Journal of Al-Azhar University Engineering Sector

Vol. 18, No. 67, April 2023, 413-422

DEVELOPMENT THE DATASET FOR AUTOMATIC TRANSLATION SYSTEM

Dahey G. Ghanem*

Department of Systems Engineering and Computers, Faculty of Engineering, Al-Azhar University, Egypt.

*Correspondence: adahey@yahoo.com

Citation:

D. G. Ghanem, "Development the dataset for automatic translation system", Journal of Al-Azhar University Engineering Sector, vol. 18, pp. 414-423, 2023.

Received: 3 September 2022

Accepted: 15 January 2023

ABSTRACT

The Automatic translation systems (ATS) for translation text have extent widely in recent years. The ARS developed to correct several types of text errors explained by the Mossop's prototype such as spelling, typographical, syntactic, semantic, word, and formal ones. The ARS need a large amount of data training in its forms. There is a shortage in German-Arabic datasets for translation and revision purposes. Building dataset is the most time-consuming and the most important part of the text translation process. We make an effort to analyze and work on this large amount of data Sentences, and the form of text free dataset on the ARS, most efforts focus on German and Arabic data. Despite the increase in the number of Arabic, users and the increase in Arabic content on ARS. Therefore, in this paper, Arabic dataset built to use in text translation purpose. This research offers the German-Arabic dataset from the Taxonomy of errors in post-editing text for growth the ARS. Our dataset gathered from A Game of Throne saga in German (GR) and Arabic (AR) saga. Our dataset consists of 65,000 bilingual sentences collected from Text. The most significant penalties of this research were the Mossop's prototype terminates to explain all errors; and the prototype had to be lengthy in demand to include the Consistency. Finally, human evaluators were employed to grade the quality of ATS outputs and to revision them. We used a Rapid Miner tool to evaluate the performance of our dataset, the dataset accuracy of 95.12%.

KEYWORDS: ATS Errors, APES Errors, ARS Errors, classification of errors in Translation text, (GR-AR) corpus

بناء جسم لغوي لنظام الترجمة الآلي

ضاحي جابر غانم*

قسم هندسة النظم والحاسبات , كلية الهندسة , جامعة الازهر, القاهرة , مصر

*البريد الاليكتروني للباحث الرئيسى : adahey@yahoo.com

الملخص

إنتشرت أنظمة المراجعة التلقائية للنص المترجم على نطاق واسع في السنوات الأخيرة. تم تطوير نظام المراجعة التلقائية للنص المترجم لتصحيح عدة أنواع من أخطاء النص المترجم التي ذكرها النموذج الأولي لموسوب مثل الإملائية ، والمطبعية ، والنحوية ، والدلالية ، والكلامية ، والشكلية. يحتاج نظام المراجعة التلقائية للنص المترجم إلى كمية كبيرة من البيانات في أشكالها لعمل تدريب عليها. هناك نقص في مجموعات البيانات الألمانية-العربية لأغراض الترجمة والمراجعة. يعد إنشاء مجموعة البيانات الجزء الأكثر استهلاكا للوقت والأكثر أهمية في عملية ترجمة النص. لقد قمنا ببذل جهدًا لتحليل هذه الكمية الكبيرة من جمل البيانات والعمل عليها وتشكيل مجموعة البيانات النصية الحالية من نظام المراجعة التلقائية للنص المترجم ، وتركز معظم الجهود على البيانات الألمانية والعربية. على الرغم من زيادة عدد المستخدمين للغة العربية وزيادة المحتوى العربي على نظام المراجعة التلقائية للنص المترجم. لذلك في هذه الورقة ، تم بناء مجموعة البيانات الألمانية-العربية لاستخدامها في أغراض ترجمة النص. يقدم هذا البحث مجموعة البيانات الألمانية العربية من تصنيف الأخطاء في نص ما بعد تصحيح الترجمة لنظام المراجعة التلقائية للنص المترجم. تم جمع مجموعة البيانات الخاصة بنا من ملحمة لعبة العروش باللغتين الألمانية والعربية وتتكون مجموعة البيانات الخاصة بنا من 65000 جملة ثنائية اللغة تم جمعها من النص. كانت أهم نتائج هذا البحث هي عجز النموذج الأولي لموسوب لشرح جميع الأخطاء ؛ وكان يجب أن يكون النموذج الأولي طويلاً ليشمل الاتساق. قمنا بتقييم صحة مجموعة البيانات الخاصة بالترجمة الألية والتدقيق بواسطة الخبراء البشريين. استخدمنا أداة Rapid Miner لتقييم أداء مجموعة البيانات الخاصة بنا وكانت دقة مجموعة البيانات 95.12٪.

