An English–Swahili parallel corpus and its use for neural machine translation in the news domain

Felipe Sánchez-Martínez,‡ Víctor M. Sánchez-Cartagena,‡ Juan Antonio Pérez-Ortiz,‡ Mikel L. Forcada,‡ Miquel Esplà-Gomis,‡ Andrew Secker,† Susie Coleman,† Julie Wall†

‡Dep. de Llenguatges i Sistemes Informàtics, Universitat d'Alacant
E-03690 Sant Vicent del Raspeig (Spain)
{fsanchez,vmsanchez,japerez,mlf,mespla}@dlsi.ua.es

†The British Broadcasting Corporation
BBC Broadcasting House, Portland Place, London, W1A 1AA. (UK)
{andrew.secker,susie.coleman,julie.wall}@bbc.co.uk

Abstract

This paper describes our approach to create a neural machine translation system to translate between English and Swahili (both directions) in the news domain, as well as the process we followed to crawl the necessary parallel corpora from the Internet. We report the results of a pilot human evaluation performed by the news media organisations participating in the H2020 EU-funded project GoURMET.

1 Introduction

Large news media organisations often work in a multilingual space in which they both publish their material in numerous languages and monitor the world's media across video, audio, printed and online sources. As regards *content creation*, one way in which efficient use is made of journalistic endeavour is the republication of news originally authored in one language into another; by using machine translation, and with the appropriate user interfaces, a journalist is able to take a news story or script, in the case of an audio or video report, and quickly obtain a preliminary translation that will be then manually post-edited to ensure it has the quality required to be presented to the audience. Concerning *news gathering*, expert monitors and journalists have to currently perform a lot of manual work to keep up with a growing amount of broadcast and social media streams of data; it is becoming imperative to automate tasks, such as translation, in order to free monitors and journalists to perform more journalistic tasks that cannot be achieved with technology.

In order to cope with these requirements, promoting both the reach of the news published to underserved audiences and the world-wide broadcasting of local information, the H2020 EU-funded project GoURMET (Global Under-Resourced Media Translation),[1] aims at improving neural machine translation (NMT) for under-resourced language pairs with special emphasis in the news domain. The two partner media organisations in the GoURMET project, the BBC in the UK and Deutsche Welle (DW) in Germany, publish news content in 40 and 30 different languages, respectively, and gather news in over 100 languages. In particular, both media partners gather news in and produce content in Swahili.

According to Wikipedia, Swahili has between 2 and 15 million first-language speakers and 90 million second-language speakers. As one of the largest languages in Africa and the recognised *lingua franca* of the East African community, BBC and DW see Swahili as an important language in which to make content available. The NMT systems described and evaluated herein can be deployed to support them in this domain specific context.

The rest of the paper is organised as follows. Next section describes the corpora we used to train our English–Swahili NMT systems in both translation directions. Section 3 then describes the crawling of the additional corpora we used and made publicly available. Section 4 describes the main linguistic contrasts between English and Swahili and the challenges they pose for building MT systems between them. Section 5 describes the resources, other than corpora, that we used to build our own systems and the technical details of the training of the NMT systems. Section 6 discussed the results of

[1] https://gourmet-project.eu/

Martins, Moniz, Fumega, Martins, Batista, Coheur, Parra, Trancoso, Turchi, Bisazza, Moorkens, Guerberof, Nurminen, Marg, Forcada (eds.)
Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, p. 299–308
Lisboa, Portugal, November 2020.

Corpus	Sent's	en tokens	sw tokens
GoURMET v1	156 061	3 334 886	2 981 699
SAWA	272 544	1 553 004	1 206 757
Tanzil v1	138 253	2 376 908	1 734 247
GV v2017q3	29 698	534 270	546 107
GV v2015	26 033	467 353	476 478
Ubuntu v14.10	986	2 486	2 655
EUbookshop v2	17	191	228
GNOME v1	40	168	170
total	623 632	8 269 266	6 948 341

Table 1: Parallel English–Swahili corpora used to train the NMT systems described in this work. *GV* stands for the GlobalVoices corpus.

automatic evaluation measures, describes a manual evaluation we are conducting and provides preliminary results. The paper ends with some concluding remarks.

