AlphaMWE-Arabic: Arabic Edition of Multilingual Parallel Corpora with Multiword Expression Annotations

Najet Hadj Mohamed [1,2*] Malak Rassem [3*] Lifeng Han [4] Goran Nenadic [4]
[1] University of Tours, France
[2] Arabic Natural Language Processing Research Group, University of Sfax, Tunisia
[3] Institute for Natural Language Processing (IMS), University of Stuttgart, Germany
[4] Department of Computer Science, University of Manchester, United Kingdom
* *co-first authors*

Abstract

Multiword Expressions (MWEs) have been a bottleneck for Natural Language Understanding (NLU) and Natural Language Generation (NLG) tasks due to their idiomaticity, ambiguity, and non-compositionality. Bilingual parallel corpora introducing MWE annotations are very scarce which set another challenge for current Natural Language Processing (NLP) systems, especially in a multilingual setting. This work presents AlphaMWE-Arabic, an Arabic edition of the AlphaMWE parallel corpus with MWE annotations. We introduce how we created this corpus including machine translation (MT), post-editing, and annotations for both standard and dialectal varieties, i.e. Tunisian and Egyptian Arabic. We analyse the MT errors when they meet MWEs-related content, both quantitatively using the human-in-the-loop metric HOPE and qualitatively. We report the current state-of-the-art MT systems are far from reaching human parity performances. We expect our bilingual English-Arabic corpus will be an asset for multilingual research on MWEs such as translation and localisation, as well as for monolingual settings including the study of Arabic-specific lexicography and phrasal verbs on MWEs. Our corpus and experimental data are available at https://github.com/aaronlifenghan/AlphaMWE

1 Introduction

Multiword Expressions (MWEs), such as "a cheap shot" (a cruel verbal attack) or "take it with a grain of salt" (regard something as exaggerated), are combinations of words that function as a single unit and have a specific meaning (Baldwin and Kim, 2010), typically regarded as a *'pain in the neck'* to Natural Language Processing (NLP) tasks, particularly in the field of machine translation (MT) (Sag et al., 2002) and information extraction (Kovačević et al., 2013; Maldonado et al., 2017). Translating MWEs accurately poses a significant challenge for statistical and neural MT systems (Han, 2022b; Han et al., 2021, 2020b). The difficulty lies in the idiomatic, colloquial or culture-specific nature of MWEs, which requires a deep understanding of their meaning, context, and cultural references (Moreau et al., 2018). Additionally, MWEs can be interpreted into multiple possible meanings, further complicating their translation. Therefore, a parallel corpus that incorporates MWE annotation is expected to be useful for improving the MT quality via system fine-tuning and error analysis. However, Arabic seems to lack a satisfactory corpus for such use. The literature describes several English–Arabic parallel corpora. However, to the best of our knowledge, none of these corpora includes the MWEs annotation.

In this paper, we describe our ongoing effort to extend AlphaMWE coordinated by Han et al. (2020a), a multilingual parallel corpus with annotation of MWEs, to the Arabic language including both standard and dialectal ones, i.e. the Egyptian and Tunisian Arabic. Arabic is a morphologically rich language and has been challenging for state-of-the-art MT systems (MILAD, 2022). Fol-

Proceedings of Recent Advances in Natural Language Processing, pages 448–457
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_050

lowing AlphaMWE, our study primarily focused on Verbal MWEs (VMWEs). A VMWE is defined as a MWE whose canonical form has a verb as its syntactic head (Markantonatou et al., 2017; Ramisch et al., 2018, 2020) with popular examples "kick the bucket", "take ... for granted", and "swallow someone's pride". We used state-of-the-art MT engines to facilitate the standard Arabic corpus creation and we will discuss the pros and cons of different MT models on MWE-related translation errors. We carried out manual post-editing and annotations by native Arabic speakers for this. Regarding dialectal Arabic corpus, we translated them from English from scratch, since the current MT models do not cover dialectal Arabic translations and the quality from MT output is too low to be useful, which also indicated the value of our corpus creation. Overall, in this work, we not only contribute to a series of parallel corpus on English-Arabic with MWE annotations but also give qualitative and quantitative analysis on MT errors facing MWEs, which we hope will be valuable for future MT research on this language pair.