الكلمات المفتاحية : أخطاء نظام الترجمة الألية ، أخطاء نظام التدقيق الألي ، تصنيف الأخطاء في نص الترجمة ، مجموعة البيانات الألمانية-العربية.

1. INTRODUCTION

1.1 The Automatic Revision System (ATS)

The task of ATS is to automatically revision ATS outputs. We have so far examined only classic baseline methods based on phrase based statistical ATS. The first system trained on the data [1]. However, this system tended to deteriorate the translation quality in terms of BLUE and TER, apparently due to the scarcity of training data. Then, our second model (b) introduced identical pairs of sentences in the target side of our corpus predictably retain grammatical fragment within outputs. By decoding the outputs using the multiple decoding path ability of ATS, this model significantly improved the naive base-line system, but the translation quality was not consistently better depending on the language pair. Finally, we introduced in the third system ARS, yet another phrase table learned from pseudo training data. Our pseudo training data obtained in the same manner; we coupled each of the decoded result to its corresponding reference translation in the corpus. Automatic revision system (ARS) is nowadays the most familiar interactive to translate and publish information to exchange of particular types of content, including text. The errors going on the translation text from the inadequate request of a translation procedure [1], in extra words, translation fault occurs every time the translation text is unsuccessful to realize its drive. It can happen at word, phrase, sentence, and textual pragmatic level. Fig. 1 shows The ATS with APES and ARS.

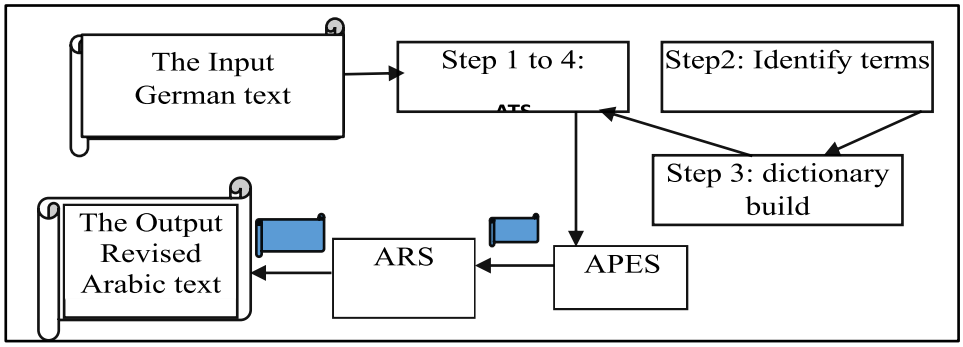

Fig. 1: The ATS with APES and ARS

Development the German-Arabic datasets for ARS is a technique complete together with data mining, machine learning, and information retrieval. It may also point out as text analysis in which significant information can be saved from the text. The text pre-processing phases are parsing, tokenization, normalization, etc. [2].

1.2 Research Objectives and Aims

Our objectives build the largest publically free accessible ATS Corpora. Full study about the impact of text pre-processing on Arabic text classification and resolve the ambiguity in the works. Provide comprehensive guidelines to help in making trade -off between accuracy and time storage requirements. The first aim of this research is to compile representative Arabic corpora that covers different text types, which will be used in this research and can used in this research and in the future as a benchmark. The corpora collected from different sources and various domains. The second aim Implement and integrate Arabic morphological analysis tools into leading open machine learning tools. The third aim was to implement Arabic morphological analysis tools applied on Arabic corpora. Apply different term weighting systems (Boolean, word count, word

count normalization, term frequency, term frequency inverse document frequency, and term pruning) on Arabic corpora.

