Building an Arabic Machine Translation Post-Edited Corpus: Guidelines and Annotation

Wajdi Zaghouani, Nizar Habash[†], Ossama Obeid, Behrang Mohit[‡], Houda Bouamor and Kemal Oflazer

Carnegie Mellon University in Qatar, [†]New York University Abu Dhabi, [‡]Ask.com

{wajdiz,oobeid,hbouamor}@cmu.edu, ko@cs.cmu.edu

[†]nizar.habash@nyu.edu, [‡]behrangm@ischool.berkeley.edu

Abstract

We present our guidelines and annotation procedure to create a human corrected machine translated post-edited corpus for the Modern Standard Arabic. Our overarching goal is to use the annotated corpus to develop automatic machine translation post-editing systems for Arabic that can be used to help accelerate the human revision process of translated texts. The creation of any manually annotated corpus usually presents many challenges. In order to address these challenges, we created comprehensive and simplified annotation guidelines which were used by a team of five annotators and one lead annotator. In order to ensure a high annotation agreement between the annotators, multiple training sessions were held and regular inter-annotator agreement measures were performed to check the annotation quality. The created corpus of manual post-edited translations of English to Arabic articles is the largest to date for this language pair.

Keywords: Post-Editing, Guidelines, Annotation

1. Introduction

In recent years, machine translation (MT) became widely used by translation companies to reduce their costs and improve their speed. Therefore, the demand for quick and accurate machine translations is growing. Machine translation (MT) systems often produce incorrect output with many grammatical and lexical choice errors. Correcting machine-produced translation errors, or MT Post-Editing (PE) can be done automatically or manually. Successful automatic post-editing approaches using manually corrected MT output were used by Elming (2006) and Simard et al. (2007). The availability of annotated resources is required for such approaches. When it comes to the Arabic language, to the best of our knowledge, there is no manually post-edited MT corpora available to build such systems. Therefore, there is a clear need to build such valuable resources for the Arabic language.

In this paper, we present our guidelines and annotation procedure to create a human corrected MT corpus for the Modern Standard Arabic (MSA). The creation of any manually annotated corpus usually presents many challenges. In order to address these challenges, we created comprehensive and simplified annotation guidelines which were used by a team of five annotators and one lead annotator. In order to ensure a high annotation agreement between the annotators, multiple training sessions were held and regular inter annotator agreement (IAA) measures were performed to check the annotation quality. To the best of our knowledge, this is the first published machine translation manual post-editing annotation effort for Arabic of this scale.

In the next sections, we review related work (Section 2), describe our corpus and the development of the guidelines (Sections 3-4), and present our annotation procedure (Section 5), than we present the annotation evaluation in Section 6, finally we conclude our work in Section 7.

2. Related Work

Large scale manually corrected MT corpora are not yet widely available due to the high cost related to building such resources. Wisniewski et al. (2014) created a corpus of machine translation errors extracted from several translation students taking part in a master program in specialized translations. The texts are translated from English to French. A portion of the corpus contains an analysis of the type of errors made by the MT system. Elming (2006) created a 265K-word English-Danish MT manually corrected corpus by a human professional translator. The full corpus covers the chemical patents domain. Simard et al. (2007) created a 500K-word corpus of manually edited French-English and English-French MT from the Canadian Job Bank website. The corpus is a collection of blocks composed of the source language texts, the machine translation output of a rule-based MT system and the final post-edited version done by a human translator. Moreover, Avramidis et al. (2014) built a corpus of human-annotated machine translations which was evaluated by professional human translators for the following three language pairs: German-English, English-German and Spanish-German.

Fishel et al. (2012) created a corpus of automatically produced translations with detailed manual translation error analysis of 576 sentences for four language pairs: English-Czech;French-German;German-English;English-Serbian.

Popescu-belis et al. (2002) produced a small corpus of 50 texts translated by students and corrected by their professors and all translation errors are annotated with their corrections in this corpus. For Arabic, we cite the effort of Bouamor et al. (2014) who created a medium scale human judgment corpus of Arabic machine translation using the output of six MT systems and a total of 1892 sentences and 22K rankings.

