A Semi-supervised Approach for a Better Translation of Sentiment in Dialectical Arabic UGT

Hadeel Saadany
Centre for Translation Studies
University of Surrey
United Kingdom
hadeel.saadany@surrey.ac.uk

Constantin Orăsan
Centre for Translation Studies
University of Surrey
United Kingdom
c.orasan@surrey.ac.uk

Emad Mohamed
RGCL
University of Wolverhampton
Wolverhampton, UK
e.mohamed2@wlv.ac.uk

Ashraf Tantawy
School of Computer Science and Informatics
De Montfort University
Leicester, UK
ashraf.tantavy@dmu.ac.uk

Abstract

In the online world, Machine Translation (MT) systems are extensively used to translate User-Generated Text (UGT) such as reviews, tweets, and social media posts, where the main message is often the author's positive or negative attitude towards the topic of the text. However, MT systems still lack accuracy in some low-resource languages and sometimes make critical translation errors that completely flip the sentiment polarity of the target word or phrase and hence delivers a wrong affect message. This is particularly noticeable in texts that do not follow common lexico-grammatical standards such as the dialectical Arabic (DA) used on online platforms. In this research, we aim to improve the translation of sentiment in UGT written in the dialectical versions of the Arabic language to English. Given the scarcity of gold-standard parallel data for DA-EN in the UGT domain, we introduce a semi-supervised approach that exploits both monolingual and parallel data for training an NMT system initialised by a cross-lingual language model trained with supervised and unsupervised modeling objectives. We assess the accuracy of sentiment translation by our proposed system through a numerical 'sentiment-closeness' measure as well as human evaluation. We will show that our semi-supervised MT system can significantly help with correcting sentiment errors detected in the online translation of dialectical Arabic UGT.

1 Introduction

Incorporating automatic translation tools by websites such as Twitter, amazon.com and booking.com has become common practice to cater for their multilingual users. In this context, sentiment preservation is of great importance because decisions about purchasing a product or service, as well as analysis of public trends, are based on accurate translation of the user's affect message. Arabic UGT constitutes a significant challenge for MT systems because it is commonly a mix of Dialectical Arabic (DA) and Modern Standard Arabic (MSA) which differ significantly on the lexico-grammatical level. Research has shown that the code-switching between DA and MSA by online users can lead to a serious mistranslation of sentiment for several reasons (Saadany and Orasan, 2020).

First, there are lexical and structural differences between the two versions of the Arabic language which cause confusion to MT systems in choosing the correct sentiment-carrying word. On the lexical level, there are polysemous words used in both MSA and DA which can have exact opposite sentiment poles. To give one example, the word 'جامد' means 'rigid' in MSA, but in DA, within the UGT domain, it often means 'great or awesome'. Hence, we find the positive Goodreads review 'كتاب جامد جدا' (A very good book)[1] is mistranslated by the online MT tool into 'A very rigid book', incorrectly reflecting a negative sentiment. The same word, however, in another book review written in MSA – 'جامده جدا طريقه المؤلف في سرد الاحداث' – is correctly translated as 'The author's way of narrating events is very rigid', rightly reflecting the dissatisfaction of the author.

Second, the Arabic writing system does not have letters for short vowels; instead short vowels are realised as diacritic symbols on or below letters. UGT commonly lacks diacritics and hence it of-

[1] https://www.goodreads.com/book/show/16031620

Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP), pages 214 - 224
December 8, 2022 ©2022 Association for Computational Linguistics

ten contains words spelled alike in MSA and DA but different in meaning due to different pronunciation. An example of these homographs is in the DA tweet[2] ‘كفايانا نصب’ where the noun ‘نصب’ commonly means ‘fraud’ in DA with the diacritic ‘fatha’ (a short /a/ sound) on the first letter; the tweet should read ‘Enough of the fraud’. The online MT system flips the negative polarity as it mistakes this word with its common homograph in MSA meaning ‘monument’, pronounced with ‘damma’ (a short /u/ sound) on its first and second letters. The mistranslation of the homograph produces a neutral statement, ‘enough monument’, which completely misses the negative polarity of the source.