Arabic Language is the fifth most widely used language in the world. It is used by more than 422 million people as a first language and by 250 million as a second language [2]. Arabic Language belongs to the Semitic language family. Semitic languages used without the vowel marks, which would indicate the short vowels. Semitic languages can get away with this because they all have a predictable root pattern system [2]. Arabic is fourth one of the top ten languages on the internet [3]. The need and attention in Arabic dataset have increased recently, due to many reasons: Arabic language is very rich with contents, there are about 184 million Arab Internet users and a large percentage of them cannot read German [3]. German is a major European language, and it has special characteristics such as a relatively free word order and a rich morphology. These characteristics mean that a parsing approach that is suitable for Arabic is not automatically so for German. While German parsers typically perform, worse than Arabic ones, the disagreement whether parsing German is an inherently harder task than parsing Arabic is still open [1]. In addition, the German contents have grown quickly in the last decade [4]. However, there is lack of language resources and text processing techniques for the German-Arabic dataset [5]. One of the difficulties that encounter this work and other research in the field of Arabic linguistics was the lack of publicly available Arabic corpus for evaluating text categorization algorithms. The Linguistic Data Consortium (LDC) provides two non-free Arabic corpora, the Arabic NEWSWIRE and Arabic Gig word corpus. Both corpora contain newswire stories.

We collected our dataset from A Game of Throne saga in German (GR) and Arabic (AR) saga, the dataset contains 65,000 sentences, collected and manually considered, and then we apply some text processing techniques which include, removing non-Arabic letters, removing word suffix and prefix, normalization, and transformation. The organization phase includes two stages, in the first, we labelled each sentence to a specific class, then we ask 10 Text users who are German-Arabic native speakers to label the sentences, and according to users' feedbacks, some sentences classification is changed.

In the remainder of this paper, we first describe the GR/AR datasets Related Works in Section 2. Then, in Section 3, we present Research Problems. Section 4 describes Procedure for Dataset Construction. Section 5 describes Dataset Building phases. Finally, Section 6 describes Conclusions and Future Works.

2. RELATED WORKS

There are various corpora to perform machine-learning systems, the corpora variations include small/large size corpus, with few and more categories. The most famous three corpora are available publically at [68]. The corpus that collected by Latifa Al-Sulaiti and Eric Atwell includes Contemporary Corpus of Arabic (CCA) [6, 7] from the University of Leeds. Their study confirms that the existing corpora are too narrowly limited in source-type and genre, and that there is a need for a freely accessible corpus of contemporary Arabic covering a broad range of text-types. The corpus contains 293 text documents belonging to 1 of 5 categories. The corpus includes 95,530 district keywords after stop words removal. Aljazeera corpus was used by [8], the corpus includes 1,500 text documents, each text document belongs one of two five categories, each category includes 300 documents. The corpus includes 55,376 district keywords after stop words removal. Khaleej-2004 corpus was collected by [9, 10] from Khaleej newspaper of the year 2004. The corpus includes 5,690 text documents. Each text document belongs to one of four categories. The corpus includes 122,062 district keywords after stop words removal. The BBC Arabic corpus from BBC Arabic website bbcarabic.com, the corpus includes 4,763 text documents. Each text document belongs to one of seven categories. The corpus contains 1,860,786 (1.8M) words and 106,733 district keywords after stop words removal. The corpus is available publically at [11]. The CNN Arabic corpus from CNN Arabic website cnnarabic.com, the corpus includes 5,070 text documents.

Each text document belongs to one of six categories. The corpus contains 2,241,348 (2.2M) words and 144,460 district keywords after stop words removal. The corpus is available publically at [12].

3. RESEARCH PROBLEMS

The following points describe the research problems:

1. The lack of availability of publically free accessible Arabic Corpora, most of related works in the literature used small in-house collected corpus.
2. The aids of using Arabic morphological tools are not addressed for Arabic Language.
3. The impact of text pre-processing on Arabic text classification using popular text classification algorithms not studied in the works.

4. PROCEDURE FOR DATASET CONSTRUCTION

We have created our ATS datasets, regarding German as the source language. We have so far regarded Arabic as the target languages, considering that the writers of these languages hold the largest number of visitors to German [12]. Following the procedure practices in ATS [13], we determined the following five-step process.

1. Collecting German sentences (src)
2. Generation of ATS Arabic outputs (hyp)
3. Manual Arabic translation (ref)
4. Manual grading of ATS Arabic output (grade)
5. Manual revision of ATS Arabic output (R).

The most time exhaustion and the most important phase of text mining is Data collection [9]. For the last three tasks, we allocated Arabic writer of the target language who also understand German.