Our corpus is a part of the Qatar Arabic Language Bank (QALB) project, a large scale manually annotated annotation project (Zaghouani et al., 2014b; Zaghouani et al.,

2015). The project goal was to create an error corrected 2M-words corpus for online user comments on news websites, native speaker essays, non-native speaker essays and machine translation output. The 100K-word machine translation portion was selected from various Wikinews English articles translated to Arabic automatically using the Google Translate tool.[1]

3. Corpus Description

We collected a 100K-word corpus of English news articles taken from the collaborative journalism Wikinews website.[2] Since Wikinews is a free-content news source, we avoided any copyrights complications. The corpus includes 520 articles with an average of 192 words per article. The articles cover mostly political news and they are selected from the completed version of articles since some the recent Wikinews articles may still be edited or updated. The original English files were in HTML format and were exported to a UTF-8 plain text standard format so it can be used later on in the annotation tool. Afterwards, the corpus collected was automatically translated from English to Arabic using the Google Translate API paid service.[3]

4. Development of the Guidelines

To obtain a consistent post-edited version of MT errors by the various annotators, clear and concise correction guidelines are needed. In order to annotate the MT corpus, we use the general annotation correction guidelines we created previously for L1 described in Zaghouani et al. (2014a; Zaghouani et al. (2014b) and we add specific MT post-editing correction rules. In the general correction guidelines we place the errors to be corrected into the following seven categories:

1. **Spelling errors:** mostly letter Yaa and hamza errors in the MT texts.
2. **Word choice errors:** a very frequent error in texts produced by MT systems.
3. **Morphology errors:** mostly related to an incorrect inflection or derivation.
4. **Syntactic errors:** the MT systems used in this project produced many cases of wrong gender and number agreement and also errors related to definiteness and wrong case and tense assignment.
5. **Proper names errors:** we observed many cases of named entities being improperly translated into Arabic.
6. **Dialectal usage errors:** the dialectal is generally not present in the MT texts.
7. **Punctuation errors:** in some cases punctuation signs appear in the wrong place.

We refer to Zaghouani et al. (2014b) for more details about these errors.

In the MT post-editing guidelines, we provide the annotators with detailed annotation procedure and explain how to deal with borderline cases in order to a increase the inter-annotator agreement. We include many annotated examples to illustrate some specific cases of machine translation correction rules as seen in the example of Figure 3 showing the original English text, Figure 4 showing a sample machine translation output and Figure 5 showing a sample human post-edited version of the sample.

Since there are equally-accurate alternative ways to edit the machine translation output, all being considered correct, some using fewer edits than others, therefore, the task can be very challenging.

We explained in the guidelines that the machine translated texts should be corrected with a minimum number of edits necessary to achieve an acceptable translation quality. However, correcting the accuracy errors and producing a semantically coherent text is more important than minimizing the number of edits and therefore, the annotators were asked to pay attention to the following three aspects: accuracy, fluency and style.

Accuracy The accuracy of the translation is very important and any missing translation should be added to ensure the coverage of the semantics of the source sentence. In all cases, the edited machine translation text should have the same meaning as the English source text. Word and phrases rearrangement over long distances is only permitted if it is really needed as the annotators are encouraged to use as much of the raw MT output as possible. Finally, the edited machine translation should not add any information that is not present in the English source text. In the same way, it should not omit any information or meaning that is in the English source text as seen in the following example in Table 1.[4]

Fluency Grammatical and agreement errors occur frequently in MT texts and they should be always corrected. Word reordering is only permitted when it is needed to correct the the syntax or meaning in the sentence as seen in the example of Table 2.

Style Since machine translation output is fully automatically generated, it has its own style that may seem unfamiliar or unnatural although, in some cases it has an acceptable word order and conveys the meaning of the English source text accurately. Therefore, in such cases the style of writing should not be modified or improved as can be seen in Table 3.

The MT post-editing annotation guidelines will be published as a technical report and will be available in the near future on the QALB project web page.[5]

5. Annotation Procedure

The annotation team consisted of a lead annotator and five annotators. The lead annotator is also the annotation workflow manager of this project. He evaluates the quality of the

[1] El Kholy and Habash (2012) present some of the major challenges for statistical machine translation from English into Arabic.

[2] https://en.wikinews.org

[3] https://cloud.google.com/translate

[4] Arabic transliteration is presented in the Habash-Soudi-Buckwalter scheme (Habash et al., 2007).