2 Monolingual and bilingual corpora

Parallel data is the basic resource required to train NMT. Additionally, it is common practice to use synthetic parallel corpora obtained by back-translating monolingual data (Sennrich et al., 2016b). This section describes the corpora we used to train the NMT systems described in Section 5.

Tables 1 and 2 describe the parallel and monolingual corpora we used, respectively. As regards parallel corpora, with the exception of GoURMET and SAWA, all of them were downloaded from the OPUS website,[2] one of the largest repositories of parallel data on the Internet.[3] We used two additional parallel corpora: the SAWA corpus (De Pauw et al., 2011), that was kindly provided by their editors, and the GoURMET corpus, that was crawled from the web following the method described in Section 3.

As regards monolingual data, only three corpora were used: the NewsCrawl (Bojar et al., 2018) for English (`en`) and for Swahili (`sw`),[4] and the GoURMET monolingual corpus for `sw`. The first two corpora were chosen because they belong to the news domain, the same domain of application of our NMT systems. Given that the size of the `sw` monolingual corpus is much smaller than the size of the `en` monolingual corpus, additional monolingual data in `sw` was obtained as a by-product of the process of crawling parallel data from the web.

[2] `http://opus.nlpl.eu/`

[3] Table 1 contains the parallel corpora available at OPUS at the time of training our systems. New corpora have been added recently, such as the large JW300 corpus (Agić and Vulić, 2019), which we did not use.

[4] `http://data.statmt.org/news-crawl/sw/`

Corpus	Sent's	Tokens
NewsCrawl (en)	18 113 311	359 823 264
NewsCrawl (sw)	174 425	3 603 035
GoURMET (sw)	5 687 000	174 867 482

Table 2: Monolingual Swahili and English corpora used to build synthetic parallel data through back-translation.

3 Crawling of additional corpora

The amount of data for `en–sw` is clearly low, even if one compares it to the amount of data available for other under-resourced language pairs, such as English–Maltese or English–Icelandic.[5] For this reason, a new corpus was crawled from the Internet (see the GoURMET corpus in Table 1). This corpus has been made publicly available.[6]

The GoURMET corpus was obtained by using Bitextor (Esplà-Gomis and Forcada, 2010; Esplà-Gomis et al., 2019), a free open/source software that allows to identify parallel content on multilingual websites. Bitextor is organised as a pipeline that performs a sequence of steps to obtain parallel data from a list of URLs; for each of these steps, Bitextor supports different approaches that require different resources. In this section, the specific configuration of Bitextor for this work is described, as well as the resulting corpora crawled from the Web.

Crawling. Crawling is the first step of the pipeline implemented in Bitextor and consists of downloading any document containing text from the websites specified by the user. We used `wget`[7] to crawl documents from 3 751 websites;[8] these websites were obtained by leveraging automatic-language-identification metadata from the CommonCrawl corpus:[9] we consider those websites with at least 5 kB of text in `en` and in `sw`.

Every website was crawled during a period of 12 hours and only documents in `en` or `sw` were kept; CLD2[10] was used for automatic language identification. Plain text was extracted from HTML/XML and, after this, sentence splitting was applied to every document. From the collection of 3 751 preselected websites, 519 were not available at the time

[5] For example, in OPUS one can find about 3M sentence pairs for English–Icelandic and 7.6M sentence pairs for English–Maltese, whereas only 1.2M are available for `en–sw`.

[6] `http://data.statmt.org/gourmet/corpora/GoURMET-crawled.en-sw.zip`

[7] `https://www.gnu.org/software/wget/`

[8] The list of crawled websites can be found in the `hosts.gz` file accompanying the corpus.

[9] `https://commoncrawl.github.io/cc-crawl-statistics/plots/languages`

[10] `https://github.com/CLD2Owners/cld2`

of crawling and, from the remaining 3 232, only 908 ended up containing data in both languages.