4.1. Collection German Sentences

First, we collected the following two sets of sentences in German that have used with our writing translation service. Arabic sentences, we randomly sampled 65,000 identical write segments that identified as German by its automatic translation module. The extracted segments, especially those in the rhetorical domain, include ungrammatical ones, non-understandable ones, and those containing inappropriate expressions with respect to social standards. We therefore asked German writer to filter out such segments. Many segments do not have a clear subject, as German is a pro-drop language; even required arguments can be missing.

4.2. Generation of ATS Outputs

The collected German segments (src) then translated by our in-house ATS, which implement a phrase based statistical ATS. The GR/AR translations obtained with the system trained on 736k sentences pairs.

4.3. Manual Translation

Reference translations manually given, referring only to the source segments (src). As each src is not attributed with its specific context, we asked the translators to imagine some context as long as it is practical considering the domain. On the contrary, we also asked to avoid adding too much content that cannot specified only from the src. For the src, this has more than one interpretation, only one translation given rather than enumerating all the possible interpretations.

4.4. Manual Classifying of ATS Output

The quality of ATS output (hyp) with respect to its source (src) graded according to a standard, which is compatible with the "Acceptability" criterion. In case the evaluator cannot understand the meaning of src, she/he allowed referring to the corresponding reference translation (ref), with an advice that it is not only the correct translation.

4.5. Manual Revision of ATS Output

Human workers asked to revision ATS outputs; produce Re in the following guidance.

1. Refer only to source and outputs. Refer also to ref if necessary.
2. Make each output grammatically and semantically appropriate with respect to its source.
3. Perform minimal revision, as we use Re for the reference.

The workers were also informed that we consider the following four revision operations equally.
Deletion of a word: Delete an unnecessary word.
Insertion of a word: Insert a missing but necessary word.
Substitution of a word: Substitute a word with another word.
Shift of a word or a phrase: Change the word order by moving a single word or a sequence.

5. DATASET BUILDING PHASES

In this paper, we will explain the overview of the dataset development process. This process is divided into three phases, data Acquisition phase, data filtering phase, and data-labelling phase. Fig. 2 shows the dataset development phases [13].

Fig. 2: Dataset development phases

The following steps describe the dataset development phases:

1. Collect data by packing the Arabic Text.
2. Filter the data collected in the previous step.
   a. Removing non-Arabic
   b. Remove the repeated sentences.
3. Manually label the filtered sentences to one of the ten categories chosen.

We chose the dataset to cover the German-Arabic dataset subjects. Figure 1 shows the process of building the proposed dataset. The algorithm used for building the dataset can be summarized as; the first phase in construction process is to build a text dataset, which involves compiling and labelling text documents into corpus. we collected our dataset from A Game of Throne saga in German (GR) and Arabic (AR) saga, the dataset contains 65,000 sentences, collected and manually considered, and then we apply some text processing techniques which include, removing non-Arabic letters, removing word suffix and prefix, normalization, and transformation [14].

5.1. Data Acquisition Phase

We have collected the 65,000 German-Arabic Text sentences in Excel browser (it can collect sentences by automatically scrolling down the Text page to show all sentences) to collect the German sentences with the equivalent Google translation Arabic sentences and compare these sentences with the Arabic sentences in A Game of Throne saga in Arabic (AR) saga. The errors between these sentences covering 12 revision constraints delivered in Mossop menu [15], separated into four collections:-

Group A – Transmission – Difficulty of sensory transmission
Group B – Gratified – Difficulties of gratified
Group C – Linguistic – Difficulties of linguistic and stylish
Group D –Appearance – Difficulties of appearance lower

The instruction to check the constraints, the operation order [16]:

1- Deliver the complete change for any ambiguity; style, terminology; and idiomatic.
2- Create a relative form for Transmission, i.e., whether the text is correct or not.
3- Deliver the whole translation check for any errors of punctuation and grammar.
4- Trendy statistics are significant in the text; prepare a distinct form.
5- Check the text for headers, footnotes, page numbering, table of insides, orientations.
6- Modified the next steps to change the necessary errors by the reviser force be lost.
7- The text saved after all the modifications.