[5] http://nlp.qatar.cmu.edu/qalb/

Original	Japan has an extensive web of highways with thousands of tunnels.
Machine Translation	اليابان لديها **على** شبكة **الإنترنت** واسعة من الطرق السريعة مع الآلاف من الأنفاق. AlyAbAn ldyhA ς lý šbkħ AlIntrnt wAsςħ mn AlTrq Alsryςħ mς AlĀlAf mn AlÂnfAq.
Post-Edit	اليابان لديها شبكة واسعة من الطرق السريعة مع الآلاف من الأنفاق. AlyAbAn ldyhA šbkħ wAsςħ mn AlTrq Alsryςħ mς AlĀlAf mn AlÂnfAq.

Table 1: Example of accuracy errors. Words in bold not present in the original text

Original	Brazil's Syrians divided over unrest.
Machine Translation	السوريون منقسمون حول **البرازيل الاضطرابات** . Alswrywn mnqsmwn Hwl AlbrAzyl AlADTrAbAt.
Post-Edit	سوريو البرازيل منقسمون حول الاضطرابات. swryw AlbrAzyl mnqsmwn Hwl AlADTrAbAt.

Table 2: Example of fluency error. Words in bold have the wrong word order.

annotation, monitor and report on the annotation progress. A clearly defined protocol is set, including a routine for the post-editing annotation job assignment and the inter-annotator agreement evaluation. The lead annotator is also responsible of the corpus selection and normalization process beside the annotation of the gold standard data to be used to compute the Inter-Annotator Agreement (IAA) portion of the corpus.

The annotators in this project are university graduates with good Arabic language background, two of them are graduate students of translation and interpretation studies. To ensure the annotation quality, an extensive training phase for each annotator was conducted. Afterwards, the annotator's performance was closely monitored during the initial period, before allowing the annotator to join the official post-editing production phase. Moreover, a dedicated online discussion group was frequently used by the annotation team to keep track of the MT post-editing questions and issues raised during the annotation process. This mechanism, proved to help the annotators and the lead annotator to have a better communication.

The annotation itself is done using the QAWI annotation tool, an in house built web annotation framework designed originally for the manual correction of errors in L1 and L2 texts (Obeid et al., 2013).

This framework includes two major components: The annotation management interface which is used to assist the lead annotator in the general work-flow process, it allows the annotation manager easily upload and organize files and projects, manage users, assign files in a batch or individually, export annotation tasks and monitor the current annotation progress by processing real time annotation progress statistics. Moreover inter-annotator agreement (IAA), evaluation metrics such as the Word Error Rate (WER) are integrated with the management interface to allow the scores to be computed and the results to be stored over time.

The MT post-editing annotation interface is the actual annotation tool (Figure 1), which allows the annotators to do the manual correction of the MT Arabic output. The interface provides the following types of corrections:

1. **Word Edit**: to correct/modify a word.
2. **Word Move**: to move words to the right location in the sentence.
3. **Add Word**: insert missing words in the text.
4. **Delete**: delete unnecessary words.
5. **Merge and Split**: to merge or split words.

All post-editing action history previously mentioned are recorded in a database and can be exported to an XML file. Figure 2 shows an example of how the annotation actions are stored in the XML annotation export file.

Finally, and in order to increase the post-editing speed and prior to the first human pass, an automatic post-editing pass is done through MADAMIRA (Pasha et al., 2014), a tool that automatically corrects common spelling errors using a prediction model based on the words in-context. MADAMIRA uses a morphological analyzer to produce, for each input word, a list of analyses specifying every possible morphological interpretation of that word, covering all morphological features of the word. Most of the errors automatically corrected are related to Ya/Alif-Maqsura, Ha/Ta-Marbuta and Hamzated Alif forms, which are common spelling errors in Arabic.[6]

6. Evaluation

6.1. Inter-Annotator Agreement

We use Word error Rate (WER) as a proxy of the Inter annotator agreement. If the WER of two different annotations of the same sentences is low, we assume there is a high agreement between them. To evaluate the MT post-editing quality, we measure the inter-annotator agreement (IAA) on randomly selected files to ensure that the annotators are consistently following the annotation guidelines. A high annotation agreement is a good indicator of the data quality.

[6]For more information on Arabic orthography and other issues of Arabic NLP, see (Habash, 2010).

