UNIVERSITY OF LEEDS

This is a repository copy of *Towards Deep Learning-Powered Chatbot for Translation Learning*.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/id/eprint/226513/

Version: Accepted Version

White Rose
university consortium
Universities of Leeds, Sheffield & York

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

Towards Deep Learning-Powered Chatbot for Translation Learning

Moneerh Aleedy[1,2(✉)], Eric Atwell[1], and Souham Meshoul[2]

[1] School of Computing, University of Leeds, Leeds LS29JT, UK
scmmale@leeds.ac.uk
[2] Information Technology Department, College of Computer and Information Sciences, Princess Nourah Bint Abdulrahman University, Riyadh, Saudi Arabia

Abstract. As a consequence of the recent advances in artificial intelligence and educational technologies, the education sector is witnessing significant changes and transformations through massive use of intelligent systems with the goal to assist students in their learning experience and teachers in delivering academic knowledge in a better way while reducing burnout and stress. Communication, knowledge acquisition, and learning are now possible anytime and anywhere through modern technologies. However, learning translation is a process that requires continuous effort, enthusiasm, and motivation from both students and instructors. While some universities adopt traditional and outdated approaches to translation instruction, others use more innovative ways. Within this context and in order to foster a student-centered translation learning approach, our work focuses on taking advantage of recent advances in machine learning to develop a chatbot to help language learners, especially translators, develop their skills by having a conversation with a chatbot and getting the appropriate Arabic translations of English sentences under study. A general framework is proposed, and a first prototype is developed using a part of bilingual corpora. Another focus of this study is the preprocessing phase used to create a sentence-based paired dataset to train the machine learning model from bilingual corpora. The preliminary results are promising.

Keywords: Artificial intelligence · Chatbots · Education · Translation learning · Machine learning · Bilingual corpora

1 Introduction

Students pass through several phases during their learning journey, starting from admission, then courses registration, academic advising, learning, exams, feedback, and finally graduation. Artificial intelligence (AI) has shown impressive successes in the field of natural language processing (NLP), and the recent success in machine learning has broadened its use. As a result, AI and NLP researchers have taken upon themselves the task of helping students during their journey. One of the AI technologies that are taking the leads recently is the educational chatbot, which is a conversational agent that responds to student queries based on its design purpose.

AI educational chatbots serve various purposes, such as responding to students' administrative inquiries about courses, objectives, learning outcomes, academic rules, and regulations. Other education chatbots are used to teach courses, conduct exams, and some chatbots just chat with the student to relieve study pressure.

Language and translation learning also benefit from the power of AI technologies by using chatbots to help learners continue conversations in English, Arabic, or many other languages. Also, improve their vocabulary, grammar, and the four language skills: listening, reading, speaking, and writing.

This paper aims to use AI and NLP techniques to create a chatbot to help the translation students and the instructors and supervisors perform a specific translation task by giving alternative Arabic phrases in a particular context for any English sentence provided.

This paper will be structured as follows: Sect. 2 reviews chatbots in language learning, translation learning and discusses challenges in terms of language and translation. Section 3 provided the research methodology used to create a translation chatbot. Finally, in Sect. 4, the conclusion and future work are given.

2   Related Work

2.1   Chatbots in Language Learning

Chatbot technologies have helped language learners practice conversation, improve their vocabulary and grammar for different languages, receive instant feedback about their achievements and motivate them by getting rewards [1, 2].

An excellent example of such a bot is **Vasya**, a Russian mobile application designed to teach English via text, audio, and video instructions. Vasya is called a chatbot, but it is actually an interactive tutorial based on a semi-communicative approach. It requires the learner to read, listen and pronounce words and phrases in English. Vasya has a good sense of humor and easily explains grammar. It's perfect for beginner-level English learners but not suitable for advanced levels. According to its creators, at more advanced levels, learning becomes very complex and requires a human tutor [2].

A virtual English language tutor known as **Andy** English Bot helps users learn and practice English. It allows its users to chat, study new words, learn grammar and play language games in the form of conversations. With Andy, users can practice daily interactions such as greeting and chatting about the weather. It is also possible to learn how to ask and answer questions. However, Andy fails to remember previous conversations and talks too fast, so; students cannot understand and respond [1, 3].