Translation Faults in a Game of Thrones; we select the 73 chapters of A Game of Thrones analyzed. From select chapters, 129 dataset errors existed. The 98 considered versions to six of the classes mined from Mossop's Prototype as shown in Table 1, a new category had to create for them: the post-editor limit Consistency [17].

Table 1: the dataset classes

Types	Number of Incidences
Expression	38
Correctness	35
Flatness	13
Wholeness	7
Reason	3
Mechanism	2
Whole	98

5.2. Dataset Filtering Phase

Making and Pre-processing the Dataset, this phase included removal of the following types of sentences non-Arabic sentences and repeated sentences; the sentence added only once. The Arabic text only be saved, and the remaining parts will remove. To begin, each line in the dataset is a tab-delimited pair consisting of a German text sequence and the translated Arabic text sequence. German called the source language and Arabic called the target language. After making the dataset, we precede with several pre-processing steps for the raw text data. For instance, we replace non-breaking space with space, convert uppercase letters to lowercase ones, and insert space between words and punctuation marks [18].

5.2.1. Tokenization

Unlike the character level tokenization in for automatic revision system, the following tokenize method tokenizes the first text sequence pairs, where each token is either a word or a punctuation mark. We add the special "<eos>" token to the end of every sequence to indicate

the end of the sequence. When a model is predicting by generating sequence token after token, the generation of the "<eos>" token can suggest that the output sequence is complete [19].

5.2.2. Loading Sequences of Fixed Length

Recall that in language modelling each example sequence, either a segment of one sentence or a span over multiple sentences, had a fixed length. In automatic revision system, each example is a pair of source and target text sequences, where the two text, sequences may have different lengths. For computational efficiency, we can still process a mini group of text sequences at one time by truncation and padding. Suppose that every sequence in the same mini batch should have the same length num_steps. If a text sequence has fewer than num_steps tokens, we will keep appending the special "<pad>" token to its end until its length reaches num_steps. Otherwise, we will truncate the text sequence by only taking its first num_steps tokens and discarding the remaining. In this way, every text sequence will be the same length to be loaded in mini batches of the same shape. Besides, we also record length of the source sequence excluding padding tokens. We build two vocabularies for both the source language and the target language separately. With word-level tokenization, the vocabulary size will be significantly larger than that using character-level tokenization. To ease this, here we treat infrequent tokens that appear less than 2 times as the same unknown ("<unk>") token. when training with target sequences, the decoder output (label tokens) can be the same decoder input (target tokens), shifted by one token; and the special beginning-of-sequence "<bos>" token will be used as the first input token for predicting the target sequence [20].

5.3. Dataset Labelling Phase

Reading the filtered sentences used in the labeling phase, we introduce the automatic revision system dataset that we will use in the system. First, we will need some new code to process our data. Unlike the language modelling, here each example consists of two separate text sequences, one in the source language and another in the target language. Finally, we define the get dataset loader to return the data iterator. Let us read the first mini batch from the German-Arabic dataset. We show a pair of source and target sequences that are processed by the above build arrays method (in the string format). To assess our labelling made during the first part, the dataset given to ten German-Arabic native speakers who additionally confirm the validity of the dataset created. The source document of the current training contains the chief eleven sections of the chief volume of the A Song of Ice and Fire story – called A Game of Thrones – that has interpreted from German text (GR) to Arabic text (AR). The document alienated into two regular texts [21]:

1- The original German text, it is the source text (ST);
2- Its translation Arabic text, these texts will be contained within the alignment system designed for texts fixing of translation faults.

The alignment system mechanism with parallel corpus [22], But earlier data is prepared, it was aligned. As clarified by [23]. Used for the arrangement of the parallel extracts of both ST and TT.

5.3.1. The Occurrences of Dataset Files

Toward brand, the translation mistakes correction, mechanically we practice Google Translation System (GTS) to translate the source German text to the target Arabic text. The output of GTS is grammatically unsuitable due to the lack of verbal rules for the language pair fact practical. Syntactic mistakes fail the fluency and adequacy of the translation and the BLEU scores [24] range only between 0.21 and 0.29, depending on test sets and numbers of reference translations. Table 2 displays the occurrences of dataset [25].