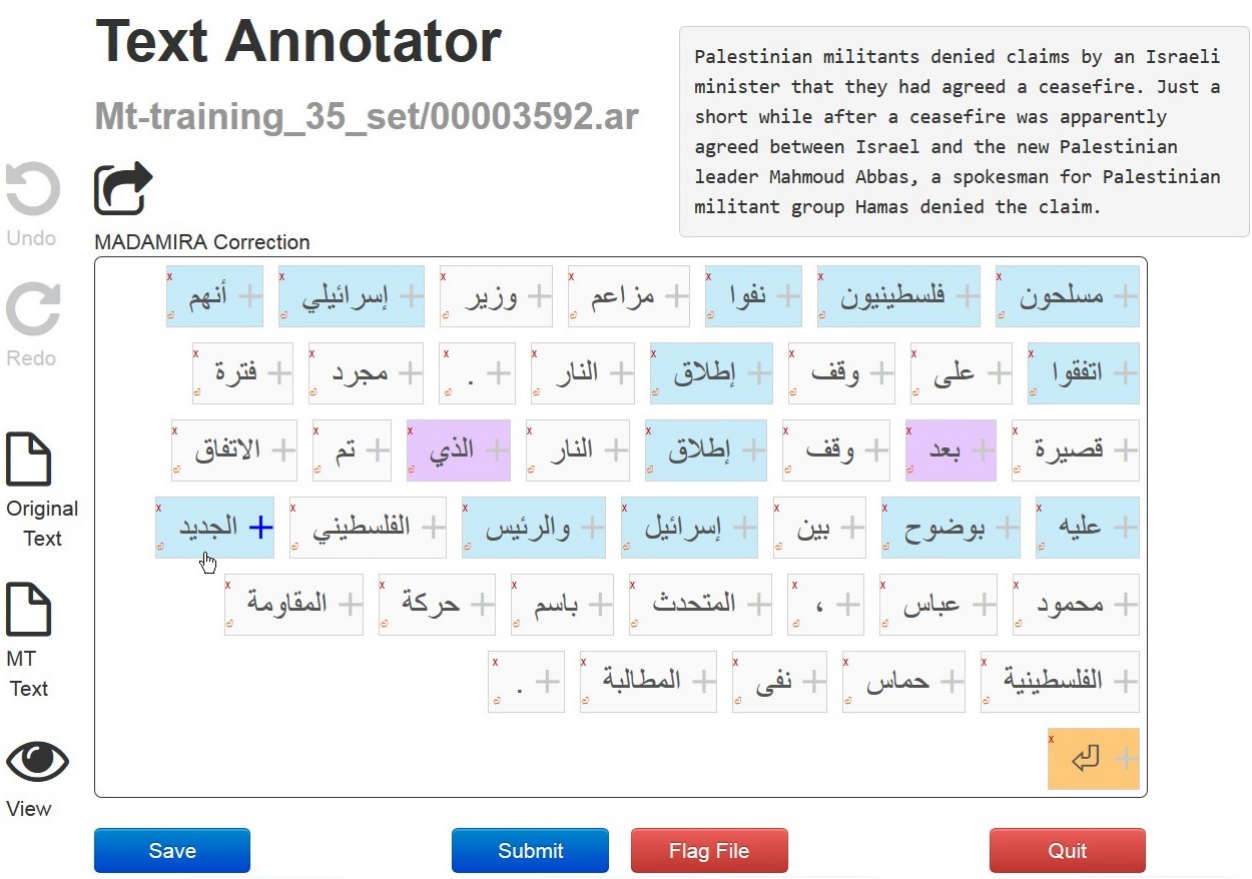

Figure 1: The MT Post-Editing annotation interface. Edited words are highlighted in blue. Inserted words are highlighted in purple.

```
<ACTION_HISTORY>
  <ACTION actionType="edit" annotatorID="23" newText="الأمريكي" passNum="1" tokenID="34" />
  <ACTION actionType="edit" annotatorID="23" newText="أنه" passNum="1" tokenID="37" />
  <ACTION actionType="edit" annotatorID="23" newText="الأسد" passNum="1" tokenID="46" />
  <ACTION actionType="edit" annotatorID="23" newText="وأضاف" passNum="1" tokenID="106" />
  <ACTION actionType="edit" annotatorID="23" newText="أن" passNum="1" tokenID="107" />
  <ACTION actionType="edit" annotatorID="23" newText="الأمن" passNum="1" tokenID="112" />
  <ACTION actionType="edit" annotatorID="23" newText="الأمريكي" passNum="1" tokenID="132" />
  <ACTION actionType="edit" annotatorID="49" newText="شخصاً" passNum="2" tokenID="6" />
  <ACTION actionType="edit" annotatorID="49" newText="وفقاً" passNum="2" tokenID="10" />
  <ACTION actionType="edit" annotatorID="49" newText="لجيش" passNum="2" tokenID="11" />
  <ACTION actionType="edit" annotatorID="49" newText="الولايات" passNum="2" tokenID="12" />
  <ACTION actionType="delete" annotatorID="49" passNum="2" tokenID="16" />
  <ACTION actionType="add_token_before" annotatorID="49" newTokenID="135" newTokenText="في" passNum="2"
  tokenID="21" />
  <ACTION actionType="delete" annotatorID="49" passNum="2" tokenID="25" />
  <ACTION actionType="edit" annotatorID="49" newText="الجثث" passNum="2" tokenID="24" />
  <ACTION actionType="add_token_before" annotatorID="49" newTokenID="136" newTokenText="التي" passNum="2"
  tokenID="30" />
  <ACTION actionType="move_before" annotatorID="49" passNum="2" targetTokenID="40" tokenID="37" />
  <ACTION actionType="edit" annotatorID="49" newText="عليها" passNum="2" tokenID="37" />
  <ACTION actionType="edit" annotatorID="49" newText="فهوة" passNum="2" tokenID="45" />
  <ACTION actionType="move_after" annotatorID="49" passNum="2" targetTokenID="49" tokenID="51" />
  <ACTION actionType="add_token_before" annotatorID="49" newTokenID="137" newTokenText="من" passNum="2"