Duolingo is one of the most popular language-learning chatting platforms. It aims to improve general language skills through games and practice speaking with chatbots. Moreover, virtual language tutors help users to learn a new language comfortably without the embarrassment of miscommunicating with native speakers, and it supports more than 23 different languages [1, 2].

GenieTutorPlus is another example of language-teaching chatbots. It automatically detects and corrects grammatical errors in learners' sentences and improves their ability to speak a foreign language by returning the corrected results and giving better expressions [4].

2.2 Translation Learning

Foreign language learners frequently use translation to facilitate language learning and acquire a new language. Although translation has played different roles in language teaching for students from different social backgrounds, most educators agree that translation is a powerful tool to help students understand foreign words and expressions more confidently and express ideas in the target language [5].

The researchers in [5] explore some strategic use of translation in English learning that enhances English skills, including reading, writing, listening, and speaking. The first strategy is learning by linking the new words to their mother-language equivalent; they found this method more effective than learning vocabulary in context. Moreover, another strategy involves using technology aids such as dictionaries, notes, electronic translation machines, etc. Finally, interaction with other students.

The approaches most frequently used in translation courses are: comparing two translations, familiarizing with collocations, expressions, and terms, conducting group discussions, editing the target text, and using translation tools such as dictionaries and general and specialized encyclopedias [6].

Furthermore, many researches on the use of parallel bilingual or multilingual corpora in classrooms have shown potential benefits for translation students. The teachers can use such combinations in the classroom to introduce students to a set of parallel texts that include cases from both the source and target languages. This can help students increase their awareness and enhance their translation skills by studying and reflecting on how professional translators can achieve different types of translation [7]. Moreover, translation teachers can also use a monolingual target language to introduce students to different styles of the target language and to encourage them to explore and observe how the target language is used authentically by native speakers. This action enhances students' knowledge and helps them to become more professional and independent [8]. The use of corpus methods to assess the quality of students' translations can be linked to other digital tools that are useful to both teachers and researchers in the field of translation. Teachers can implement digital language databases (online collection), online dictionaries based on parallel corpora, or machine translation platforms in their courses. The motivational potential of using these tools, on the part of students, is very high, as in all cases when didactic methods intersect with real-life interests (for example, comparing Google Translate with the target set of texts for classes) [9].

In a study on 47 undergraduate EFL students enrolled in a translation course in two Moroccan Universities, the study measured the students' abilities in English-Arabic and Arabic-English translation. The students' ability in English-Arabic translation is greater than their ability in the Arabic-English version. However, the overall scores that students got in English-Arabic are above average and below average in Arabic-English, suggesting that these acceptable values are not at the expected level of ability in both types of translation [5]. This indicates a strong need to enrich this field with research to develop students' translation levels.

2.3 Challenges

Chatbots are developed for many languages such as English, Arabic, French, etc. However, each language has its own sentence structure, punctuation rules, and the use of

spaces, which is a barrier for current chatbots to deal with [10, 11]. For English chat-bots, the authors [10] mentioned several points that affect the efficiency of chatbot conversations; the most important ones include the inability to recognize grammatical errors and similar meanings questions. Moreover, information retrieval from a database is not realistic; two questions may look different in terms of words but have the same meaning, so similarity measures should be used to eliminate differences [10, 12].

On the other hand, the Arabic language chatbots have many linguistic complexities because of the different types of Arabic language: Classical, Modern, and Colloquial. First, classical Arabic, which is used in the Quran, is more complex in grammar and vocabulary. The second type is Modern Arabic, which is considered the official language of the Arab countries and is used in everyday language, education, and media. Finally, in Colloquial Arabic, the third type, the grammar, and vocabulary are less sophisticated. However, different Arab countries have different Colloquial Arabic, and most people use it in their everyday spoken conversations and informally writing [13]. Moreover, Arabic writers make very common mistakes in spelling some problematic letters, such as Alf Hamza and Ta Marbouta. Morphological richness Arabic words are influenced by a large number of features such as gender and numbers. Also, verbs, adjectives, and pronouns are all gender-specific, requiring the chatbot to have two different responses systems for male and female users. As in other languages, Arabic has its own set of unique dialogue expressions, for example, while the English greeting expression "good morning" gets the answer "good morning", the Arabic equivalent greeting " صباح الخير" "Morning of Goodness" gets the answer " صباح النور" "Morning of Light" [14, 15, 16]. Moreover, Arabic is considered a pro-drop language in which the subject of the verb can be identified implicitly in its morphology; the subject is embedded in the verb, unlike in English. For example, the sentence " ذهبت إلى الحديقة" in Arabic can be expressed as "I went to the park". The subject she and the verb went are represented in Arabic by the singular verb " ذهبت" [17]. A collaboration between linguists and computer scientists is required to overcome these challenges.