The rest of this paper is organised as follows: Section 2 describes previous work dedicated to parallel Arabic corpora and compares our contribution to the state of the art. Section 3 is a brief introduction to the Arabic language. Section 4 explains the construction of the AlphaMWE-Arabic corpus and the qualitative evaluation using examples from MT outputs. In Section 5, we offer more quantitative and statistical analyses of the data annotation process using the human-in-the-loop metric HOPE (Gladkoff and Han, 2022). Finally, Section 6 concludes our paper and discusses perspectives for future work.

2 Related Work

The development of machine translation for low-resource languages is a widely studied challenge in NLP (Ortega et al., 2021). Many efforts have been made to create effective MT models. To train these models, various parallel resources have been proposed.

Ziemski et al. (2016) created the United Nations Parallel Corpus, which consists of over 2 million words of parallel texts in 6 official languages, including English and Arabic. Another work that includes Arabic is the multilingual parallel corpus MultiUN (Chen and Eisele, 2012). It extends the United Nations Parallel Corpus by including texts from various sources such as the United Nations and other international organisations.

In addition, several researchers have undertaken efforts to construct resources for Arabic dialects. Boujelbane et al. (2013) built a bilingual dictionary that utilised explicit knowledge about the relationship between Tunisian Arabic and Modern Standard Arabic (MSA). Wael and Nizar (2012) translated dialectal Arabic to MSA as a bridge to translate to English. Bouamor et al. (2014) created a multi-dialectal Arabic parallel corpus that contains 2000 sentences in MSA, Egyptian, Tunisian, Jordanian, Palestinian, and Syrian Arabic.

However, this previous work mainly focused on the creation of lexical and grammatical parallel resources using either manual or automatic methods, without *annotation* of linguistic phenomena such as MWEs.

To address this, there have been numerous studies aiming at creating monolingual corpora annotated with verbal MWEs, such as the PARSEME shared task corpora (Ramisch et al., 2020). PARSEME is a multilingual initiative that targets the parsing of MWEs in over 26 different languages, including MSA (Hadj Mohamed et al., 2022), but they are not parallel data. To extend this effort, AlphaMWE (Han et al., 2020a) not only focuses on the creation of *multilingual parallel* corpora but also incorporates the *annotation of MWEs* in both the source and target languages. So far, 4 languages are covered in AlphaMWE, namely English, Chinese, Polish, and German. However, as we discussed earlier, there is a lack of such parallel corpora for Arabic, even though it is one of the most spoken and used languages. In this work, we develop an Arabic edition of AlphaMWE including both the standard language and dialectal varieties.

3 On Arabic Language

The term "Arabic language" today can refer to either Modern Standard Arabic (MSA) or various spoken vernaculars referred to as Arabic dialects. The classical form of MSA is used in religious texts, poetry, and formal writing, whereas the dialectal form is used in everyday and colloquial conversation. We give in this section a brief overview of MSA specificities.

Firstly, in MSA, there are no capital letters and the use of punctuation marks is not widely adopted in current Arabic texts (at least not reg-

ularly). Secondly, Arabic tends to use long and complex sentences with *right-to-left* writing, making it common to find an entire paragraph without any punctuation. Thirdly, as a Semitic language, Arabic has a complex morphology. Indeed, it uses *concatenative morphology (agglutinated or compound words)*, where words are formed via a sequential concatenation process[1]. For example, the sentence *'then they will write it'* is presented in Arabic as one word فسيكتبونها. In addition, the Arabic language has some words that can add diacritical marks on top or below them to form new words that have new pronunciations and meanings, of which the new pronunciation is similar to the ones from the original word. As a result, texts without diacritical marks are highly susceptible to ambiguity. For instance, the word/symbol علم (pronunciation: Alam) can be diacritised in 9 different forms (Maamouri et al., 2006) including عِلم ("science", pronunciation: Elm), عَلَم ("flag", pronunciation: Alam), and عَلَّم ("he taught", pronunciation: Ellem), etc. Finally, another special aspect of Arabic is its flexible word order, where the rearrangement of certain words in a sentence does not affect its meaning. This is because the language uses case markers, particles, and other linguistic tools to clarify the connections between words, resulting in a more flexible syntax compared to languages with a more rigid word order. For example, *"the boy went to the school"* can be written in Arabic in three forms: الولد ذهب إلى المدرسة (the boy went to the school), ذهب الولد إلى المدرسة (went the boy to the school), and إلى المدرسة ذهب الولد (to the school went the boy). These unique features make Arabic a challenging language for NLP tasks.