The third problem is that the way sentiment is expressed by the DA used in UGT is different than the structured DA data that is commonly used to train DA-EN NMT systems (e.g. Zbib et al. (2012); Bouamor et al. (2014); Elmahdy et al. (2014); Meftouh et al. (2015); Bouamor et al. (2018)). Some of the main differences is that UGT typically contains profanity and aggressive words that are not to be found in the available dialectical data. Moreover, the DA used on online platforms such as Twitter usually contains unusual orthography to express emotions or to obfuscate aggression and, at times, nuanced words that are understood only within context. A review of the literature shows that the authentic parallel datasets for DA-EN consist mainly of hand-crafted structured data which significantly differ from this type of noisy DA used in UGT. On the other hand, there is a considerable number of large parallel MSA-EN datasets in various domains (e.g OPUS[3] open-source parallel MSA-EN datasets include UN documents, TEDx talks, subtitles, news commentary, etc.). Since DA in the UGT domain has peculiar qualities and since it differs on the lexico-grammatical level from Standard Arabic and, at times, same words can have opposite sentiment in the two versions, the freely available MSA datasets are not optimal for translating sentiment in UGT written in a dialectical version.

Given the scarcity of any substantial gold-standard DA-EN data within the UGT domain, we propose to improve the transfer of sentiment in Arabic UGT by training a semi-supervised NMT system where we leverage the relatively large gold-standard MSA-EN data with DA monolingual data from the UGT domain. We take advantage of pre-training a cross-lingual language model with both a Masked Language Modelling (MLM) objective and a Translation Language Modelling (TLM) objective for creating a shared embedding space for English, MSA and DA. We show that initialising our NMT model with these cross-lingual pretrained word representations has a significant impact on the translation performance in general and on the transfer of sentiment in particular. In this research, therefore, we make the following contributions:

- We introduce a semi-supervised AR-EN NMT system trained on both parallel and monolingual data for a better translation of sentiment in Arabic UGT.
- We introduce an empirical evaluation method for assessing the transfer of sentiment between Arabic and English in the UGT domain.
- We make our compiled dataset, crosslingual language models and semi-supervised NMT system publicly available[4].

To present our contributions, the paper is divided as follows: Section 2 provides a summary of relevant approaches to supervised and unsupervised MT as well as research attempts for the translation of DA. Section 3 describes our semi-supervised NMT system set up and its requirements. Section 4 presents the experiments we conducted on our compiled datasets as well as the assessment methods used to evaluate the improvement of sentiment translation in DA UGT. Finally, Section 5 presents our conclusions on the different experiments and the limitations of the study.

2 Related Work

The earliest attempt to solve the problem of translating DA has been introduced by Zbib et al. (2012). They created the largest existing parallel data for DA to English which is relied upon in most MT research for DA. The dataset consists of around 250k parallel sentences. They used Mechanical Turk to translate sentences from DA to EN. Most of the DA is in the Levantine and Egyptian dialects, but none of the texts used belong to the UGT domain. They show that when translating the dialectical test sets, the DA-EN MT system performs 6.3

[2] https://twitter.com/Abdullahehemidy/status/221985043793444865, Accessed: Aug 2022

[3] https://opus.nlpl.eu/

[4] https://tinyurl.com/bdfh8e4m

and 7.0 BLEU points higher than an MT system trained on a 150M-word MSA-EN parallel corpus. Another approach to solve the data scarcity problem was introduced by Salloum and Habash (2013) who propose pivoting to MSA instead of directly translating from DA to EN. They transform DA sentences into MSA by a large number of hand-written morphosyntactic transfer rules.

There have been other attempts to create DA-EN and DA-MSA parallel datasets such as the multi-dialectical MDC and MADAR datasets (Bouamor et al., 2014, 2018), the QCA speech corpus (Elmahdy et al., 2014), and the PADIC parallel corpus which includes five dialects and MSA, but not English (Meftouh et al., 2015). These datasets, however, are relatively too small (max 14.7k parallel sentences) and differ considerably from the UGT domain. Since the problem of DA-EN scarcity of data still exists up to the time of writing this research, the most recent attempts to improve the translation of DA to English have focused either on augmenting the available datasets by bootstrapping techniques (Abid, 2020) or on training with the large available MSA datasets and fine-tuning on the smaller DA datasets (Sajjad et al., 2020).

A recent research line in MT which has been introduced to overcome the sparsity of gold-standard parallel data for low-resource languages is unsupervised MT which relies solely on monolingual data of the source and target languages in training (Lample et al., 2017, 2018; Artetxe et al., 2017). The key idea is to build a common latent space for two languages (or more) which can be used to reconstruct a sentence in a given language from a noisy version of it (Vincent et al., 2008), or to obtain the translated sentence by using a back-translation procedure (Sennrich et al., 2015a). The use of high quality cross-lingual word embeddings pretrained by state-of-the-art cross-lingual language models to initialise the unsupervised MT systems has recently contributed to a significant improvement in their performance (Lample and Conneau, 2019; Artetxe et al., 2019; Conneau et al., 2020). In this research, we combine both methods of supervised and unsupervised MT to compensate for the sparsity of the DA-EN data from the UGT domain. Our semi-supervised system is explained in the following section.