  tokenID="57" />
  <ACTION actionType="edit" annotatorID="49" newText="والجو" passNum="2" tokenID="58" />
  <ACTION actionType="edit" annotatorID="49" newText="عندما" passNum="2" tokenID="60" />
  <ACTION actionType="delete" annotatorID="49" passNum="2" tokenID="59" />
</ACTION_HISTORY>
```

Figure 2: Extract of output file showing the correction action history.

Original	It's been five years since pro-democracy protests started.
Machine Translation	**انها كانت** خمس سنوات منذ أن بدأت الاحتجاجات المؤيدة للديمقراطية. AnhA kAnt xms snwAt mnð Ân bdÂt AlAHtjAjAt Almŵydħ lldymqrATyħ.
Post-Edit	لقد مرت خمس سنوات منذ أن بدأت الاحتجاجات المؤيدة للديمقراطية. lqd mrt xms snwAt mnð Ân bdÂt AlAHtjAjAt Almŵydħ lldymqrATyħ.

Table 3: Example of machine translation unnatural but acceptable style shown in the words in bold. No correction is needed in this case.

	Raw vs Gold	**IAA**$_{Round1}$	**IAA**$_{Round2}$
QALB L1 Corpus	24.45	3.80	N/A
QALB L2 Corpus	37.64	14.67	3.35
QALB MT Corpus	31.75	16.87	4.92

Table 4: Comparison between the MT corpus and the L1 and L2 corpus with the percentage of changes from the RAW output against the gold output and the inter-annotator agreement (IAA) on all 'words' in terms of average WER (Punctuation is ignored). Round1 is basic IAA comparing two annotations starting from raw output text. Round2 starts with the output of Round1.

Original	'Traditional museums are run by the old people.'
Machine Translation	يتم تشغيل المتاحف التقليدية التي كتبها كبار السن. ytm tšγyl AlmtAHf Altqlydyħ Alty ktbhA kbAr Alsn. 'Traditional museums are functioning that were written by old people.'
Annotator A (Good)	المتاحف التقليدية تشتغل من طرف كبار السن. AlmtAHf Altqlydyħ tštγl mn Trf kbAr Alsn. 'Traditional museums are working by old people.'
Annotator B (Acceptable)	كبار السن يديرون المتاحف التقليدية. kbAr Alsn ydyrwn AlmtAHf Altqlydyħ. 'Old people are directing traditional museums.'
Annotator C (Bad)	تدار المتاحف التقليدية من قبل كبار السن tdAr AlmtAHf Altqlydyħ mn qbl kbAr Alsn. 'Traditional museums are run by the old people.'