4 AlphaMWE-Arabic

Following AlphaMWE (Han et al., 2020a), we used the PARSEME corpus for English as the source language. The PARSEME corpus is well established and provides a clear process of tagging and categorisation. The English corpus used in the PARSEME shared task was created by Walsh et al. (2018), where 832 VMWEs were manually annotated across 7,437 sentences taken from various topics and domains, such as news, literature, and IT documents[2]. Overall there are around 750 sentences extracted from the source PARSEME English corpus that include VMWE tags by AlphaMWE[3]. Furthermore, AlphaMWE divided these 750 sentences into 5 portions (by files) with the same size, i.e. around 150 sentences each for cross-validation and system-tuning purposes. We followed this process for the creation of our three corpora: Modern Standard Arabic, Tunisian Arabic, and Egyptian Arabic. We will first introduce the workflow for creating standard Arabic MSA including the usage of MT; then we introduce the ones for the dialectal varieties.

4.1 AlphaMWE-MSA

For MSA, we translated the English source using a MT system in the loop of our process. We favoured the use of the "MT plus post-editing (MT+PE)" as the preferred option, rather than translating from scratch via native speakers. Henceforth, the translation process is more efficient and the creation of the Arabic corpus was made more easily. This, in turn, allowed us to quantitatively evaluate the results and then finally, post-edit the output to obtain our human gold standard. This pipeline will be further elaborated in the next subsections. Our MSA corpus yielded 2,700 tokens. Our two native Arabic speakers who carried out the post-editing work include one Masters student from Egypt and one Ph.D. student from Tunisia both studying NLP abroad for their degrees as fluent English speakers. Following the AlphaMWE creation workflow (Han et al., 2020a), the post-editing was cross-validated by having them double-check on each other's first edit edition. The amount of annotation, translation, and evaluation work measured by time is around 15+ hours each.

4.1.1 MT Systems Comparison

We compared different MT systems on the English-to-Arabic translation including GoogleMT (Vaswani et al., 2017; Johnson et al., 2017) and Systran Translate[4]. We give some examples of our comparisons in Figure 4.1.1, where we used the colours green, red, yellow, and magenta to indicate that categories of well-translated,

[1] Agglutination is the process, common in Arabic, of adjoining clitics from simple word forms to create more complex forms.

[2] https://gitlab.com/parseme/parseme_corpus_en PARSEME English corpus

[3] https://github.com/poethan/AlphaMWE It includes parallel Chinese, German, and Polish ↔ English

[4] https://www.systran.net/en/translate/

wrong, correct but unnatural and skipped. We qualitatively evaluated the translation samples and from which, we have the following findings:

- 1) when Systran MT output makes mistakes, the errors are very severe, such as adding context out of the blue, while GoogleMT's output still makes some sense when it is wrong. For instance, in sentence 2 (Figure 4.1.1), the phrase "jerked the paper out of view" was translated by Systran MT into a completely different context أزاغ الورقة خجلاً (azāgh al-wārakah khajalan / lit. 'deflected the paper shyness') 'he deflected the paper with shyness'.

- 2) Systran has more correct translations on entities. For example, the word *"copyright"* in sentence 5 (Figure 4.1.1) is correctly translated by Systran MT to حقوق النشر و التأليف while Google MT translated it as حقوق المؤلف ("the right of the author"). Although Systran MT performs reasonably well on some translations, as shown in the previous example, Google MT still performs better in terms of semantic accuracy overall.

Our thought is that we want to reduce the workload for the professional post-editing step, and we are keen to know more about how MT makes errors and mistakes when translating MWEs and verbal idioms. Therefore, we choose GoogleMT as our engine with the following rationale: a) entity errors can be fixed more easily than out-of-the-blue errors; b) we can get more examples of how MT fails in translating MWE-related content and these examples can be valuable for future research such as on guiding MT development.

4.1.2 Workflow Examples

We illustrate our workflow process with the following example sentence (Sentence 1). Firstly, we carried out the automatic translation for the Arabic target direction using Google Translate (output in Sentence 2). Then, we post-edited the output with annotation of the relevant target side VMWEs that are in line with the source English ones as shown in the example (Sentence 3). Finally, we evaluated the Google translation quality using the HOPE metric (Gladkoff and Han, 2022). The HOPE methodology is used to assess the quality of the Google Translation, taking into account expert post-editing annotations and a scoring model that assigns error penalty points based on error severity and category (Charalampidou and Gladkoff, 2022). In our example, Google Translate was unable to preserve the idiom of the original statement. As a result, the sentence's idiomatic meaning is lost in some cases.