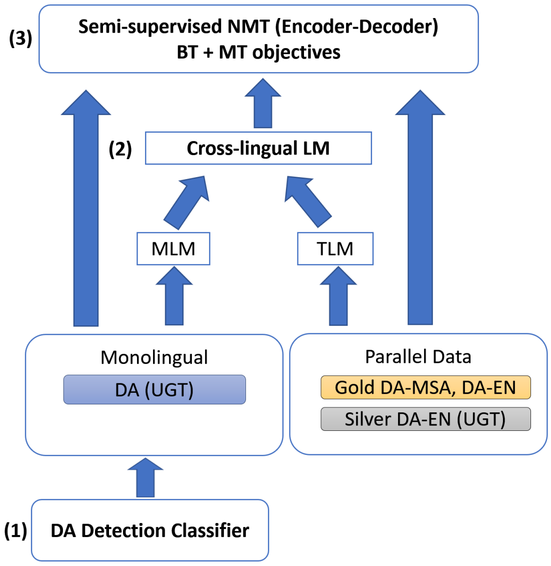

Figure 1: Semi-supervised NMT system

3 Semi-supervised NMT System Set Up

3.1 Cross-Lingual Language Model

Due to their lexico-grammatical differences, we treat dialectical and standard Arabic as two distinct languages. Hence, we construct a multi-directional NMT system between the permutations of DA-MSA-EN with the objective of obtaining the highest translation accuracy in the DA-EN direction. The setup of this system is shown in Figure 1. For constructing our semi-supervised NMT system we require the following data:

1. MSA-EN clean parallel data usually used for training NMT,
2. MSA-DA clean parallel data from any domain,
3. DA-EN silver-standard parallel data from the UGT domain with sentiment lexicon infused, and
4. DA monolingual data from the UGT domain.

It should be noted that the Arabic UGT is not written in DA per se, it is usually a mix of DA and MSA. Since we are treating DA and MSA as two distinct languages, we need to extract only the DA instances from the UGT dataset. For this purpose, we build our own DA detection classifier as per step (1) in Figure 1.

In step (2), we pretrain a cross-lingual language model to initialise our NMT system. We follow Lample and Conneau (2019) approach to train a cross-lingual language model with the combination of the following two objectives:

Masked Language Model (MLM): The MLM we train has a similar objective to BERT (Devlin et al., 2018) masking technique but adopting Lample and Conneau (2019)'s approach by including the use of text streams of an arbitrary number of sentences (truncated at 200 tokens) instead of pairs of sentences. We optimise the MLM objective on the MSA and EN source data as well as the DA monolingual data mentioned in data requirements 1, 2, and 4 above.

Translation Language Model (TLM): We use the TLM objective to improve cross-lingual training where the language model is trained on the gold-standard parallel sentences (i.e MSA-EN and MSA-DA in data requirements 1 and 2 above). The training is achieved by randomly masking words in both the source and target sentences. Thus, to predict a word masked in a DA sentence, for example, the model can either attend to surrounding DA words or to the EN/MSA side of the parallel data if the DA context is not sufficient to infer the masked DA word. By relying on the parallel data, the TLM objective helps in the alignment of embedding spaces across the three languages.

3.2 Semi-Supervised Machine Translation

To maximally exploit the similarity between the DA and MSA, we use the embeddings from the cross-lingual model we trained in step (2) of the experiment to initialise the encoder and decoder of the NMT system instead of random initialisation (step (3) in Figure 1). We train our system with both supervised and unsupervised NMT objectives. The unsupervised objective is achieved by a back-translation (BT) objective optimised by a round-trip translation of the UGT monolingual data. So a sentence s in DA monolingual data is translated to EN, and then back-translated with the objective of generating s. As for the supervised objective, we use the normal Machine Translation (MT) objective on our gold and silver parallel data. The compilation of the data requirements for our model is explained in the next section.