Document alignment. In this step, documents that are likely to contain parallel data are identified. Bitextor supports two strategies for document alignment: one based on bilingual lexicons and another based on MT. The last option was not feasible in this work as no high-quality MT system between `sw` and `en` was available; therefore, the first one was used. This method combines information from bilingual lexicons, the HTML structure of the documents, and the URL to obtain a confidence score for every pair of documents to be aligned (Esplà-Gomis and Forcada, 2010). The bilingual lexicon used was automatically obtained from the word alignments obtained with `mgiza++` (Gao and Vogel, 2008) for the following corpora: EUBookshop v2, Ubuntu and Tanzil (see Table 1). A total of 180 520 pairs of documents were obtained by using this method.

Sentence alignment. In this step, aligned documents are segmented and aligned at the sentence level. Two sentence-alignment tools are supported by Bitextor: Hunalign (Varga et al., 2007) and BLEUalign (Sennrich and Volk, 2010). We used Hunalign because BLEUalign requires an MT system to be available. The same bilingual dictionary used for document alignment was provided to Hunalign in order to improve the accuracy of the alignment. After applying Hunalign, 2 051 678 unique segment pairs were obtained.

Cleaning. Bicleaner[11] (Sánchez-Cartagena et al., 2018) was used to clean the raw corpora obtained after sentence alignment. Cleaning implies removing the noisy sentence pairs that are either incorrectly aligned or not in the expected languages.[12] Bicleaner cleaning models require some language-dependent resources:

- Two probabilistic bilingual dictionaries, one for each direction for the language pair, built from the corpora used to build the bilingual lexica for document alignment.
- A parallel (ideally clean) corpus to train the regressor used to score the segment pairs in the raw corpus: the preexisting GlobalVoices v2015 parallel corpus was used, as Bicleaner requires parallel data used to train the dictionaries and the regressor to be different.
- A collection of pairs of segments that are wrongly aligned to train a language model: following Bicleaner's documentation, this collection was obtained from the raw parallel corpus by applying the "hard rules" implemented in Bicleaner.

Bicleaner was used to score all the sentence pairs in the raw corpus with two different scores: one coming from the regressor, which may be interpreted as the probability that the pair of sentences are parallel, and one coming from the language model, which is the probability that one of the sentences in the pair is malformed. After sampling a small fraction of the corpus, the score thresholds were set to 0.68 and 0.5, respectively. The resulting parallel corpus consisted of 156 061 pairs of segments.

In addition to the parallel corpus obtained after cleaning, a large amount of Swahili monolingual data was obtained as a by-product of crawling and released as a monolingual corpus. Monolingual data cleaning consisted of discarding those sentences not deemed fluent enough to be used for NMT training. Sentences were ranked by perplexity computed by a character-based 7-gram language model and only the 6 million sentences with the lowest perplexity were kept. The language model was trained[13] on the concatenation of the `sw` side of the parallel corpora listed in Table 1, excluding GoURMET. Moreover, those sentences that were automatically identified not to be in `sw`,[14] or contained more numeric or punctuation characters than alphabetic characters were also discarded.

4 Contrasts and challenges for MT

Swahili belongs to a very large African language family, the Niger–Congo family, and more specifically to the Bantu group. Swahili is currently written in the Latin script, with no diacritics; the apostrophe is used in the seldom-occurring combination *ng'* which represents the sound of *ng* in *singer* (not *finger*); one common example is *ng'ombe*, ('cow').

Swahili is morphologically and syntactically quite different from English, in spite of the fact that both are subject–verb–object languages. Swahili verb morphology is rich and agglutinative, and a

[11] `https://github.com/bitextor/bicleaner/`

[12] This additional language checking is required as document-level language identification may be too general and small fragments in other languages can be included in the sentence-aligned corpus.

[13] The language model was trained with KenLM (Heafield, 2011) with modified Kneser-Ney smoothing (Ney et al., 1994).