(1) (*Source*)
But she did not **give me any time of day**.
lit.[5] 'But she did not pay me any attention' || 'she ignored me'

(2) (*Google Translation*)

لي	تحدد	لم	لكن-ها
lī	taḥaded	lm	lakenn-hā
for-me	pick.3.FEM.PAST	not	but-her

ال-يوم	من	ال-وقت
al-līūm	mn	al-wāqt
the-day	of	the-time

lit. But she did not pick for me the time of the day

(3) (*Human Gold Standard*)

إهتمام	أي	**تعيرني**	لم	لكنها
ihtemām	aī	**ta'īrnī**	lm	lakenn-hā
attention	any	pay.3.FEM.PAST	not	but-her

lit. 'But she did not pay me any attention' || 'she ignored me'

The different types of MT errors in HOPE are described in (Gladkoff and Han, 2022) as follows:
Mistranslation (MIS): Translation distorts the meaning of the source, and presents mistranslation or accuracy error.
Style (STL): Translation has poor style, but is not necessarily ungrammatical or formally incorrect.
Terminology (TRM): Incorrect terminology, inconsistency of translation of entities (forms, sections, etc.)
Impact (IMP): The translation falls short in clearly conveying the intended message (even if it may be accurate word-for-word, a good translation should not rely solely on literal equivalence and should have a clear expression of the central idea).
Missing Required Adaptation (RAM): The source has errors and needs correction, or requires significant adaptations for the target but the translator failed to make these changes.
Ungrammatical (UGR): Translation is ungrammatical - needs to be fixed to convey the meaning

[5]We follow the PARSEME corpus format, using 'lit.' as starting mark followed by literal translation and meaning equivalent translation.

	Source	*Google MT*	*Source*	*SysTran MT*
1	This was generally a grim place, filled with dust and people with nowhere to go, but now, with the rush hour at full force, it had been taken over by men and women with briefcases, books, and newspapers.	كان هذا مكانًا كئيبًا بشكل عام ، مليئًا بالغبار والناس ليس لديهم مكان يذهبون إليه ، ولكن الآن ، مع ساعة الذروة بكامل قوتها ، استولى عليها الرجال والنساء مع حقائبهم وكتبهم وصحفهم.	This was generally a grim place, filled with dust and people with nowhere to go, but now, with the rush hour at full force, it had been taken over by men and women with briefcases, books, and newspapers.	كان هذا بشكل عام مكانا قاتما، مليئا بالغبار والناس الذين ليس لديهم مكان يذهبون إليه، ولكن الآن، مع ساعة الذروة بكل قوة، سيطر عليها الرجال والنساء حافظات الملفات، الكتب، والصحف.
2	He had made it to the third or fourth paragraph when the man turned slowly toward him, gave him a vicious stare, and **jerked the paper out of view**.	لقد وصل إلى الفقرة الثالثة أو الرابعة عندما استدار الرجل ببطء تجاهه ، وأعطاه نظرة شريرة ، ونفض الورقة بعيدًا عن الأنظار.	He had made it to the third or fourth paragraph when the man turned slowly toward him, gave him a vicious stare, and jerked the paper out of view.	وكان قد وصل إلى الفقرة الثالثة أو الرابعة عندما التفت الرجل نحوه ببطء، ونظر اليه نظرة قاسية، وأزاغ الورقة خجلا.
3	The chair was comfortable, and the beer had gone slightly to his head.	كان الكرسي مريحًا ، وكانت الجعة قد انحرفت قليلاً إلى رأسه.	The chair was comfortable, and the beer had gone slightly to his head.	كان الكرسي مريحا، والبيرة وصلت قليلا إلى رأسه
4	It seemed to Quinn that Stillman's body had not been used for a long time and that all its functions had been relearned, so that motion had become a conscious process, each movement broken down into its component submovements, with the result that all flow and spontaneity had been lost.	بدا لكوين أن جسد ستيلمان لم يتم استخدامه لفترة طويلة وأن جميع وظائفه قد تم تعلمها مرة أخرى ، بحيث أصبحت هذه الحركة عملية واعية ، حيث تم تقسيم كل حركة إلى عناصرها الفرعية ، مما أدى إلى تدفق كل شيء وتلقائي. قد ضاعت.	It seemed to Quinn that Stillman's body had not been used for a long time and that all its functions had been relearned, so that motion had become a conscious process, each movement broken down into its component submovements, with the result that all flow and spontaneity had been lost.	ويبدو للوهلة الأولى أن جسد ستيلمان لم يستخدم منذ وقت طويل، وأن جميع وظائفه قد أعيدت، فأصبحت الحركة عملية واعية، وانقسمت كل حركة إلى حركات فرعية مكونة لها، مما أدى إلى فقد كل التدفق والعفوية.
5	Addressing the OAS, Amin had provoked laughter and applause among the delegates by saying that the hostages were as comfortable as they could be in the circumstances surrounded by explosives.	وكان أمين قد أثار الضحك والتصفيق بين المندوبين بقوله إن الرهائن كانوا مرتاحين بقدر ما يمكن أن يكونوا مرتاحين في الظروف التي تحيط بهم المتفجرات.	Addressing the OAS, Amin had provoked laughter and applause among the delegates by saying that the hostages were as comfortable as they could be in the circumstances surrounded by explosives.	وفي كلمته امام منظمة الدول الامريكية ، آثار امين الضحك والتصفيق بين المندوبين قائلا ان الرهائن كانوا مرتاحين بقدر ما يمكنهم في الظروف التي تحيط بها المتفجرات.
6	Copyright and the EU's principle of free competition should be taken into account in the televising of sports as of other events.	يجب أن يؤخذ حق المؤلف ومبدأ الاتحاد الأوروبي للمنافسة الحرة في الاعتبار في البث التلفزيوني للألعاب الرياضية كما هو الحال في الأحداث الأخرى.	Copyright and the EU's principle of free competition should be taken into account in the televising of sports as of other events.	وينبغي أن تؤخذ حقوق التأليف والنشر ومبدأ الاتحاد الأوروبي بشأن المنافسة الحرة في الاعتبار في البث التلفزيوني للرياضة كأحداث أخرى.