4 Experiment and Results

4.1 Data Compilation

As explained in Section 3.1, we need gold and silver standard parallel data for DA, MSA and EN as well as DA from the UGT domain. For the gold standard DA-EN data, we use the MDC (Bouamor et al., 2014) and the MADAR (Bouamor et al., 2018) corpora which consist of $\approx 33k$ parallel sentences where the DA side comprises Egyptian, Syrian, Palestinian, Jordanian and Tunisian dialects. Although this corpus has diverse dialects, it differs from the noisy DA used in UGT as it contains handcrafted sentences written for a traveler's guide. We, therefore, use two other gold DA-EN datasets that are closer to the UGT domain. The first is compiled by Abid (2020) consisting of 18k sentences created for the evaluation of the DA-EN translation by native speakers of Egyptian and Levantine dialects. The second is the Sentiment After Translation (SAT) corpus (Salameh et al., 2015) which consists of 1200 manually translated tweets from Levantine. The latter is the only gold-standard DA-EN UGT data we are aware of. As for the MSA-EN gold-standard data, we opt for diversifying the domain. Thus, we use 2M sentences from the Opus UN multilingual, and 1M from mixed Opus which is extracted from TEDx talks and subtitles (Tiedemann, 2012).

For the monolingual data, we compiled UGT datasets that were used as benchmarks for Arabic sentiment detection tasks to guarantee that they have a sentiment content. The monolingual datasets comprise tweets (Gamal et al., 2019), Goodreads reviews (LABR dataset) (Aly and Atiya, 2013) and the Arabic Online Commentary (AOC) (Zaidan and Callison-Burch, 2011). To extract the DA instances from these datasets, we build a DA detection classifier by fine-tuning a Roberta-XLM (Conneau et al., 2019) model on the Arabic Online Commentary (AOC) dataset (Zaidan and Callison-Burch, 2011). The AOC is composed of 3M MSA and dialectal comments created by extracting reader commentary from the online versions of three Arabic newspapers which have a high degree (about half) of dialectal content. From the 3M comments in this dataset, only 108,173 comments are labelled via crowdsourcing. We use the labelled comments for training our DA classifier. We randomly shuffle the labelled dataset and split it into 80% training (Train), 10% validation (Dev), and 10% test (Test). The accuracy of the model on the test set reached 92% which assured a satisfactory extraction of the DA instances from the monolingual dataset.

As for the silver-standard dataset, we have noticed that Google Translate, which is the ad hoc MT

Data Type	**Corpus**	**Domain**	**No. Sentences**
Gold MSA-EN	Multi-UN Mixed OPUS	UN Documents TEDx, Subtitles	2M 1M
Gold DA-MSA	MADAR MDC	Traveler's Guide	60K
Gold DA-EN	MADAR MDC (Abid, 2020)	Traveler's Guide Subtitles Wiki Fables	90K
Silver DA-MSA	AOC LABR SAT + NileULEX lexicon	(Back translation) Tweets Goodreads reviews Online comments	166K
Silver DA-EN	AOC LABR SAT + NileULex lexicon	(Automatic Translation) Tweets Goodreads reviews Online Comments	166K
Monolingual DA	AOC LABR SAT + NileULex lexicon	Tweets Goodreads reviews Online Comments	166K
Total Sentences			**3.648M**

Table 1: Distribution of the datasets used for training and their particular domains

system on different UGT platforms such as Twitter, translates English into standard Arabic. We leveraged this feature by translating our monolingual Arabic dialectical dataset into English and then back translated it into Arabic. This round-trip translation produced a synthetic parallel data of DA-EN-MSA. We expected that this synthetic dataset would contain a large number of mistranslated sentiment-carrying dialectical expressions and idioms that are commonly used in Arabic UGT. To alleviate the effect of these errors, we opted for correcting these DA expressions by infusing a lexicon of DA positive/negative phrases commonly used in UGT. For this purpose, we manually translated into MSA and English the NileULex (El-Beltagy, 2016) sentiment lexicon which consisted of DA phrases and idioms extracted from DA tweets. The lexicon consisted of 1000 positive and negative phrases that were found to be frequently used in tweets. We replaced these idioms with their correct translations in the MSA and EN side of the data. The sentence distribution of our datasets is shown in Table 1.

4.2 Training Details

4.3 Semi-supervised NMT system

Lample and Conneau (2019) have shown that the alignment of embedding spaces across languages that share the same alphabet and a significant fraction of vocabulary proves to be effective in cross-lingual tasks such as MT. Since this precisely applies to DA and MSA in our experiment, we use the synthesised and gold parallel datasets as well as the monolingual datasets described in the previous section to build the crosslingual language model for DA, MSA and EN. We use both the monolingual and the parallel data to train our model with a Translation objective (TLM) used in combination with a masking objective (MLM). Before training, the data is preprocessed by Moses tokeniser (Koehn et al., 2007). We use fastBPE[5] to learn BPE codes and split words into subword units. Since shared vocabulary has also proved to improve the performance of multilingual models on downstream cross-lingual tasks (Lample and Conneau, 2019; Conneau et al., 2020), we chose to have a shared subword vocabulary for all datasets. The BPE codes are learned on the concatenation of sentences sampled by applying a BPE model (Sennrich et al., 2015b) directly on raw text data for all languages. We apply the BPE coding on a network vocabulary size of 20000. We remove sentence pairs which contain empty lines or lines with a length longer than 200 tokens.