[14] Automatic language identification was carried out by using CLD3: `https://github.com/google/cld3`

large number of morphologically-marked nominal genders participate in nominal and verbal agreement. Table 3 provides a summary of the main linguistic contrasts between en and sw; some examples are from Perrott (1965) and the table is mostly based on https://wals.info.

The challenges to build an MT system for news translation between en and sw are twofold. On the one hand, parallel corpora are rather scarce. On the other hand, a number of challenges stem from the linguistic divergences between the two languages:

- The absence of definite and indefinite articles in sw may make the generation of grammatical en tricky.
- Genders in sw do not mark sex (in fact, all nouns designating people are in the same gender or class); generating the correct en 3rd-person pronouns and possessives may be challenging.
- When translating into sw, the presence of many noun classes and their agreement inside noun phrases and with verbal affixes may be an important obstacle.
- Swahili interrogatives have to be reordered when translating to en.
- Fortunately, most word-order differences seem to occur locally (basically inside the noun phrase). This may only be a problem for longer noun phrases.

5 Neural machine translation model

This section describes the steps followed to build en→sw and sw→en NMT systems from the corpora described in Section 2. We firstly describe corpora preprocessing and give details about the NMT architecture used and the process followed to choose it. Secondly, we present the strategies followed in order to take advantage of monolingual corpora and to integrate linguistic information into the NMT systems.

5.1 Corpus preparation

In order to properly train NMT systems, we need a development corpus to help the training algorithm decide when to finish, and a test corpus that allows us to estimate the quality of the systems.

We obtained both of them from the GlobalVoices parallel corpus. We randomly selected 4 000 parallel sentences from the concatenation of GlobalVoices-v2015 and GlobalVoices-v2017q3, and split them into two halves (with 2 000 sentences each), which were used respectively as development and test corpora. The half reserved to be used as test corpus was further filtered to remove the sentences that could be found in any of the monolingual corpora.

The remaining sentences from GlobalVoices-v2015 and GlobalVoices-v2017q3, together with the other parallel corpora listed in Table 1 were deduplicated to obtain the final parallel corpus used to train the NMT systems.

All corpora were tokenised with the Moses tokeniser (Koehn et al., 2007) and truecased. Parallel sentences with more than 100 tokens in either side were removed. Words were split in sub-word units with byte pair encoding (BPE; Sennrich et al. (2016c)). Table 4 reports the size of the corpora after this pre-processing.

5.2 Neural machine translation architecture

We trained the NMT models with the Marian toolkit (Junczys-Dowmunt et al., 2018). Since training hyper-parameters can have a large impact in the quality of the resulting system (Lim et al., 2018), we carried out a grid search in order to find the best hyper-parameters for each translation direction. We explored both the Transformer (Vaswani et al., 2017) and recurrent neural network (RNN) with attention (Bahdanau et al., 2014) architectures. Our starting points were the Transformer hyper-parameters[15] described by Sennrich et al. (2017) and the RNN hyper-parameters[16] described by Sennrich et al. (2016a).

For each translation direction and architecture, we explored the following hyper-parameters:

- Number of BPE operations: 15 000, 30 000, or 85 000.
- Batch size: 8 000 tokens (trained on one GPU) or 16 000 tokens (trained on two GPUs).
- Whether to tie the embeddings for both languages (Press and Wolf, 2017)

We trained a system for each combination of hyper-parameters, using only the parallel data described above. Early stopping was based on perplexity on the development set and patience was set to 5. We selected the checkpoint that obtained the

[15] https://github.com/marian-nmt/marian-examples/tree/master/wmt2017-transformer