Figure 1: MT Output Comparisons between GoogleMT and Systran. Green: well translated, Red: wrong translation, Yellow: correct but unnatural and Magenta: skipped.

properly.

Proofreading Error (PRF): Linguistic error which does not affect the accuracy or meaning transfer, but needs to be fixed.

Proper Name (PRN): Named entity translation error.

We added two new error types to accommodate our post-editing and evaluation tasks on English-to-Arabic MT output:

MWE Missed Chance (MMC): Indicate when the MT output on source MWEs is either wrong semantically or correct translation but without using the corresponding correct MWEs in the target (in the situation when there is indeed such MWE in target).[6]

Skipped Word (SKP): Highlight when the MT system failed to translate a certain word that was important to the context.

In the scoring calculation of HOPE, there are score ranges from 0 to 16 (0, 1, 2, 4, 8, 16) indicating none, minor, medium, major, severe, and critical errors assigned to each error type. Then the overall penalty score of a segment or sentence (PSS) is used to classify the MT output into - correct (unchanged): PSS score 0, good enough: PSS score 1-to-4, or with major errors (requiring fixing): PSS score 5+.

Table 1 gives an example of evaluating these error types and their scores using our Source (1), MT output (2), towards the correct translation (3). In this example, the existing error types include MMC and IMP, with their severity level of 8 and 16 and the overall sentence level penalty point is 24. This indicates that the MT output sentence belongs to sentences with a Major error category.

In Section 5, we will report the statistical errors over all 150 segments.

4.2 Dialectal Arabic

As we previously mentioned, dialectal Arabic is used in everyday conversation, and with the explosion of social media it is inevitable that a great amount of the linguistic data digitally available is Dialectal. MWEs are also more prevalent in dialectal Arabic due to the idiomatic nature of informal speech. However, there are a few challenges with Dialectal Arabic. Firstly, it is not standardised, meaning there is no standard spelling which may incur multiple readable spelling variations of the same word. Secondly, very little work has been done on dialects in the context of MWEs. Thirdly, and perhaps most importantly, there is a large number of different dialects when it comes to Arabic. We focus in our work on both Egyptian and Tunisian Arabic.

Since there is no MT system that translates into Dialectal Arabic we opted to translate the source from scratch. Our Tunisian Arabic corpus contains 2,495 tokens and our Egyptian Arabic corpus contains 2,055 tokens.

[6]In the situation when there is no corresponding MWE in the target, we add the literal translation in place of no real MWE.