For training our cross-lingual model, we use a transformer architecture with 1024 hidden units, 8 heads, and a dropout rate of 0.1. We use the Adam optimiser (Kingma and Ba, 2014) for optimisation, a linear warm-up (Vaswani et al., 2017) and learning rates varying from 10^{-4} to 5.10^{-4}. For the MLM and TLM objectives, we use streams of 200 tokens and train on mini-batches of size 32. For the TLM objective, we also sample mini-batches of 32 tokens composed of sentences with similar lengths.

[5] https://github.com/glample/fastBPE

We use the averaged perplexity over languages as a stopping criterion for training the cross-lingual models.

We then use the pretrained embedding vectors in our crosslingual language model to initiate the semi-supervised NMT system trained on our gold and synthetic parallel datasets as well as the larger monolingual datasets. As explained in Section 3.2, the NMT system is trained with an MT objective for the three languages, DA-EN-MSA, simultaneously. We use the permutations of the three languages DA, EN, MSA taken two at a time. It is also trained with an unsupervised BT objective by maximising the back translation accuracy of the monolingual UGT dataset. For machine translation, we train on a 6 layer transformer and we increase the maximum token length to 200 to accommodate for MSA relatively long sentences. For the semi-supervised NMT system, we use the BLEU score of the DA-EN direction as the stopping criteria. We train for 100 epochs with an epoch size of 100k sentences. The training of the language model and the semi-supervised NMT system was conducted on 3 24GB GeForce RTX 3090 GPUs for a period of 9 days.

4.3.1 Baseline Models

We aimed to experiment with two alternative set ups where the monolingual UGT data is not included in training. The first is a supervised baseline model trained on the gold-standard MSA-EN and DA-EN datasets as well as the silver DA-EN dataset. We also concatenated our manually translated sentiment lexicon to the training data. For this baseline, DA and MSA are indiscriminately treated as one source language. We aimed to see how far concatenating DA and MSA data can improve the sentiment translation of DA into English in the UGT dataset. In the second set up, we followed similar research approaches (Salloum and Habash, 2013; Sajjad et al., 2020) which overcome the sparsity of DA data by pivoting to MSA as an intermediary step in the DA-EN MT pipeline. Thus, we build a DA-MSA MT system trained on the gold-standard DA-MSA datasets and then translated the MSA output into English. For translating into English, we used Marian open-source pretrained AR-EN MT model[6]. We call this latter model the Pivoting model. For both the baseline and the Pivoting model, we trained two NMT systems by replicating the same preprocessing technique of our semi-supervised model. Thus, we trained an unsupervised BPE encoding model for source and target data and split words into subword units. We set the maximum vocabulary size to 20000. The two models were trained using a transformer for both the encoding and decoding layers with 8 heads of self-attention and with an inner feed-forward layer of size 2048 and a batch size of 4096 sentences. We used the Adam optimiser with learning rate 2 and initialised training with 4000 warm up steps. We trained for 100k steps.

4.4 Results

For evaluation, we aimed to assess our proposed models' ability not only to produce quality translations but more importantly to transfer the UGT sentiment correctly from DA to EN. Therefore, we conducted different types of evaluation techniques on two test sets: a held-out DA-EN test set (180 parallel sentences) and a hand-crafted test set (50 sentences) selected from the monolingual DA dataset of tweets and book reviews. The hand-crafted dataset contained carefully chosen tweets and reviews with DA negative and positive expressions which constitute a challenge for available MT systems such as Google API (see Appendix A for some examples). A professional translator created a reference to the hand-crafted set. Both evaluation sets were translated by our baseline, the Pivoting model, the semi-supervised system proposed in this paper and Google Translate. We devised both human and automatic sentiment evaluation measures to assess how far the model is capable of maintaining the correct polarity of the source text for both test sets. The sacrebleu metric (Post, 2018) was also used as to assess whether the quality of the translation is balanced with the preservation of sentiment by our proposed models. Details of the experiment evaluations are presented in the next sections and examples of the semi-supervised DA-EN model output as compared to the ad hoc online MT tool for Twitter are included in Appendix A of this paper.

