Automatically generating questions to support the acquisition of particle verbs: evaluating via crowdsourcing

Maria Chinkina[1], Simón Ruiz[2], and Detmar Meurers[3]

Abstract. We integrate insights from research in Second Language Acquisition (SLA) and Computational Linguistics (CL) to generate text-based questions. We discuss the generation of wh- questions as functionally-driven input enhancement facilitating the acquisition of particle verbs and report the results of two crowdsourcing studies. The first study shows that automatically generated questions are comparable to human-written ones. The second study investigates different types of questions, their perceived quality, and the responses they elicit.

Keywords: automatic question generation, crowdsourcing, particle verbs.

1. Introduction

Questioning is habitually used by language teachers to test comprehension and check understanding of grammar and vocabulary. As argued in Chinkina and Meurers (2017), questions can facilitate the acquisition of different linguistic forms by providing a kind of functionally-driven input enhancement, i.e. by ensuring that the learner notices and processes the form. The CL task of automatic Question Generation (QG) has explored different types of questions: from factual (Heilman, 2011) to deeper ones (Labutov, Basu, & Vanderwende, 2015). For this study, we generate text-based wh- questions and gap sentences targeting particle verbs as they represent a considerable learning load (Schmitt & Redwood, 2011). For instance, given the source text (1), our system generated the question item (1a).

1. University of Tübingen, Tübingen, Germany; maria.chinkina@uni-tuebingen.de
2. University of Tübingen, Tübingen, Germany; simon.ruiz-hernandez@uni-tuebingen.de
3. University of Tübingen, Tübingen, Germany; detmar.meurers@uni-tuebingen.de

How to cite this article: Chinkina, M., Ruiz, S., & Meurers, D. (2017). Automatically generating questions to support the acquisition of particle verbs: evaluating via crowdsourcing. In K. Borthwick, L. Bradley & S. Thouësny (Eds), *CALL in a climate of change: adapting to turbulent global conditions – short papers from EUROCALL 2017* (pp. 73-78). Research-publishing.net. https://doi.org/10.14705/rpnet.2017.eurocall2017.692

(1) Source text[4]: Cancellations "ticked up slightly and unexpectedly" in early April amid press coverage about the coming increases, the Netflix letter said.

(1a) Computer: According to the Netflix letter, what did cancellations do? Cancellations _____ slightly and unexpectedly in early April amid press coverage about the coming increases.

Given a sentence parsed using Stanford CoreNLP (Manning et al., 2014), our algorithm detects particle verbs, identifies syntactic components, and applies transformation rules to generate a question.

The performance of QG systems is commonly assessed by human judges – from university students (Zhang & VanLehn, 2016) to crowd workers (Heilman & Smith, 2010). Using crowdsourcing to compare computer-generated and human-written questions seemed like a logical next step in this line of research. Thus, we conducted two crowdsourcing studies[5] to answer the following research questions:

- Are computer-generated questions perceived as similar to human-written ones in terms of well-formedness and answerability?
- Are wh- questions with a gap sentence perceived better in terms of well-formedness and answerability than open-ended wh- questions?
- Do wh- questions with a gap sentence elicit more particle verbs than open-ended wh- questions?

2. Study 1

2.1. Methodology

The goal of this study was to evaluate our question generation system against the gold standard of human-written questions. Given a corpus of 40 news articles, an English teacher and our system each produced 69 questions targeting particle verbs. Questions (2a) and (2b) below are examples of well-formed human-written and computer-generated questions.

4. http://www.reuters.com/article/us-netflix-results/netflix-customer-growth-slows-amid-price-hike-shares-plunge-idUSKCN0ZY2H4
5. http://crowdflower.com

(2) Source text[6]: Beijing's drive to make the nation a leader in robotics through its "Made in China 2025" initiative launched last year has set off a rush as municipalities up and down the country vie to become China's robotics center.

(2a) Human: What has the "Made in China 2025" initiative done since it was launched last year? It has _____ a rush for municipalities to become China's robotics center.

(2b) Computer: According to the article, what has Beijing's drive done? Beijing's drive has _____ a rush as municipalities up and down the country vie to become China's robotics center.

To acquire high-quality judgements from proficient English speakers, we limited the countries participating in our crowdsourcing study to English-speaking and some European ones (e.g., Sweden, the Netherlands). We also included so-called test questions to ensure the contributors understood the task at hand and were able to tell well-formed from ill-formed questions.

In the study, the participants were presented with a source text one to three sentences long and a question about it. They had to rate each question on two separate five-point scales (well-formedness and answerability). Additionally, the participants were required to answer the question and to make a guess as to whether it was produced by an English teacher or a computer. We collected 1380 judgements from 364 contributors.

2.2. Results

We first calculated the IntraClass Correlation (ICC) between the contributors' ratings. As the ICC was smaller than .1 (.08 for well-formedness and .09 for answerability), we could ignore the dependencies among the observations and use a simple t-test.

The results showed that human-written questions were slightly better-formed than computer-generated ones (Cohen's d=0.13, t=2.06, p=.03). On the answerability scale, the results were non-significant (d=0.02, t=-0.42, p=.1). To quantify the similarity of the two types of questions, we conducted equivalence tests (d=0.5, alpha level of .05). All results were statistically significant on both scales ($p \leq .001$),

6. http://www.reuters.com/article/us-china-debt-robotics-insight/chinas-robotics-rush-shows-how-its-debt-can-get-out-of-control-idUSKCN10E0EV

which indicates that computer-generated and human-written questions are equivalent given the aforementioned parameters[7]. A mixed-effects model revealed a strong correlation between rating a question high and thinking it was human-written (well-formedness: d=0.8, t=17.12, p<0.001; answerability: d=0.7, t=11.71, p<0.001). This indicates that participants expect automatically generated questions to be more ungrammatical and unnatural.