Error type	score	severity
MMC	8	Severe
MIS	0	None
STL	0	None
TRM	0	None
IMP	16	Critical
UGR	0	None
PRF	0	None
SKP	0	None
Sum	24	Major

Table 1: An Evaluation Example using Source Sentence (1) and MT Output Sentence (2) Toward the Correct Translation (3) using HOPE Metric (including each error type and overall sentence level)

5 Statistical MT Error Analysis

Evaluation of HOPE tasks can be carried out both with and without a final human-generated reference translation. Regardless, the evaluator assesses errors based on the HOPE quality metrics and assigns a score based on the severity of the errors using a penalty points system. In this task, we will generate a gold standard translation for the purpose of our open-source parallel corpus creation. Figure 2 shows the statistics from the HOPE metric on the 'aa' portion, one of the five files (*aa* to *ae*) included in the AlphaMWE corpus, on the percentage of MT output sentences that falls into 'un-changed (correct)', 'minor errors', and 'major errors'. The scoreboard shows that only 35% of MT output sentences are correct, and there are 44% and 21% of sentences having minor and major errors.

Table 2 shows more details and statistics of each error type from the HOPE metric evaluation. From Table 2, we can see that the largest ratio of error type is IMP, i.e. the "Impact" error. Then the

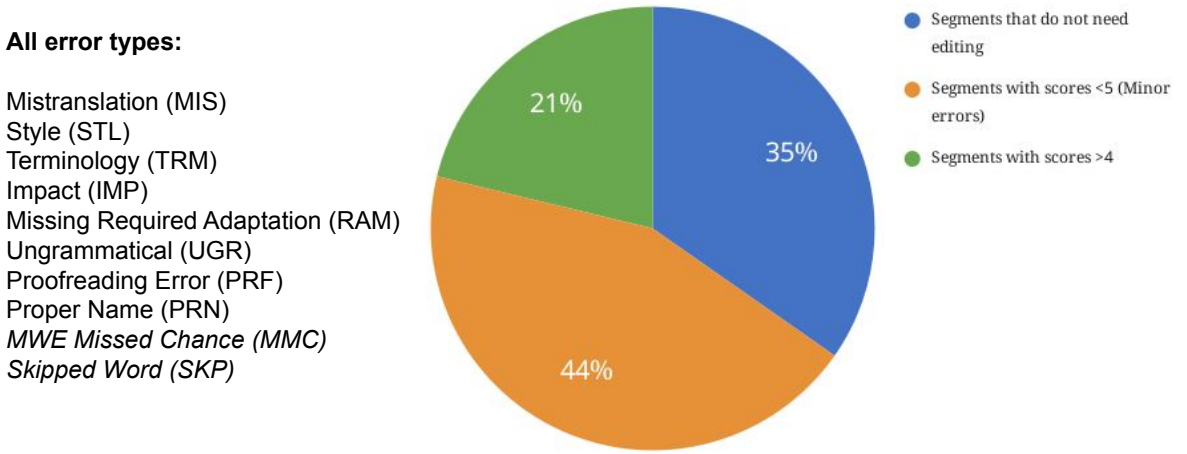

Figure 2: Evaluation Results using HOPE Metric on 150 Segments including Correct Translations (blue, PSS score = 0), Minor Errors (orange, 1 < PSS < 5), and Major Errors (green, PSS > 5).

Error type	MMC	MIS	STL	TRM	IMP	UGR	PRF	SKP	All	PPS
Total Penalty Scores	76	68	69	39	114	37	46	6	455	
Ratio out of total segments	17%	15%	15%	9%	25%	8%	10%	1%		3.03

Table 2: Penalty Score Ratios of Each Error Type and Average Penalty Scores of Each Segment from 150 Segments using HOPE Metric. The 'total penalty score' is the sum of all penalty score values from the same specific error type across all 150 segments. The 'Ratio out of total segments' values are calculated by the specific penalty scores divided by the All sum (455), except for the last value in the bottom right corner PPS (Penalty Points per Segment) which is calculated by all Penalty scores divided by all segment numbers, i.e. 455/150. The Average Penalty Point per Segment is 3.03 overall for all tested segments.

newly added error type MMC, i.e. "MWE Missed Chance", has 17% of all error weight. The next most common error types are followed by MIS, STL, and PRF representing "Mistranslation, Style, and Proofreading Error". On average, each segment received 3.03 penalty points.

6 Discussion and Future Work

To bridge the gap in the parallel corpus of English-Arabic with MWE annotations, we created AlphaMWE-Arabic, an Arabic edition of the AlphaMWE corpus. This is another step further to facilitate low-resource language processing including dialectal ones and can be useful to both multilingual and monolingual MWE-focused research.