4.5 Translation Quality

Although there are benchmark datasets for the translation of DA into English (Bouamor et al., 2018; Meftouh et al., 2015; Sajjad et al., 2020), none belongs to the UGT domain. Accordingly, due to discrepancy in domain for our test data, we could not compare our results to any of these research experiments. We compare the BLEU scores

[6] https://nlp.johnsnowlabs.com/2021/01/03/translate_ar_en_xx.html

	SAM Score	**Average SAM Score**	**Human Evaluation**			**BLEU**
Model	**Test Set**	**Test Set**	**Hand-crafted Set**			**Test Set**
			H1	**H2**	**H3**	
Baseline	10.52	0.18	1.53	1.38	1.51	12.12
Pivoting MS-DA-EN	10.95	0.14	2.26	2.5	3	11.87
Google Translate	9.14	0.16	3.32	**3.28**	3.33	26.98
Semi-supervised MT	**5.26**	**0.10**	**4**	3.26	**4.35**	**32.29**

Table 2: Evaluation results for sentiment-closeness measure, human evaluation, and BLEU on test sets. The best scores are in bold.

of the held-out test set for outputs of the baseline, the Pivoting model, Google API and the semi-supervised MT model. As can be seen in Table 2, the BLEU score of the semi-supervised system is 5.31 points higher than Google Translate system and both the baseline and the Pivoting model fall far behind. This indicates that the quality of translation improves with our semi-supervised approach. However, despite the higher scores achieved by our system, research has shown that the BLEU metric may not be optimal for assessing how far the MT models transfer the sentiment correctly (Saadany and Orasan, 2021). The reason is that due to its restrictive exact matching to the reference, BLEU does not accommodate for importance n-gram weighting which may be essential in assessing sentiment-critical n-grams. For this reason, we conduct two types of sentiment-focused measures, automatic and manual, on our test sets. The sentiment assessment is explained in the next section.

4.5.1 Sentiment Quality

The first method is a Sentiment-Aware Measure (SAM) which evaluates the sentiment distance between the MT output (the hypothesis) and the reference translation in English. SAM is calculated by using the SentiWord dictionary of prior polarities (Gatti et al., 2015). SentiWord is a sentiment lexicon that combines the high precision of manual lexica and the high coverage of automatic ones (covering 155,000 words). It is based on assigning a 'prior polarity' score for each lemma-POS in both SentiWordNet and a number of human-annotated sentiment lexica (Baccianella et al., 2010; Warriner et al., 2013). The prior polarity is the out-of-context positive or negative score which a lemma-POS evokes.

We assume that SAM is proportional to the distance between the sentiment scores of the unmatched words in the system translation of the DA source and the reference in English, the higher the distance the greater the SAM score. To calculate the SAM score, we designate the number of remaining mismatched words in the hypothesis and reference translation by m and n, respectively. We calculate the total SentiWord sentiment score for the lemma-POS[7] of the mismatched words in the translation and reference sentences using a weighted average of the sentiment score of each mismatched lemma-POS. The weight of a hypothesis mismatched word w_h and a reference mismatched word w_r is calculated based on the sentiment score of its lemma-POS, s, as follows:

$$w_h^i = |s_i| \qquad i = 1, 2, \ldots, m. \tag{1}$$

$$w_r^i = |s_i| \qquad i = 1, 2, \ldots, n. \tag{2}$$

Then the total sentiment score for hypothesis S_h and reference S_r is given by:

$$S_h = \sum_{i=1}^{m} \alpha_i s_i, \quad \alpha_i = \frac{w_h^i}{\sum_{i=1}^{m} w_h^i} \tag{3}$$

$$S_r = \sum_{i=1}^{n} \beta_i s_i, \quad \beta_i = \frac{w_r^i}{\sum_{i=1}^{n} w_r^i} \tag{4}$$

The normalised SAM score is given by:

$$p = \frac{|S_r - S_h|}{2} \tag{5}$$

As seen from equation (5), SAM is interpreted as a translation cost. Thus, a lower SAM score indicates a shorter distance from the sentiment score of the source, and hence a better translation. As illustrated by Table 2, the semi-supervised NMT system maintains the lowest sentiment distance as it records the lowest total SAM score for the test set (5.26). Moreover, the average SAM score between the hypothesis of the semi-supervised model and

[7] We use spaCy V3.1 library to assign the lemma-POS of each token.

reference is also the lowest (0.10). Compared to the other models, the lower SAM scores indicate that the semi-supervised model is more capable of maintaining the sentiment polarity of the individual tokens of the source DA tweet or review as it shows the least sentiment discrepancy between its hypothesis and the reference translation.