3 Proposed Framework

The framework we proposed in this paper consists of two main stages offline, and online. In the offline stage, we first understand the main domain of the research. Then we study the data and understand all of its components. After that, we manipulate the data to make it ready to use as conversational data. The last phase in the offline stage is working on the machine learning model, which includes planning, building, and evaluating.

On the other hand, direct communication with the user occurs in the online stage. The user uses the interface to send the question to the chatbot, which is then processed and fed to the trained model to get the appropriate response and send it back to the user. Figure 1 illustrates the general framework.
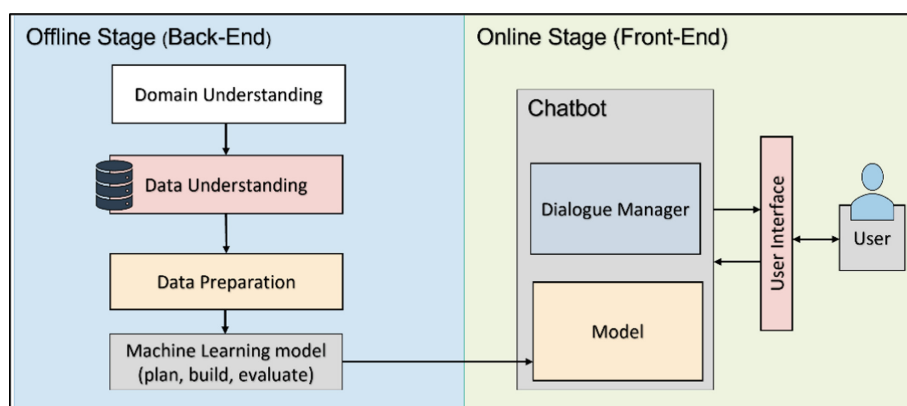
Fig. 1. The general framework

3.1 Domain Understanding

A series of interviews were conducted with a member in the Translation Department at the language college at Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia, to identify the needs in the teaching and learning Translation from English to Arabic languages. Moreover, the meetings illustrated the challenges that faculty members face and discussed the best way to use AI technologies to support translation teaching and learning based on the student, teacher, and supervisors' needs. As a result, the following three main problems are identified:

- Give suggestions for alternative phrases translations.
- Highlights common translation mistakes.
- Assessing the students' translation projects based on students' previous translation scores from Saudi Learner Translator Corpus (SauLTC).

This paper will focus on the first problem: giving suggestions for alternative phrases translations.

3.2 Data Understanding

The proposed corpus is the Saudi Learner Translator Corpus (SauLTC) [18], part of the "Design and Build the Saudi Bilingual Blog for Translation Learners" project funded by PNU. It is an English-Arabic translation corpus composed of three sub-corpora and three participants profiles.

The first corpus is English student source text; the second and third corpora include two versions of the translations to the Arabic language: the draft translation, which is the student's first attempt to translate the source text on her own; and secondly, the post-feedback final translation.

The collection of SauLTC involves three types of participants. The first and primary type of the participant is the student who produced the translation. The student profile contains information related to the student translator, the source, and the translated text.

The second type of participant is the instructor, who provided feedback on the students' drafts and later assessed the students' final submission. The instructor information includes educational background, experience in teaching translation, and experience in supervising translation projects.

The third type of the participant is the alignment verifier who double-checked the automatic sentence alignment.

The current SauLTC corpus has 366 translations for 366 students, and they were under the supervision of 48 instructors and 23 alignment verifiers. The source texts are chapters or booklet extracts from different contexts. In addition, the students are allowed to find their own texts to translate. The only caveat is that they had to obtain the instructor's approval of the book title before beginning implementation. Table 1 shows a list of the contexts available in the database [18].