[16] https://github.com/marian-nmt/marian-examples/tree/master/training-basics

Feature	Value in English	Value in Swahili	Examples
Coding of plurality in nouns	Plural suffix	Plural prefix	***ki**chwa* ('head'), ***vi**chwa* ('heads'); ***ji**cho* ('eye'), ***ma**cho* ('eyes')
Number of categories encoded in a single-word verb	Few (number, person, tense)	Many ("STROVE", that is, number and person of **s**ubject, **t**ense, aspect and mood, optional **r**elatives, number and person of **o**bject, **v**erb root, and optional **e**xtensions)	*nimekinunua kitabu* 'I have bought the book', where: *ni* 'I', subject; *me*, present perfect; *ki*, 'it', object; *nunua*, 'buy', verb root.
Definite articles	Definite word distinct from demonstrative	Demonstrative (seldom) used as definite article	*kitabu* ('book', 'the book', 'a book').
Noun Phrase Conjunction	*And* different from *with*	*And* identical to *with*	*Lete chai na maziwa* ('Bring tea and milk'); *Yesu alikuja na Baba yake* ('Jesus came with his Father').
Inflectional morphology	Suffixing	Mainly prefixing	*kitabu* ('book'), *vitabu* ('books'); *nilinunua* ('I bought'), *ulinunua* ('You bought'); but *jenga* ('build'), *jengwa* ('be built')
Reduplication	No productive reduplication	Productive full and partial reduplication	*Mimi ninasoma kitabu* 'I am reading the book'; *mimi ninasomasoma kitabu* 'I am reading the book bit by bit'
Number of genders	Three, sex-based, only in 3rd person singular pronouns and possessives	Many, not based on sex (called *classes*)	*kitabu* 'book' (*ki-vi*-class): plural *vitabu* 'books' ; *mtoto* 'child' (*m-wa*-class): plural *watoto* 'children' ; etc. Note that adjectives and verbs have to agree: *kitabu kidogo* 'small book', *vitabu vidogo* 'small books'; *mtoto mdogo* 'small child', etc.
Order of genitive and noun	No dominant order	Noun–genitive	*gari la mama* 'Mom's (*mama*) car (*gari*)'; *paa la nyumba* 'The roof (*paa*) of the house (*nyumba*)'.
Order of adjetive and noun	adjective–noun	noun–adjective	*mtoto mdogo* 'small child', lit. 'child small'
Order of demonstrative and noun	demonstrative–noun	noun–demonstrative	*gari hili* 'this car', lit. 'car this'
Order of numeral and noun	numeral–noun	noun–numeral	*vitabu viwili* ('two books', lit. 'books two')
Expression of Pronominal Subjects	Obligatory pronouns in subject position	Subject affixes on verb	*Nilinunua* ('I bought'), *ulinunua* ('You bought')
Negation	Particle or construction	Negative form of verb	*Ninasoma* ('I am reading'), *Sisomi* ('I am not reading'); *Unasoma* ('You are reading'), *husomi* ('You are not reading');
Position of Interrogative Phrases in Content Questions	Initial interrogative phrase	Not initial interrogative phrase	*Unasoma vitabu* ('You are reading books'); *Unasoma nini?* ('What are you reading', lit. 'you are reading what?')
Polar questions	Change in word order, use of auxiliaries	No change in word order	*Amesoma* ('He has read'); *Amesoma?* ('Has he read?')
Comparative	Comparative form of adjective ('-er') or 'more'	Absolute form of adjective	*Virusi ni ndogo* ('A virus is small') *Virusi ni ndogo kuliko bakteria* ('A virus is smaller than a bacterium', lit. 'A virus is small where there is a bacterium')
Predicative Possession	'have'	conjunctional ('to be with')	*Nina swali* ('I have a question', lit. 'I-am-with question')

Table 3: A summary of linguistic contrasts between English and Swahili.

highest BLEU (Papineni et al., 2002) score on the development set.

We obtained the highest test BLEU scores for en→sw with an RNN architecture, 30 000 BPE operations, tied embeddings and single GPU, while the highest ones for sw→en were obtained with a Transformer architecture, 30 000 BPE operations, tied embeddings and two GPUs.