For the second evaluation, we aimed to conduct a focused assessment of the ability of each model to transfer sentiment in challenging examples. We, therefore, conducted a human evaluation on the smaller hand-crafted dataset that consisted of UGT DA examples that constitute a challenge to online MT systems. We asked three native speakers of Arabic, who are also near native in English, to scale from 1 to 5 how far the sentiment expressed in the source DA tweet or the online review is preserved. We provided each human annotator with four translations of the source produced by the baseline, the Pivot model, the semi-supervised system and Google Translate. The average scores of the three annotators (H1, H2, H3) for each output is recorded in Table 2. As can be seen from the scores, the average performance of the semi-supervised model is slightly higher than Google Translate for Annotator H1 and H3, but lower for annotator H2. The baseline and the Pivoting model, however, are performing around 2 scales below the average according to all annotators. Overall, the automatic and manual sentiment evaluation of the four systems indicate that the semi-supervised MT system is more competent in preserving the sentiment of the source DA text.

4.6 Error Analysis

We conducted an error analysis on the mistranslation of sentiment by extracting the translations that received the lower scores by the human annotators in the hand-crafted dataset. It was observed that the aggressive DA examples in tweets were generally missed by Google API, the baseline as well as the Pivoting model. For example, the aggression in the DA tweet ‘يخرب بيتك يا سعد الدين’ (Go to hell Saadu-deen) is missed in the output of the Google API – ‘*your house will be destroyed, Saadu-deen*’ – as it provides a literal meaning to the DA offensive curse ‘يخرب بيتك’ (Go to hell). The semi-supervised model output, on the other hand, correctly transfers the offensive message as it translates the tweet with a similarly aggressive curse: ‘*Damn you Saadu-deen*’ (See Ex3 and Ex4 in Appendix A for similar aggressive tweets).

Moreover, the UGT monolingual data used for training the semi-supervised model had a positive effect in improving the translation of problematic structures such as negation particles which were realised as clitics added to the stem of the word. For example, the negation in the tweet ‘منصحش اي حد يشتريها’ (I would not advise anyone to buy it) is correctly transferred by the hypothesis of the semi-supervised model whereas Google API produces the wrong translation: ‘*I advise anyone to buy it*’, and the Pivoting model produces a similarly wrong meaning: ‘*Anybody buys it*’ (See also Ex3 in Appendix A). It was also noticed that the baseline performed well on structured DA-EN data but the translation quality was significantly degraded with the DA test data of tweets and online reviews. This substantiates our hypothesis that the available DA-EN structured data are not optimum for building a robust DA-EN system capable of translating the UGT domain.

Finally, it was noticed that are several examples where the sentiment gist of the source is transferred despite structural errors. For example, the human annotators marked the hypothesis of the semi-supervised model ‘*We are backwardness in us*’ as correctly transferring the negative sentiment despite the ill-formed structure. The correct reference of this tweet is ‘*Backwardness is in us*’. This trade-off between sentiment accuracy and translation fluency is evident in a number of hypotheses produced by the semi-supervised model (See Ex4, Ex5, Ex6 in Appendix A for similar examples).

5 Conclusion

In this research, we tackled the intricate problem of translating sentiment in different Arabic dialects in the UGT domain such as tweets and online reviews. We overcome the problem of the scarcity of gold-standard parallel data by training an NMT model with both a supervised and an unsupervised objective functions using monolingual as well as parallel data. We compared this model to a baseline that was trained solely on parallel data and a DA-EN MT model where we pivoted on MSA as an intermediary step. Our semi-supervised model showed improved performance over these two models not only in terms of translation quality but specifically in the preservation of the sentiment polarity of the source. We also conducted automatic and manual evaluation of the models' performance and proposed a lexicon-based metric that takes into account the sentiment distance between the source and the MT output. Overall, our error analysis has revealed that despite some structural inaccuracies the semi-supervised model is more capable of transferring the correct sentiment specifically in aggressive tweets. Future research will address the challenge of trading off translation fluency for sentiment accuracy to improve the translation of sentiment-oriented Arabic online content.

References

Wael Abid. 2020. The SADID Evaluation Datasets for Low-Resource Spoken Language Machine Translation of Arabic Dialects. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6030–6043.