Table 5: Example of multiple post-editing corrections of an MT sentence

The IAA is measured over all pairs of annotations to compute the AWER (Average Word Error Rate). In this evaluation, the WER measures the post-editing errors against all words in the text, the lower the WER between two annotations, the higher is their agreement (Snover et al., 2006).

The IAA results shown in Table 4 include the results obtained in the current work as well as the results from the previous work described in Zaghouani et al. (2014b) for L1 corpus and in Zaghouani et al. (2015) for L2 corpus. We included the results from previous work to be able to compare IAA scores across the different genres.

The IAA results for the MT corpus are computed over 20 files (2,980 words) post-edited by at least three different annotators for the MT corpus and over 200 files (10,288 words) for the L1 corpus and finally 20 files (3,188 words) for the L2 corpus.

Table 4 shows the number of changes done over the whole corpus measured in WER between the raw text and the edited text. We observe that on average 31.75% of text was changed for the MT corpus. Secondly, we present the IAA numbers in terms of AWER in two evaluation rounds. In the first IAA round, the post-edited text is compared to a post-edited text made by a second pool of three annotators. The IAA of 16.87 obtained for the round 1 could be explained by the relatively high level of changes in the text and also by the difficult nature of the MT post-editing task in general. In order to measure the fluency agreement of the post-edited text, we performed a second round of IAA in which the output text of the first round was provided to second pool of three annotators in order to measure their agreement on the correction done during the first round of annotation in term of IAA. The low average WER of 4.92 obtained show a high agreement with the post-editing done in the first round between three annotators. The results obtained with the MT are comparable to those obtained with the L2 corpus, this can be explained by the difficult nature of both corpora and the multiple acceptable corrections for both.

Original

Hurricane Ismael was a weak, but deadly Pacific hurricane that killed over one hundred people in northern Mexico in September of the 1995 Pacific hurricane season.
It developed from a persistent area of deep convection on September 12, and steadily strengthened as it moved to the north-northwest. Ismael attained hurricane status on September 14 while located 210 miles (340 km) off the coast of Mexico. It continued to the north, and after passing a short distance east of Baja California it made landfall on Topolobampo in the state of Sinaloa with winds of 80 mph (130 km/h). Ismael rapidly weakened over land, and dissipated on September 16 over northwestern Mexico.
The remnants entered the United States and extended eastward into the Mid-Atlantic States. Offshore, Ismael produced waves of up to 30 feet (9 m) in height. Hundreds of fishermen were unprepared for the hurricane, which was expected to move more slowly, and as a result 52 ships were wrecked, killing 57 fishermen. On land, Ismael caused 59 deaths in mainland Mexico and resulted in $26 million in damage (1995 USD, $39.7 million 2012 USD). The hurricane destroyed thousands of houses, leaving 30,000 people homeless. Moisture from the storm extended into the United States, causing heavy rainfall and localized moderate damage in southeastern New Mexico.

Figure 3: Sample text in original English version.

Machine Translation

وكان الاعصار اسماعيل ضعيفة، ولكن القاتل الإعصار الذي قتل المحيط الهادئ أكثر من مائة شخص في شمال المكسيك في سبتمبر من موسم 1995 إعصار المحيط الهادئ. أنها وضعت من منطقة الحمل الحراري العميق استمرار في 12 سبتمبر، وعزز بشكل مطرد، حيث إنها انتقلت إلى الشمال والشمال الغربي. بلغ إعصار وضع إسماعيل في 14 سبتمبر بينما يقع 210 أميال (340 كيلومترا) قبالة سواحل المكسيك. استمرت في الشمال، وبعد اجتياز مسافة قصيرة شرق ولاية باجا كاليفورنيا من اليابسة على Topolobampo في ولاية سينالوا مع رياح 80 ميلا في الساعة (130 كم / ساعة).
اسماعيل ضعفت بسرعة على الأرض، وتبدد في 16 سبتمبر على شمال غرب المكسيك.
دخلت فلول الولايات المتحدة وامتد شرقا إلى الولايات منتصف الأطلسي. في الخارج، أنتجت موجات من إسماعيل يصل إلى 30 قدما (9 م) في الارتفاع. كان مئات الصيادين غير مستعدة للإعصار، الذي من المتوقع أن تتحرك ببطء أكثر، ونتيجة لذلك دمرت 52 سفينة، مقتل 57 صيادا. على الأرض، تسبب إسماعيل 59 حالة وفاة في المكسيك البر الرئيسي وأسفرت عن 26 مليون دولار في الضرر (1995 USD، 2012 $ 39700000 دولار أمريكي). دمر الاعصار الالاف من المنازل، وترك 30،000 شخص بلا مأوى. مدد الرطوبة من العاصفة إلى الولايات المتحدة، مما تسبب في هطول أمطار غزيرة وأضرارا متوسطة محلية في جنوب شرق نيو مكسيكو.