5.3 Leveraging monolingual data

Once the best hyper-parameters were identified, we tried to improve the systems by making use of the monolingual corpora via back-translation. Back-translation (Sennrich et al., 2016b) is a widespread method for integrating target-language (TL) monolingual corpora into NMT systems. The quality of a system trained on back-translated data is usually

Corpus	Sentences	en tokens	sw tokens
parallel	424 821	7 536 537	6 191 959
NewsCrawl (en)	40 000 000	796 199 072	-
NewsCrawl (sw)	414 598	-	8 377 157
GoURMET mono (sw)	5 687 000	-	174 867 482
development	2 000	41 726	42 037
test (en-sw)	1 863	41 097	41 188
test (sw-en)	1 969	43 149	43 174

Table 4: Size of the corpora used to build the NMT systems after preprocesing. For the en NewsCrawl corpus, only the size of the subset that has been used for training is displayed. Token counts were calculated before BPE splitting.

correlated with the quality of the system that translates the TL monolingual corpus into the source language (SL) (Hoang et al., 2018, Sec. 3). We took advantage of the fact that we are building systems for both the en→sw and sw→en directions and applied an iterative back-translation (Hoang et al., 2018) algorithm that simultaneously leverages monolingual sw and monolingual en data. It can be outlined as follows:

1. With the best identified hyper-parameters for each direction we built a system using only parallel data.
2. en and sw monolingual data were back-translated with the systems built in the previous step.
3. Systems in both directions were trained on the combination of the back-translated data and the parallel data.
4. Steps 2–3 were re-executed 3 more times. Back-translation in step 2 was always carried out with the systems built in the most recent execution of step 3, hence the quality of the system used for back-translation improved with each iteration.

The sw monolingual corpus used in step 2 was the GoURMET monolingual corpus. The en monolingual corpus was a subset of the NewsCrawl corpus, the size of which was duplicated after each iteration. It started at 5 million sentences.

Since the sw NewsCrawl corpus was made available near the end of the development of our MT systems, it could not be used during the iterative back-translation process. Nevertheless, we added it afterwards: the sw NewsCrawl was back-translated with the last available sw→en system obtained after completing all the iterations, concatenated to the existing data for the en→sw direction and the MT system was re-trained.

5.4 Integrating linguistic information

In addition to the corpora described above, linguistic information encoded in a more explicit representation was also employed to build the MT systems. In particular, we explored the *interleaving* (Nadejde et al., 2017) of linguistic tags in the TL side of the training corpus with the aim of enhancing the grammatical correctness of the translations.

Morphological taggers were used to obtain the interleaved tags added to the training corpus. The sw text was tagged with TreeTagger (Schmid, 2013). We used a model[17] trained on the Helsinki Corpus of Swahili.[18] The en text was tagged with the Stanford tagger (Qi et al., 2018), which was trained on the English Web Treebank (Silveira et al., 2014).

Figure 1 shows examples of en→sw and sw→en training parallel sentences with interleaved tags. While the tags returned by the sw tagger were just part-of-speech tags, en tags contained also morphological inflection information. Interleaved tags are removed from the final translations produced by the system.

6 Evaluation

This section reports the scores obtained on the test corpus using automatic evaluation metrics. It then describes the manual evaluation we are conducting at the time of writing these lines and provides preliminary results.

6.1 Automatic evaluation

Table 5 shows the BLEU and chrF2++ scores, computed on the test set, for the different steps in the development of the MT systems. All systems were trained with the hyper-parameters described in Section 5.2. As a reference, we also show the scores obtained by the translation obtained with Google Translate[19] on 6th March 2020 using the web interface.

It is worth noting the positive effect of adding monolingual data during the iterative back-translation iterations and that interleaved tags also help to improve the systems according to the automatic evaluation metrics.