Table 2: The occurrences of dataset

The input German text	The Arabic text result by ATS	The Arabic text	Errors category	Error parameter
Der winter komATS	الشتاء قادم	الشتاء قادم	Post-editing	Consistency
Gared gehörte seit vierzig Jahren der Nachtwache an, als Mann und schon als Junge, und er war es nicht gewohnt, dass man sich über ihn lustig machte.	كان جارد حارساً ليلياً لمدة أربعين عاماً ، كرجل وكصبي ، ولم يكن معتاداً على السخرية منه.	لقد قضي الشيخ أربعين عاما كاملة مع حرس الليل,منذ إلتحق بهم وهو صبي. ولم يكن يروق له أن يستخف به الأخرين	Language and style	Idiom
Vier Jahre war er auf der Mauer. Als man ihn zum ersten Mal auf die andere Seite geschickt hatte, waren ihm all die alten Geschichten wieder eingefallen, und fast war ihm das Herz in die Hose gerutscht.	كان على الحائط لمدة أربع سنوات. في المرة الأولى التي تم إرساله فيها إلى الجانب الآخر ، كانت جميع القصص القديمة قد عادت إليه ، وكان قلبه ينزلق في سرواله.	ويوم أرسلوه وراء الجدارللمرة الأولي وجد الحكايات القديمة تتدفق من ذاكرته وشعر بأمعائه تتقلص	Meaning transfer	Completeness
»Ich wette, die hat er alle eigenhändig gemeuchelt, der Mann«, hatte Gared in der Kaserne beim Wein erklärt, »hat den kleinen Biestern die Hälse umgedreht, unser großer Krieger. « Alle hatten in sein Lachen mit eingestimATS.	أراهن أنه جعل كل "، من يديه باليد ، الرجل ،" أوضح غاريد في الثكنات فوق الخمر ، "تحولت رقاب الوحوش الصغيرة ، محاربنا العظيم." لقد انضم الجميع إلى ضحكته.	أراهن أن محاربنا العظيم قتلهم جميعا بنفسه. أراهن أنه كسر أعناقهم الصغيرة	Language and style	Idiom
Was gibt es da?	ما هو هناك؟	من هناك ؟	Language and style	Smoothness
Will trat an den Baum, einen gewölbten, graugrünen Wachbaum, und begann zu klettern.	صعد إلى الشجرة ، شجرة حراسة خضراء رمادية مقببة ، وبدأ في الصعود.	إتجه إلي شجرة الحارس الضخمة ذات الأفرع المقنطرة واللون الأخضروالرمادي وبدأ يتسلق	Meaning transfer	Accuracy
Der Andere zögerte. Will sah seine Augen, dunkler und blauer, als Menschenaugen jemals sein konnten, ein Blau, das brannte wie Eis. Sie richteten sich auf das Langschwert, das dort oben bebte, betrachteten das Mondlicht, das kalt über das Metall lief. Einen Herzschlag lang wagte er zu hoffen.	الآخر تردد. سوف يرى عينيه ، أكثر قتامة وأكثر زرقة من العين البشرية يمكن أن تكون ، زرقاء تحترق مثل الثلج. كانوا يستهدفون الفستان الطويل الذي يرتعد هناك ، يراقبون ضوء القمر البارد فوق المعدن. لدقات قلب تجرأ على الأمل.	توقف الأخر ,ورأي ويل عينيه . كانتا ذات لون أزرق شديد العمق,يحرق كالجليد,أعمق وأكثر زرقة من أي عين بشرية ,وقد ثبتت نظراتهما علي السيف الطويل الذي يرتفع مرتجفا في يد صاحبه,وراقبتا نور القمر البارد يجري علي المعدن,وللحظة جرؤ ويل علي الأمل	Content	Logic
Er nannte sie »kleine Prinzessin« und manchmal »Mylady«, und seine Hände waren weich wie altes Leder.	دعاها "الأميرة الصغيرة" وأحيانا "سيدتي" ، وكانت يديه ناعمة مثل الجلود القديمة.	وقال أمرا :"قفي مكانك ", ثم "إستديري. نعم عظيم. تبدين "بهية كملكة" وكان يمرر ..." يده علي صفحة الماء	Language and style	Mechanics
Heute war der schlimmste von allen. Kalter Wind wehte von Norden her und ließ die Bäume rascheln, als wären sie lebendig. Den ganzen Tag schon schien es Will, als würden sie beobachtet, von etwas Kaltem, Unerbittlichem.	اليوم كان الأسوأ من كل شيء. هبت الرياح الباردة من الشمال وجعلت الأشجار تسقط كما لو كانوا أحياء. طوال اليوم ، يبدو أن يراقب شيئا باردا ولا يرحم.	أم اليوم فالأسوء علي الإطلاق ,فالرياح الباردة تهب من الشمال وتجعل حفيف الأشجار كأنه حركة عشرات الكائنات , وصاحب ويل طويل اليوم شعور بأن ثمة شئ ما يراقبه , شيئا باردا قاسيا لا يكن له أي مودة	Post-editing	Case of Consistency - Converted Units of Measurement