Figure 4: Machine translated version of the sample text.

6.2. Error Analysis

There will be always cases of MT post-editing disagreement, as there is often many ways to correct a given translation. With our guidelines, we try our best to reduce the inconsistency in the annotation. In Table 2, we show an example of disagreement among the annotators including a case of two acceptable corrections. For instance, Annotator C added the unnecessary word بينما bynmA ‘while’ which

Post-Edited


وكان الإعصار إسماعيل ضعيفا، ولكنه الإعصار القاتل الذي أودى بحياة أكثر من مائة شخص في شمال المكسيك في سبتمبر من موسم 1995 لأعاصير المحيط الهادئ. وقد تشكل من منطقة حمل حراري عميق متواصلة في 12 سبتمبر، واشتد بشكل مطرد حيث انتقل إلى الشمال والشمال الغربي.

بلغ إسماعيل درجة إعصار في 14 سبتمبر، حين كان على بعد 210 أميال (340 كيلومترا) قبالة سواحل المكسيك. وامتدّ ناحية الشمال، وبعد اجتياز مسافة قصيرة شرق باجا بولاية كاليفورنيا، وصل إلى اليابسة في طوبولوبامبو من ولاية سينالوا برياح تبلغ سرعتها 80 ميلا في الساعة (130 كم / ساعة). ضعف إسماعيل بسرعة على الأرض، وتبدد في 16 سبتمبر فوق شمال غرب المكسيك.

وصلت آثاره الولايات المتحدة وامتدت شرقا إلى ولايات منتصف الأطلسي. قبالة الساحل، أحدث إسماعيل أمواجا يصل ارتفاعها إلى 30 قدما (9 م).

كان مئات الصيادين غير مستعدين للإعصار الذي كان المتوقع له أن يتحرك ببطء أكثر، مما دمر 52 سفينة، وقتل 57 صيادا.

على اليابسة، تسبب إسماعيل في 59 حالة وفاة، في بر المكسيك الرئيسي، وأسفر عن 26 مليون دولار خسائر (39.7 مليون دولار USD، سنة 1995). دمر الإعصار الآلاف من المنازل، وخلّف 30،000 شخص بلا مأوى.

وامتدت الرطوبة بفعل العاصفة إلى الولايات المتحدة، مما تسبب في هطول أمطار غزيرة، ووقوع أضرار محلية متوسطة في جنوب شرق نيو مكسيكو.

Figure 5: Post-Edited version of the sample text.

was not present in the original sentence or the MT output, moreover, she kept the word اليورانيوم AlywrAnywm 'Uranium' wrongly present in the MT output. On the other hand, the annotator A produced the perfect translation while the annotator B produced an acceptable one.

7. Conclusions

We have presented in detail the methodology used to create a 100K-word English to Arabic MT manually post-edited corpus, including the development of the guidelines as well as the annotation procedure and the quality control procedure using frequent inter-annotator measures. The created guidelines will be made publicly available and we look forward to distribute the post-edited corpus in a planned shared task on automatic error correction and getting feedback from the community on its usefulness as it was in the previous shared tasks we organized for the L1 and L2 corpus (Mohit et al., 2014; Rozovskaya et al., 2015).

We believe that this corpus will be valuable to advance research efforts in the machine translation area since manually annotated data is often needed by the MT community. We believe that our methodology for guideline development and annotation consistency checking can be applied in other projects and other languages as well. In the future, we plan to increase the size of the corpus and also to add other corpus domains.