During our creation, we introduced two new error types to the HOPE metric, and the experimental results show that MWE-related errors have a big ratio out of all error types. This reflects that the current state-of-the-art MT systems are still far from reaching human parity as they falsely claimed sometimes, which was partially due to their limited evaluation setting (Läubli et al., 2018; Graham et al., 2020; Han, 2022b).

For the standard Arabic corpus, we had two native speakers who carried out the post-editing and annotation. The corpus quality was ensured by cross-validation, i.e. having the second person check on the output from the other person's first edit. However, to quantitatively measure the inter-annotator agreement (Gladkoff et al., 2023) levels, in the future, we plan to design some experiments on calculating how much chance they agree with each other on the MT output quality and on the post-editing, e.g. to target MWEs vs non-MWEs.

Following the AlphaMWE open-source project, we plan to extend our corpus to a larger size and launch an open research project call where researchers can contribute and volunteer for the extension of the English-Arabic corpus. We have shared tasks in mind by contributing our corpus as a MWE-focused MT challenge, e.g. using human-

in-the-loop MT evaluation metric HilMeMe that looks into MWEs (Han, 2022a).

Limitations

In this work, we prepared a small-sized parallel corpus of English-Arabic with multiword expression (MWE) annotations, around 750 sentences directed from AlphaMWE (Han et al., 2020a). While we think this is an important step towards such kinds of resources, we do believe the size of our corpus can be enlarged via further development, such as recruiting volunteering professionals from translation backgrounds. Regarding dialectal Arabic, we offered Tunisian and Egyptian ones with the resources available. However, we can expect more dialectal Arabic to be added to this work if more native speakers are available. We used a human-in-the-loop metric HOPE to evaluate the GoogleMT output which gives a relatively transparent output on how many percents of the errors were made and how many percents of automatic translations fall into minor errors vs major errors. In a possible extensive investigation, we can apply more metrics to generate more diverse evaluation outputs, including fully automatic metrics.

Ethics Statement

There are no ethical issues with the work we carried out in this paper, including the corpus we created. Our Arabic corpus is translated from the source English one that has been validated and checked by the PARSEME shared task organisers and released publicly in 2018 (Ramisch et al., 2020, 2018).

Acknowledgement

We thank Haifa Alrdahi for her helpful advice and input on the Arabic language. We thank the PARSEME (PARSing and Multi-word Expressions) orgnisation https://typo.uni-konstanz.de/parseme/ for open research of their shared task corpus. We thank anonymous reviewers for their valuable comments. This work has been partially supported by grant EP/V047949/1 "Integrating hospital outpatient letters into the healthcare data space" (funder: UKRI/EPSRC).

References

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing, Second Edition*, pages 267--292. CRC Press, Taylor and Francis Group, Boca Raton, FL. ISBN 978-1420085921.

Houda Bouamor, Nizar Habash, and Kemal Oflazer. 2014. A multidialectal parallel corpus of arabic. In *LREC*, pages 1240--1245.

Rahma Boujelbane, Mariem Ellouze Khemekhem, and Lamia Belguith Hadrich. 2013. Mapping rules for building a Tunisian dialect lexicon and generating corpora. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*.

Parthena Charalampidou and Serge Gladkoff. 2022. A case of application of a new human mt quality evaluation metric in the emt classroom. In *New Trends in Translation and Technology (NeTTT) Conference*, pages 161 – 165.

Yu Chen and Andreas Eisele. 2012. Multiun v2: Un documents with multilingual alignments. In *LREC*, pages 2500--2504.

Serge Gladkoff and Lifeng Han. 2022. HOPE: A task-oriented and human-centric evaluation framework using professional post-editing towards more effective MT evaluation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 13--21, Marseille, France. European Language Resources Association.

Serge Gladkoff, Lifeng Han, and Goran Nenadic. 2023. Student's t-distribution: On measuring the inter-rater reliability when the observations are scarce.

Yvette Graham, Barry Haddow, and Philipp Koehn. 2020. Statistical power and translationese in machine translation evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 72--81, Online. Association for Computational Linguistics.

Najet Hadj Mohamed, Cherifa Ben Khelil, Agata Savary, Iskandar Keskes, Jean-Yves Antoine, and Lamia Hadrich-Belguith. 2022. Annotating verbal multiword expressions in Arabic: Assessing the validity of a multilingual annotation pro-

cedure. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1839--1848, Marseille, France. European Language Resources Association.

Lifeng Han. 2022a. Hilmeme: A human-in-the-loop machine translation evaluation metric looking into multi-word expressions.

Lifeng Han. 2022b. *An investigation into multi-word expressions in machine translation*. Ph.D. thesis, Dublin City University.