CONCLUSION

Aiming to promote the research on quality estimation (QE) of ATS outputs, especially for those among German languages, we have created new datasets for the German to Arabic translation tasks. This paper described the process of corpus creation and observations from the created datasets. We also presented our benchmarking experiments using the created datasets, for all of the tasks in our worry: word-level QE, two variants of sentence-level QE, and ATS. Although the methods examined in this paper could be far from the state-of-the-art, we confirmed that the performance of these tasks can be improved by introducing features and pseudo training data that had been proven useful in the literature.

We believe that this corpus will be valuable to advance re-search efforts in the machine translation area since the ATS community often needs manually annotated data. We believe that our methodology for guideline development and annotation consistency checking can applied in other projects and other languages as well. To evaluate our dataset Rapid Miner tool used and the performance achieved in terms of accuracy was 95.12% with the SVM model. Our dataset will help researchers in the field of Arabic text processing.

Following the appearance of neural ATS, we are now working on extending the datasets with translations of such systems. We are planning to improve the performance on the QE/ ARS tasks, and to investigate applications of the technologies, including enhancing the functionality of our writing translation service, and filtering automatically collected parallel sentences. In the future, we plan to maximize the use of this dataset by using it in improving ATS between German and Arabic in both directions. In the future, we plan to increase the size of the corpus and to add other corpus domains.

REFERENCES

[1] Lin, Z.; Madotto, A.; Winata, G.I.; Xu, P.; Jiang, F.; Hu, Y.; Shi, C.; Fung, P. BiToD: A Bilingual Multi-Domain Dataset For Task-Oriented Dialogue Modeling. arXiv 2021, arXiv:2106.02787.

[2] Xue, L.; Constant, N.; Roberts, A.; Kale, M.; Al-Rfou', R.; Siddhant, A.; Barua, A.; Raffel, C. ATS5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, online, 15–20 June 2020; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 483–498. [CrossRef]

[3] Qin, L.; Xu, X.; Che, W.; Zhang, Y.; Liu, T. Dynamic Fusion Network for Multi-Domain End-to-end Task-Oriented Dialog. In Proceedings of the 58th annual Meeting of the Association for Computational Linguistics, online, 5–10 July 2020; pp. 6344–6354. [CrossRef]

[4] Ham, D.; Lee, J.-G.; Jang, Y.; Kim, K.-E. End-to-End Neural Pipeline for Goal-Oriented Dialogue Systems using GPT-2. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, online, 5–10 July 2020; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; Volume 2, pp. 583–592. [CrossRef]

[5] Peng, B.; Li, C.; Li, J.; Shayandeh, S.; Liden, L.; Gao, J. SOLOIST: Building Task Bots at Scale with Transfer Learning and Machine Teaching. Trans. Assoc. Comput. Linguist. 2021, 9, 807–824. [CrossRef]

[6] Yang, Y.; Li, Y.; Quan, X. UBAR: Towards Fully End-to-End Task-Oriented Dialog Systems with GPT-2. arXiv 2020, arXiv:2012.03539.

[7] Wang, W.; Zhang, Z.; Guo, J.; Dai, Y.; Chen, B.; Luo, W. Task-Oriented Dialogue System as Natural Language Generation. arXiv 2021, arXiv:2108.13679.

[8] Peng, B.; Zhu, C.; Li, C.; Li, X.; Li, J.; Zeng, M.; Gao, J. Few-shot Natural Language Generation for Task-Oriented Dialog. arXiv 2020, arXiv:2002.12328.