Mohamed Aly and Amir Atiya. 2013. Labr: A large scale Arabic book reviews dataset. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 494–498.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2019. An effective approach to unsupervised machine translation. *arXiv preprint arXiv:1902.01313*.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2017. Unsupervised neural machine translation. *arXiv preprint arXiv:1710.11041*.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*, volume 10, pages 2200–2204.

Houda Bouamor, Nizar Habash, and Kemal Oflazer. 2014. A Multidialectal Parallel Corpus of Arabic. In *LREC*, pages 1240–1245.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, et al. 2018. The MADAR Arabic dialect corpus and lexicon. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Samhaa R. El-Beltagy. 2016. NileULex: A phrase and word level sentiment lexicon for Egyptian and Modern Standard Arabic. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2900–2905, Portorož, Slovenia. European Language Resources Association (ELRA).

Mohamed Elmahdy, Mark Hasegawa-Johnson, and Eiman Mustafawi. 2014. Development of a tv broadcasts speech recognition system for Qatari Arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3057–3061.

Donia Gamal, Marco Alfonse, El-Sayed M El-Horbaty, and Abdel-Badeeh M Salem. 2019. Twitter benchmark dataset for Arabic sentiment analysis. *Int J Mod Educ Comput Sci*, 11(1):33.

Lorenzo Gatti, Marco Guerini, and Marco Turchi. 2015. Sentiwords: Deriving a high precision and high coverage lexicon for sentiment analysis. *IEEE Transactions on Affective Computing*, 7:409–421.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, pages 177–180.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-based & neural unsupervised machine translation. *arXiv preprint arXiv:1804.07755*.

Karima Meftouh, Salima Harrat, Salma Jamoussi, Mourad Abbas, and Kamel Smaili. 2015. Machine translation experiments on PADIC: A parallel Arabic dialect corpus. In *The 29th Pacific Asia conference on language, information and computation*.

Matt Post. 2018. A call for clarity in reporting BLEU scores. *arXiv preprint arXiv:1804.08771*.

Hadeel Saadany and Constantin Orasan. 2020. Is it Great or Terrible? Preserving Sentiment in Neural Machine Translation of Arabic Reviews. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 24–37.

Hadeel Saadany and Constantin Orasan. 2021. BLEU, METEOR, BERTScore: Evaluation of Metrics Performance in Assessing Critical Translation Errors in Sentiment-oriented Text. *TRITON 2021*, page 48.

Hassan Sajjad, Ahmed Abdelali, Nadir Durrani, and Fahim Dalvi. 2020. AraBench: Benchmarking Dialectal Arabic-English Machine Translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5094–5107.

Mohammad Salameh, Saif Mohammad, and Svetlana Kiritchenko. 2015. Sentiment after translation: A case-study on Arabic social media posts. In *Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 767–777.

Wael Salloum and Nizar Habash. 2013. Dialectal Arabic to English machine translation: Pivoting through modern standard Arabic. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 348–358.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015a. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015b. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *LREC*, volume 2012, pages 2214–2218. Citeseer.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior research methods*, 45(4):1191–1207.

Omar Zaidan and Chris Callison-Burch. 2011. The Arabic online commentary dataset: an annotated dataset of informal Arabic with high dialectal content. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 37–41.

Rabih Zbib, Erika Malchiodi, Jacob Devlin, David Stallard, Spyros Matsoukas, Richard Schwartz, John Makhoul, Omar Zaidan, and Chris Callison-Burch. 2012. Machine translation of Arabic dialects. In *Proceedings of the 2012 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 49–59.

A Appendix

Ex1	Source Google Translate Our System Reference	سحلنى slay me pissed me off He made me quite angry
Ex2	Source Google Translate Our System Reference	اسفين جدا very wedged very sorry We are very sorry
Ex3	Source Google Translate Our System Reference	الله يحفظك مبحبش اكدب انا May God protect you, I love you May God protect you, I don’t like to lie May God protect you, I don’t like to lie
Ex4	Source Google Translate Our System Reference	الله لايوفقه God doesn’t help him God does not grant him success May God not grant him success.
Ex5	Source Google Translate Our System Reference	معليش خلينا شوي نتكلم يعني يسرق احسن هيك؟ OK let’s talk a little, I mean steal the best heck? Sorry, let’s talk a little he steals the best like this? Let’s just talk a bit, so does he better steal like this?
Ex6	Source Google Translate Our System Reference	بدون زعل فكونا من الكلام Without getting upset, let’s talk Without getting upset, so be from talking Without getting upset, so be it from talking