[17] Available at https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

[18] https://korp.csc.fi/download/HCS/a-v2/hcs-a-v2-dl

[19] https://translate.google.com/

en (SL): he 's studying law at No@@ tre D@@ ame .
sw (TL): VFIN A@@ nj@@ ifunza N sheria PRON huko PROPNAME No@@ tre PROPNAME D@@ ame

sw (SL): A@@ nj@@ ifunza sheria huko Notre Dame
en (TL): PRON|Nom|Masc|Sing|3|Prs he AUX|Ind|Sing|3|Pres|Fin 's VERB|Pres|Part studying NOUN|Sing law ADP at PROPN|Sing No@@ tre PROPN|Sing D@@ ame

Figure 1: Examples of parallel sentences after interleaving linguistic tags. The @@ symbol is placed at the end of each BPE sub-word when it is not the last sub-word of a token. The tag corresponding to the morphological analysis of a token is interleaved before the first sub-word unit of the token.

Strategy	it.	BLEU	chrF++
en→sw			
only parallel	-	22.23	46.34
iter. backt.	1	25.59	50.08
iter. backt.	2	26.22	50.91
iter. backt.	3	26.36	51.09
iter. backt.	4	26.58	51.39
+ NewsCrawl	4	26.77	51.46
+ NewsCrawl + tags	4	27.42	52.11
Google Translate	-	23.24	48.80
sw→en			
only parallel	-	22.66	44.62
iter. backt.	1	29.29	51.19
iter. backt.	2	29.70	51.82
iter. backt.	3	29.99	51.98
iter. backt.	4	30.19	52.10
+ tags	4	30.55	52.72
Google Translate	-	30.36	53.32

Table 5: Automatic evaluation results obtained for the different development steps of the MT systems: *only parallel* stands for the systems trained only on parallel data with the best hyper-parameters; *iter. backt.* represents systems obtained after iteratively back-translating monolingual data (iteration number is shown in column *it.*); *+NewsCrawl* means that the sw NewsCrawl corpus was back-translated and added; and *+ tags* indicates that TL linguistic tags were interleaved.

Finally, our system clearly outperforms Google Translate for the en→sw direction, while their performances are close for the opposite direction. We noticed that the sw→en Google Translate system improved dramatically since we built our systems, which suggests that their systems may be trained on data that was not available at OPUS website at that time.

6.2 Manual evaluation

Manual evaluation requires the use of humans to give subjective feedback on the quality of translation, either directly or indirectly. All manual evaluation undertaken within the GoURMET project uses in-domain data, i.e. test data derived from news sources. Two types of subjective evaluation have been selected and applied in order to generate the most insight for the media partners:

- *Direct assessment* (Graham et al., 2016a; Graham et al., 2016b) (DA) is used to test en→sw. This corresponds to the content creation use case which will use translation predominantly in this direction, and where the correctness of the translation is key.
- *Gap filling* (Forcada et al., 2018) (GF) is used to test sw→en. This corresponds to the media monitoring use case which will use translation almost exclusively in this direction and where getting the gist of the meaning of a sentence is enough to fulfil the use-case, perfect translation of sentence structure is less important.

Custom interfaces were created to support both evaluations; see figures 2 and 3 for DA and GF, respectively.

Evaluators were recruited from within the media partner organisations to complete the DA and GF tasks. Evaluators were required to have an excellent level of comprehension in the TL (i.e. sw for DA and en for GF) and precedence was given to journalists who write exclusively or predominately in one of the two target languages.

Media partners (BBC, DW) prepared test data using previously published articles. For DA this consisted of 205 sentences drawn at random from six different articles originally published in en by DW. The test data was further augmented with 5 sentences written in the TL by a human and used as *calibration* examples resulting in a total of 210 sentences shown to each evaluator in random order. All evaluators were asked to rate the quality of the translated sentence on a sliding scale from 0 to 100 for two criteria according to the statement "*For the pair of sentences below read the text and state how much you agree that: Q1) The black text adequately expresses the meaning of the grey text and Q2) The black text is a well written phrase or sentence that is grammatically and idiomatically correct*". The ratings for the first five sentences were discarded as practice evaluations while the results for the five sentences used for calibration were discarded, leaving 200 pairs of results for each evaluator. Four evaluators completed the task.