[9] Wu, C.-S.; Hoi, S.C.H.; Socher, R.; Xiong, C. TOD-BERT: Pre-trained Natural Language Understanding for Task-Oriented Dialogue. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), online, 16–20 November

2020; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 917–929. [CrossRef]
[10] Madotto, A.; Liu, Z.; Lin, Z.; Fung, P. Language Models as Few-Shot Learner for Task-Oriented Dialogue Systems. arXiv 2020, arXiv:2008.06239.
[11] Campagna, G.; Foryciarz, A.; Moradshahi, M.; Lam, M. Zero-Shot Transfer Learning with Synthesized Data for Multi-Domain Dialogue State Tracking. arXiv 2020, arXiv:2005.00891.
[12] Zhang, Y.; Ou, Z.; Hu, M.; Feng, J. A Probabilistic End-To-End Task-Oriented Dialog Model with Latent Belief States towards Semi-Supervised Learning. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), online, 16–20 November 2020; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 9207–9219. [CrossRef]
[13] Kulhánek, J.; Hudeˇcek, V.; Nekvinda, T.; Dušek, O. AuGPT: Dialogue with Pre-trained Language Models and Data Augmentation. arXiv 2021, arXiv:2102.05126.
[14] Kim, S.; Yang, S.; Kim, G.; Lee, S.-W. Efficient Dialogue State Tracking by Selectively Overwriting Memory. arXiv 2020, arXiv:1911.03906. [CrossRef]
[15] Kumar, A.; Ku, P.; Goyal, A.; Metallinou, A.; Hakkani-Tur, D. MA-DST: Multi-Attention-Based Scalable Dialog State Tracking. Proc. Conf. AAAI Artif. Intell. 2020, 34, 8107–8114. [CrossRef]
[16] Heck, M.; van Niekerk, C.; Lubis, N.; Geishauser, C.; Lin, H.-C.; Moresi, M.; Gaši´c, M. TripPy: A Triple Copy Strategy for Value Independent Neural Dialog State Tracking. arXiv 2020, arXiv:2005.02877.
[17] Li, S.; Yavuz, S.; Hashimoto, K.; Li, J.; Niu, T.; Rajani, N.; Yan, X.; Zhou, Y.; Xiong, C. CoCo: Controllable Counterfactuals for Evaluating Dialogue State Trackers. arXiv 2020, arXiv:2010.12850.
[18] Wang, D.; Lin, C.; Liu, Q.; Wong, K.-F. Fast and Scalable Dialogue State Tracking with Explicit Modular Decomposition. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Mexico City, Mexico, 6–11 June 2021; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 289–295. [CrossRef]
[19] Liu, Z.; Winata, G.I.; Lin, Z.; Xu, P.; Fung, P. Attention-informed mixed-language training for zero-shot cross-lingual task-oriented dialogue systems. Proc. AAAI Conf. Artif. Intell. 2020, 34, 8433–8440. [CrossRef]
[20] Zhang, Y.; Ou, Z.; Yu, Z. Task-oriented dialog systems that consider multiple appropriate responses under the same context. Proc. AAAI Conf. Artif. Intell. 2020, 34, 9604–9611. [CrossRef]
[21] Al-Ajmi, A.H.; Al-Twairesh, N. Building an Arabic Flight Booking Dialogue System Using a Hybrid Rule-Based and Data Driven Approach. IEEE Access 2021, 9, 7043–7053. [CrossRef]
[22] Bendjamaa, F.; Nora, T. A Dialogue-System Using a Qur'anic Ontology. In Proceedings of the 2020 Second International Conference on Embedded & Distributed Systems (EDiS), Oran, Algeria, 3 November 2020; pp. 167–171.
[23] Al-Ghadhban, N.; Al-Twairesh, D. Nabiha: An Arabic dialect chatbot. Int. J. Adv. Comput. Sci. Appl. Int. J. Adv. Comput. Sci. Appl. 2020, 11, 452–459. [CrossRef]
[24] Mayeesha, T.T.; Sarwar, A.M.; Rahman, R.M. Deep learning based question answering system in Bengali. J. Inf. Telecommun. 2020, 5, 145–178. [CrossRef]
[25] Naous, T.; Hokayem, C.; Hajj, H. Empathy-driven Arabic Conversational Chatbot. In Proceedings of the Fifth Arabic Natural Language Processing Workshop, Barcelona, Spain, 8 December 2020; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 58–68.