Table 1. SauLTC contexts

Context	Number of texts in the database under each context
Health	109
Psychology	41
Self-help	66
Business	51
Parenting	12
Language	12
Religion	1
Education	28
Biography	1
Autobiography	20
Nutrition	11
Management	1
Fiction	7
Social sciences	1
Sciences	3
History	2
Total	**366**

3.3 Data Preparation

We received the corpus as an SQL dump file, and then we extracted the data in CSV format to make it easier to study and analyze. We used python to extract the data, and as a result, we got 19 tables. As this research focuses on giving alternative phrases in Arabic, we explore the file table in detail, see Table 2.

Table 2. File table details.

Column name	Description
ID	Row sequence number
Type	Type of translation (SOURCE, DRAFT, FINAL)
GroupNumber	Translation file group number
TransID	Translator ID
Version	Translation version
Year	Translation submission year
Text	The English or Arabic text
Genre	The classification of the text

After extracting the file table, we explore the data and select the "Health" context because it has the largest translated texts. The database contains 109 different translation texts under the Health category. In addition, the following steps are performed using python and jupyter Notebook:

- Select Health context texts.
- Remove the unnecessary columns (ID, groupNumber, version, Year).
- Remove the draft text rows since we need the final revised Arabic translation from the English source text.
- Modifying the data to create English-Arabic pairs.
- Split the source and target texts into sentences based on the ".".
- Remove numbers.
- Replace """ with """ in all texts.
- Expanding English contractions.
- Remove punctuation and New Line(\n).
- Lowercase all English text.
- Manually check and correct the divided sentences.

Table 3 shows a sample of the data after applying preprocessing steps.

Table 3. Sample of the data after applying the preprocessing steps.

	English	Arabic
0	the stress of misunderstanding i have met countless people with autism who misread social situations and behavior and even after someone tries explaining what they had not grasped they still do not understand	الضغوط الناتجة من سوء الفهم لقد التقيت بالعديد من الناس المصابين بالتوحد والذين يفهمون الأوضاع والتصرفات الاجتماعية بشكل خاطئ والذين لن يفهموا مهما حاول الناس الشرح لهم مالم يستطيعوا إدراكه
1	enduring that experience again and again takes its toll	ومرورهم بهذه التجربة مرارا وتكرارا يسبب لهم الحزن
2	knowing i am supposed to understand this but no matter how hard i try i cannot causes frustration unhappiness and anxiety	ومعرفتي بأني مهما بذلت من جهد شديد لأفهم لن أستطيع يسبب لي التعاسة والحزن والقلق
3	many react by shutting down in the face of social encounters or simply avoiding them	ويتعامل العديد من المصابين بالتوحد مع المرض عن طريق الانغلاق عن اللقاءات الاجتماعية أو تجنبها
4	some turn inward and experience depression	والبعض الأخر يعايش الاكتئاب وينغلق على نفسه
5	selfesteem suffers as they ask why do not i understand this what is wrong with me am i stupid social understanding is only one kind of being smart	كما أنهم يعانوا من عدم الثقة بالنفس ويتساءلون لماذا أنا وما لا أفهم هذا ؟ ما العيب الذي أعاني منه؟ هل أنا غبي الذكاء الاجتماعي إلا نوع واحد من أنواع الذكاء
6	you can be brilliant in many other ways and still struggle with grasping facial expressions and other subtle cues in social situations	ومن الممكن أن تكون ذكيا بعدة طرق أخرى ومع ذلك ستعاني من عدم فهم بعض تعابير الوجه والإشارات الدقيقة في المواقف الاجتماعية
7	social understanding requires what howard gardner famous for his theory of multiple intelligences called interpersonal intelligence	ويتطلب الذكاء الاجتماعي ماسماه هاورد قاردنر بالذكاء الشخصي والذي عرف بنظريته عن الذكاءات المتعددة
8	a person with strengths in this area can assess the emotions desires and intentions of others across different social situations	والشخص الذي لديه نقاط قوة في هذا المجال قادر على أن يقيم العواطف ورغبات الآخرين في مختلف المواقف الاجتماعية
9	of course someone who struggles with interpersonal intelligence can demonstrate intelligence in for example music math or solving complex puzzles	وبالطبع الشخص الذي لديه ذكاء شخصي من الممكن أن يظهر لديه عبقرية في الموسيقى والرياضيات أو حل الألغاز الصعبة
10	aware of their difficulty many children apologize for themselves almost habitually—even without understanding what they are apologizing for	و دائما مايعتذر الأطفال المصابين بالتوحد عن أنفسهم وذلك لوعيهم بمشكلتهم – حتى إذا لم يفهموا ما يعتذرون عنه
11	they may understand social rules in extremes of black and white	وقد تبدو القوانين الأجتماعيه لديهم أما بيضاء أو سوداء
12	they are making every effort to get it right and if they suspect they have not said the right thing or acted the right way their instinct is to blurt out sorry sorry no matter how many times parents or teachers reassure them they come to expect that they will make mistakes so they automatically apologize	ويبذل هؤلاء الأطفال ما بجهدهم لجعل الأمور صحيحة حتى إذا شعروا بأنهم لم يتكلمون أو يتصرفون بالطريقة الصحيحة دائما تجبرهم فطرتهم على قول أسف أسف مهما طمئنهم والديهم أو أساتذتهم ودائما مايشعرون بأنهم سيخطئون لذلك يقومون بالاعتذار مسبقا
13	living in a state of constant confusion about even ordinary social interactions can mean that when situations arise that are unanticipated or truly unfamiliar it is likely the child will react in unexpected or extreme ways	في الأحداث ويعيشون في حالة من التشتت حتى الأجتماعية الطبيعية مما يعني أن الطفل عادة عندما يواجه أي حدث غير متوقع أو غريب فمن المتوقع أنه سيتصرف بطريقة غير لائقة أو غير مألوفة