3. Study 2

3.1. Methodology

In the second crowdsourcing study, we wanted to find out i) whether adding a gap sentence to an otherwise open-ended wh- question improves its rating, and ii) whether wh- questions with a gap sentence elicit more particle verbs than open-ended wh- questions. Given the 40 news articles used in the first study, we generated 60 questions and included two types of each question in the dataset – a wh- question with and without a gap sentence. We did not intend to evaluate our system in this study and excluded all ungrammatical or unanswerable questions. In the end, the data consisted of 96 human-written and 96 computer-generated questions.

To imitate a study with non-proficient English learners, we selected contributors with a high reliability but did not limit the participation based on their level of English. The participants were required to answer the questions and rate them on two separate five-point scales (well-formedness and answerability). We collected 960 responses from 477 contributors.

3.2. Results

The agreement among non-proficient English speakers was moderate. The ICC was 0.34 and 0.37 for well-formedness and answerability, respectively, so we opted for mixed-effect models. The results showed that adding a gap sentence improved both well-formedness (d=0.133, t=2.27, p<.01) and answerability (d=0.14, t=2.33, p<.05). To investigate which types of questions elicited more particle verbs, we randomly selected 20% of the responses, excluded nonsensical and non-English answers, and annotated and analysed the remaining questions. We found that

7. In equivalence testing, the null and the alternative hypothesis are reversed. Therefore, statistically significant results indicate that the two samples are equivalent.

the questions containing an additional gap sentence elicited more particle verbs (d=0.16, t=2.97, p<.01) and more correct responses (d=0.12, t=2.5, p=.01).

4. Conclusions

The results of two crowdsourcing studies showed that computer-generated questions are comparable to human-written ones. We also found that the addition of a gap sentence to a wh- question significantly improves its perceived well-formedness and answerability. Moreover, the responses elicited by wh- questions with a gap sentence contain significantly more correct answers, particle verbs among them, than those elicited by open-ended wh- questions.

From the CL perspective, these findings imply that QG systems can benefit from leveraging different types of questions. Combining a wh- question with a more specific gap sentence helps avoid the pitfalls of the two question types: it maximises the grammaticality and minimises the ambiguity of a question while keeping the task communicative. Such combined question items also elicit more target linguistic forms, which is crucial for functionally-driven input enhancement, as discussed in Chinkina and Meurers (2017).

Interestingly, the participants associated the well-formedness and answerability of a question with it being human-written. This shows that people, and teachers in particular, are often not aware of the state-of-the-art in CL technology and could benefit more from intelligent computer-assisted language learning tools.

5. Acknowledgements

This research was supported by the LEAD Graduate School & Research Network [GSC1028], a project of the Excellence Initiative of the German federal and state governments. We would like to thank our LEAD colleagues Michael Grosz and Johann Jacoby for sharing their expertise in statistical analysis.

References

Chinkina, M., & Meurers, D. (2017). Question generation for language learning: from ensuring texts are read to supporting learning. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*.

Heilman, M. (2011). *Automatic factual question generation from text.* Doctoral dissertation, Carnegie Mellon University.

Heilman, M., & Smith, N. A. (2010). Rating computer-generated questions with Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language data with Amazon's Mechanical Turk* (pp. 35-40).

Labutov, I., Basu, S., & Vanderwende, L. (2015). Deep questions without deep understanding. In *Proceedings of ACL* (pp. 889-898). https://doi.org/10.3115/v1/P15-1086

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of ACL* (pp. 55-60). https://doi.org/10.3115/v1/P14-5010

Schmitt, N., & Redwood, S. (2011). Learner knowledge of phrasal verbs: a corpus-informed study. In F. Meunier, S. De Cock, G. Gilquin & M. Paquot (Eds), *A taste for corpora: in honour of Sylviane Granger* (pp. 173-209). John Benjamins Publishing. https://doi.org/10.1075/scl.45.12sch

Zhang, L., & VanLehn, K. (2016). How do machine-generated questions compare to human-generated questions? *Research and Practice in Technology Enhanced Learning, 11*(7), 1-28. https://doi.org/10.1186/s41039-016-0031-7

Research-publishing.net

Published by Research-publishing.net, not-for-profit association
Contact: info@research-publishing.net

CALL in a climate of change: adapting to turbulent global conditions – short papers from EUROCALL 2017
Edited by Kate Borthwick, Linda Bradley, and Sylvie Thouësny

Typeset by Research-publishing.net

Cover design based on © Josef Brett's, Multimedia Developer, Digital Learning, http://www.eurocall2017.uk/, reproduced with kind permissions from the copyright holder.

Cover layout by © Raphaël Savina (raphael@savina.net)
Photo "frog" on cover by © Raphaël Savina (raphael@savina.net)

Fonts used are licensed under a SIL Open Font License

ISBN13: 978-2-490057-04-7 (Ebook, PDF, colour)
ISBN13: 978-2-490057-05-4 (Ebook, EPUB, colour)
ISBN13: 978-2-490057-03-0 (Paperback - Print on demand, black and white)
Print on demand technology is a high-quality, innovative and ecological printing method; with which the book is never 'out of stock' or 'out of print'.

British Library Cataloguing-in-Publication Data.
A cataloguing record for this book is available from the British Library.

Legal deposit: Bibliothèque Nationale de France - Dépôt légal: décembre 2017.