Acknowledgements

We would like to thank the anonymous reviewers for their valuable comments and suggestions. We also thank all our dedicated annotators: Noor Alzeer, Hoda Fathy, Hoda Ibrahim, Anissa Jrad and Jihene Wafi. This publication was made possible by grants NPRP-4-1058-1-168 from the Qatar National Research Fund (a member of the Qatar Foundation). The statements made herein are solely the responsibility of the authors.

8. References

Avramidis, E., Burchardt, A., Hunsicker, S., Popovic, M., Tscherwinka, C., Torres, D. V., and Uszkoreit, H. (2014). The taraxu corpus of human-annotated machine translations. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, pages 2679–2682. European Language Resources Association (ELRA), 5.

Bouamor, H., Alshikhabobakr, H., Mohit, B., and Oflazer, K. (2014). A human judgement corpus and a metric for arabic MT evaluation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 207–213.

El Kholy, A. and Habash, N. (2012). Orthographic and morphological processing for english–arabic statistical machine translation. *Machine Translation*, 26(1-2):25–45.

Elming, J. (2006). Transformation-based corrections of rule-based mt. In *Proceedings of the EAMT 11th Annual Conference*.

Fishel, M., Bojar, O., and Popovic, M. (2012). Terra: a collection of translation error-annotated corpora. In *Proceedings of the 8th International Conference on Language Resources and Evaluation. International Conference on Language Resources and Evaluation (LREC-12), 8th, May 23-25, Istanbul, Turkey*. European Language Resources Association (ELRA), 5.

Habash, N., Soudi, A., and Buckwalter, T. (2007). On Arabic Transliteration. In A. van den Bosch et al., editors, *Arabic Computational Morphology: Knowledge-based and Empirical Methods*. Springer.

Habash, N. (2010). *Introduction to Arabic Natural Language Processing*. Morgan & Claypool Publishers.

Mohit, B., Rozovskaya, A., Habash, N., Zaghouani, W., and Obeid, O. (2014). The first qalb shared task on automatic text correction for arabic. In *Proceedings of the EMNLP Workshop on Arabic Natural Language Processing*, page 39.

Obeid, O., Zaghouani, W., Mohit, B., Habash, N., Oflazer, K., and Tomeh, N. (2013). A Web-based Annotation Framework For Large-Scale Text Correction. In *The Companion Volume of the Proceedings of IJCNLP 2013: System Demonstrations*, Nagoya, Japan, October.

Pasha, A., Al-Badrashiny, M., Kholy, A. E., Eskander, R., Diab, M., Habash, N., Pooleery, M., Rambow, O., and Roth, R. (2014). MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, Reykjavik, Iceland.

Popescu-belis, A., King, M., and Benantar, H. (2002). Towards a corpus of corrected human translations.

Rozovskaya, A., Bouamor, H., Habash, N., Zaghouani, W., Obeid, O., and Mohit, B. (2015). The second qalb shared task on automatic text correction for arabic. In *Proceedings of the ACL-IJCNLP Workshop on Arabic Natural Language Processing*, page 26.

Simard, M., Goutte, C., and Isabelle, P. (2007). Statistical phrase-based post-editing. In *Proceedings of the North American Association for Computational Linguistics: Human Language Technologies Conference*.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of AMTA*, pages 223–231.

Wisniewski, G., Kubler, N., and Yvon, F. (2014). A corpus of machine translation errors extracted from translation students exercises. In *International Conference on Language Resources and Evaluation (LREC 2014)*, page 4 pages, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Zaghouani, W., Habash, N., and Mohit, B. (2014a). The qatar arabic language bank guidelines. *Technical Report CMU-CS-QTR-124, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, September.*

Zaghouani, W., Mohit, B., Habash, N., Obeid, O., Tomeh, N., Rozovskaya, A., Farra, N., Alkuhlani, S., and Oflazer, K. (2014b). Large scale arabic error annotation: Guidelines and framework. In *International Conference on Language Resources and Evaluation (LREC 2014)*.

Zaghouani, W., Habash, N., Bouamor, H., Rozovskaya, A., Mohit, B., Heider, A., and Oflazer, K. (2015). Correction annotation for non-native arabic texts: Guidelines and corpus. In *Proceedings of the Association for Computational Linguistics Fourth Linguistic Annotation Workshop*, pages 129–139.