanto, R. H. and Bryant, C.: The CoNLL-2014 Shared Task on Grammatical Error Correction, *Proceedings of CoNLL Shared Task*, pp. 1–14 (2014).

[12] Ng, H. T., Wu, S. M., Wu, Y., Hadiwinoto, C. and Tetreault, J.: The CoNLL-2013 Shared Task on Grammatical Error Correction, *Proceedings of CoNLL Shared Task*, pp. 1–12 (2013).

[13] Sakaguchi, K., Arase, Y. and Komachi, M.: Discriminative Approach to Fill-in-the-Blank Quiz Generation for Language Learners, *Proceedings of ACL*, pp. 238–242 (2013).

[14] Sawai, Y., Komachi, M. and Matsumoto, Y.: A Learner Corpus-based Approach to Verb Suggestion for ESL, *Proceedings of ACL*, pp. 708–713 (2013).

[15] Wu, S.-H., Liu, C.-L. and Lee, L.-H.: Chinese Spelling Check Evaluation at SIGHAN Bake-off 2013, *Proceedings of SIGHAN Workshop on Chinese Language Processing*, pp. 35–42 (2013).