Lifeng Han, Gareth Jones, and Alan Smeaton. 2020a. AlphaMWE: Construction of multilingual parallel corpora with MWE annotations. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 44--57, online. Association for Computational Linguistics.

Lifeng Han, Gareth Jones, and Alan Smeaton. 2020b. MultiMWE: Building a multi-lingual multi-word expression (MWE) parallel corpora. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2970--2979, Marseille, France. European Language Resources Association.

Lifeng Han, Gareth Jones, Alan Smeaton, and Paolo Bolzoni. 2021. Chinese character decomposition for neural MT with multi-word expressions. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 336--344, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339--351.

Aleksandar Kovačević, Azad Dehghan, Michele Filannino, John A Keane, and Goran Nenadic. 2013. Combining rules and machine learning for extraction of temporal expressions and events from clinical narratives. *Journal of the American Medical Informatics Association*, 20(5):859--866.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791--4796, Brussels, Belgium. Association for Computational Linguistics.

Mohamed Maamouri, Seth Kulick, and Ann Bies. 2006. Diacritization: A challenge to arabic treebank annotation and parsing. In *Proceedings of the International Conference on the Challenge of Arabic for NLP/MT*, pages 35--47.

Alfredo Maldonado, Lifeng Han, Erwan Moreau, Ashjan Alsulaimani, Koel Dutta Chowdhury, Carl Vogel, and Qun Liu. 2017. Detection of verbal multi-word expressions via conditional random fields with syntactic dependency features and semantic re-ranking. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 114--120, Valencia, Spain. Association for Computational Linguistics.

Stella Markantonatou, Carlos Ramisch, Agata Savary, and Veronika Vincze, editors. 2017. *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*. Association for Computational Linguistics, Valencia, Spain.

KHALED MILAD. 2022. Comparative evaluation of translation memory (tm) and machine translation (mt) systems in translation between arabic and english. In *New Trends in Translation and Technology (NeTTT) Conference*, pages 142--151.

Erwan Moreau, Ashjan Alsulaimani, Alfredo Maldonado, Lifeng Han, Carl Vogel, and Koel Dutta Chowdhury. 2018. Semantic reranking of CRF label sequences for verbal multiword expression identification. In *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, pages 177 – 207. Language Science Press.

John Ortega, Atul Kr. Ojha, Katharina Kann, and Chao-Hong Liu, editors. 2021. *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT2021)*. Association for Machine Translation in the Americas, Virtual.

Carlos Ramisch, Silvio Ricardo Cordeiro, Agata Savary, Veronika Vincze, Verginica Barbu Mititelu, Archna Bhatia, Maja Buljan, Marie Candito, Polona Gantar, Voula Giouli, Tunga Güngör, Abdelati Hawwari, Uxoa Iñurrieta, Jolanta Kovalevskaitė, Simon Krek, Timm Lichte, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Behrang QasemiZadeh, Renata Ramisch, Nathan Schneider, Ivelina Stoyanova, Ashwini Vaidya, and Abigail Walsh. 2018. Edition 1.1 of the PARSEME shared task on automatic identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 222--240, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Carlos Ramisch, Agata Savary, Bruno Guillaume, Jakub Waszczuk, Marie Candito, Ashwini Vaidya, Verginica Barbu Mititelu, Archna Bhatia, Uxoa Iñurrieta, Voula Giouli, Tunga Güngör, Menghan Jiang, Timm Lichte, Chaya Liebeskind, Johanna Monti, Renata Ramisch, Sara Stymne, Abigail Walsh, and Hongzhi Xu. 2020. Edition 1.2 of the PARSEME shared task on semi-supervised identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 107--118, online. Association for Computational Linguistics.

Ivan A Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *Computational Linguistics and Intelligent Text Processing: Third International Conference, CICLing 2002 Mexico City, Mexico, February 17--23, 2002 Proceedings 3*, pages 1--15. Springer.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Conference on Neural Information Processing System*, pages 6000--6010.

Salloum Wael and Habash Nizar. 2012. Elissa: A Dialectal to Standard Arabic Machine Translation System. . In *In: Proc. 24th International. Conference on Computational Linguistics, COLING*.

Abigail Walsh, Claire Bonial, Kristina Geeraert, John Philip McCrae, Nathan Schneider, and Clarissa Somers. 2018. Constructing an annotated corpus of verbal mwes for english. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 193--200.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The united nations parallel corpus v1. 0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530--3534.