Table 3. (*continued*)

14	to an observer the behavior can look rash sudden or inexplicable but it is often the result of frustration and anxiety that has been building in the child for some time	ومن وجهه نظر الملاحظ أن هذا التصرف قد يبدو سريع ومفاجئ و لا يمكن تفسيره ولكن عادة يكون نتيجة الإحباط والقلق الذي تجمع داخل الطفل منذ وقت

3.4 Modeling

The infrastructure used for this experimentation involved using python programming language and jupyter Notebook. Before doing any complex modeling, we need to transform the dataset into a numerical format suitable for training. We apply the word tokenization and lemmatize, and then we use doc2bow to convert a list of words into the Bag of Words (BOW) format. We split all the text in the dataset into an array of tokens (words), then we built a vocabulary dictionary that contains all words in the dataset and their corresponding index value. The array of words is then converted to an array of indexes.

In this experiment, we created a retrieval-based chatbot called "TranslatorBot" that uses TF-IDF to match user English sentences with corpora's most relevant Arabic sentence. It is built from scratch using NLTK and GENSIM Libraries.

3.5 Preliminary Result

To make a conversation with the TranslatorBot, the student starts typing in the textbox. The chatbot responds with a message asking the student to enter the English text he/she wants to find alternatives in the Arabic language. To end the conversation, the student typed "bye". Figure 2 shows a sample conversation with the chatbot.

Fig. 2. Conversation with the Chatbot

The preliminary results are promising and give us a clear insight into how the conversation will go through. In addition, TranslatorBot will help students find solutions to

their translation problems by quickly reviewing terms and phrases related to the translations they are looking for anytime and anywhere. Moreover, when the chatbot responds with the most relevant Arabic sentences based on the student sentences in English, the students will learn how to translate these sentences, and their translation skills will be improved. The purpose of this chatbot is not to give translations but to show how to translate sentences in different ways.

4   Conclusion and Future Work

This paper highlights two main problems, the first one is creating a chatbot for teaching translation, and the other one is dealing with bilingual corpora. We encourage future researchers to investigate the impact of using deep learning models to solve such problems. Moreover, extracting pair sentences accurately to train the chatbot is very important, as it affects the chatbot performance. For that reason, we also suggest investigating the different methods of creating sentences-based bilingual corpora.

References

1. Na-young, K., Cha, Y.J., Kim, H.-S.: Future English learning: Chatbots and artificial intelligence. Multimedia-Assist. Lang. Learn.**22**, 32–53 (2019)
2. Dokukina, I., Gumanova, J.: The rise of chatbots-new personal assistants in foreign language learning. Procedia Comput. Sci. **169**, 542–546 (2020). https://doi.org/10.1016/j.procs.2020.02.212
3. Smutny, P., Schreiberova, P.: Chatbots for learning: a review of educational chatbots for the Facebook messenger. Comput. Educ. **151**, 103862 (2020). https://doi.org/10.1016/j.compedu.2020.103862
4. Choi, S.-K., Kwon, O.-W., Kim, Y.-K.: Computer-assisted English learning system based on free conversation by topic. In: Borthwick, K., Bradley, L., Thouësny, S. (eds.) CALL in a climate of change: adapting to turbulent global conditions – short papers from EUROCALL 2017, pp. 79–85. Research-publishing.net (2017). https://doi.org/10.14705/rpnet.2017.eurocall2017.693
5. Akki, F., Larouz, M.: A comparative study of English-Arabic-English translation constraints among EFL students. Int. J. Linguist. Transl. Stud. **2**(3), 33–45 (2021). https://doi.org/10.36892/IJLTS.V2I3.163
6. Jafari, O.: How approaches to teaching English can be used for teaching translation. Trans. J. **17**(2), (2013). http://translationjournal.net/journal/64teaching.htm
7. Al Hassan, A.: A corpora-driven approach for the Sudanese EFL translation classroom: moving beyond bilingual dictionaries and intuition. ADAB 31–47 (2015). https://doi.org/10.46673/1311-000-035-013
8. Classroom, T.: Moving towards a corpus-driven pedagogy for Omani translation major students. Arab World English J. **12**(1), 40–58 (2021). https://doi.org/10.24093/AWEJ/VOL12NO1.4

9. Mădălina Chitez, L.P.: Digital methods in translation studies: using corpus data to assess trainee translations. B.A.S. British Am. Stud. **26**(26), 241–251 (2020)
10. Nuruzzaman, M., Hussain, O.K.: A survey on Chatbot implementation in customer service industry through deep neural networks. In: Proceedings - 2018 IEEE 15th International Conference on e-Business Engineering, ICEBE 2018, pp. 54–61, December 2018. https://doi.org/10.1109/ICEBE.2018.00019
11. Ayanouz, S., Abdelhakim, B.A., Benhmed, M.: A smart chatbot architecture based NLP and machine learning for health care assistance. ACM Int. Conf. Proc. Ser. (2020). https://doi.org/10.1145/3386723.3387897
12. Niranjan, M., Saipreethy, M.S., Kumar, T.G.: An intelligent question answering conversational agent using Naïve Bayesian classifier (2012). https://doi.org/10.1109/ICTEE.2012.6208614
13. Aljameel, S.S., O'Shea, J.D., Crockett, K.A., Latham, A., Kaleem, M.: Development of an Arabic conversational intelligent tutoring system for education of children with ASD. In: 2017 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications, CIVEMSA 2017 - Proceedings, pp. 24–29, July 2017. https://doi.org/10.1109/CIVEMSA.2017.7995296
14. Ali, D.A., Habash, N.: Botta: an Arabic dialect Chatbot, pp. 208–212 (2016)
15. Al-Ghadhban, D., Al-Twairesh, N.: Nabiha: an Arabic dialect chatbot. Int. J. Adv. Comput. Sci. Appl. **11**(3), 452–459 (2020). https://doi.org/10.14569/IJACSA.2020.0110357
16. AlHumoud, S., Al Wazrah, A., Aldamegh, W.: Arabic Chatbots: a survey. Int. J. Adv. Comput. Sci. Appl. **9**(8), 535–541 (2018). https://doi.org/10.14569/IJACSA.2018.090867
17. Alkhatib, M., Shaalan, K.: The key challenges for Arabic machine translation. In: Shaalan, K., Hassanien, A.E., Tolba, F. (eds.) Intelligent Natural Language Processing: Trends and Applications. SCI, vol. 740, pp. 139–156. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-67056-0_8
18. Al-Harthi, M., Al-Saif, A.: The design of the SauLTC application for the English-Arabic learner translation corpus. In: Proceedings of the 3rd Workshop on Arabic Corpus Linguistics, pp. 80–88 (2019). Accessed 8 Dec 2021. https://aclanthology.org/W19-5610