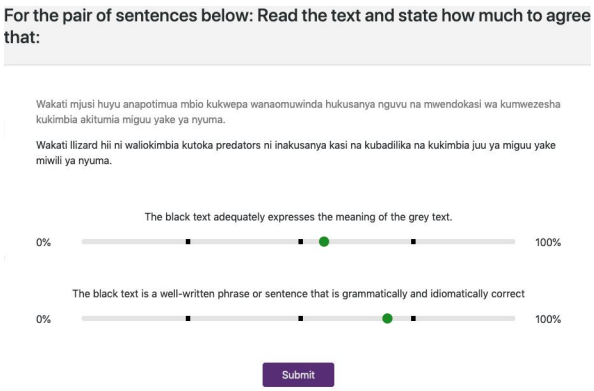

Figure 2: Custom Direct Assessment interface.

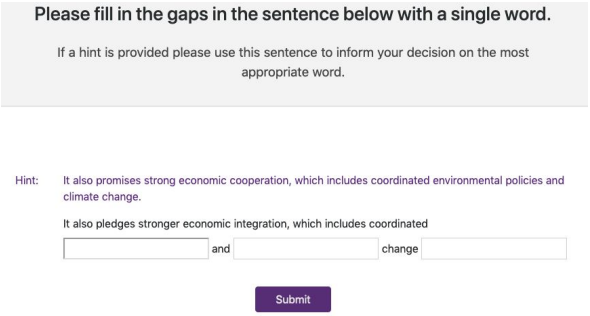

Figure 3: Custom Gap Filling interface.

Hint type	Success rate
No hint	26.45%
Google	60.60%
GoURMET	54.34%

Table 6: Gap-filling success rates for each hint type

For GF 30 sentences were selected from six different articles originally published in sw by DW. Each sentence was translated into en by a professional translator and it was ensured that once translated, each sentence was 15 words or more in length. For each sentence in en, 20% of the content words were removed, making sure there were no two consecutive gaps, typically leaving between 1 and 8 missing words in each sentence, averaging 2.67, for a total of 70 different missing-word problems. Each sentence in sw was translated into en by the GoURMET MT system described here, and Google Translate. The work of seventeen human evaluators was collected and their work on each of the 30 sentences was evaluated in three different ways: one evaluator saw the gapped sentence with no hint; one evaluator saw the gapped sentence with the GoURMET MT output as a hint; finally, one evaluator saw the gapped sentence with the Google Translate output as a hint. A total of 210 different missing-word/hint type configurations were therefore evaluated by an average of 17/3=5.67 evaluators. Sentences were distributed in such a way that no evaluator ever saw the same sentence twice. The GF evaluation requires the evaluator to fill in the missing words using the hint (if present). The accuracy is simply a *success rate*: the fraction of gaps correctly filled.

6.3 Manual evaluation results

Gap-filling (GF) success rates are shown in Table 6. As may be seen, Google Translate seems to be more helpful in this gisting task than the system created in this paper. To get an idea of how significant this difference is, Figure 4 shows a box-and-whisker plot of the distribution of success rates for each hint type by evaluator. As may be seen, the boxes for Google Translate and GoURMET clearly overlap, meaning that the difference in usefulness is not significant. However we also notice a slight overlap between the GoURMET success-rate distribution and that when there is no hint (NONE); this overlap does not occur with Google Translate.

Direct asssesment (DA): evaluators 1 and 2 scored the calibration sentences with values close to the expected ones (0 or 100 depending on the sentence), but evaluators 3 and 4 provided relatively inconsistent scores. Besides that, there is a weak positive correlation among the evaluators' answers (Pearson correlation coefficients between 0.22 and 0.46 for Q1, and between 0.24 and 0.49 for Q2, the highest values corresponding to evaluators 1 and 2 in both cases). Consequently, Table 7 shows the average score per evaluator. Unfortunately, these scores do not allow us to extract reliable conclusions.

7 Concluding remarks

We have described the development and evaluation of an NMT system to translate in the news domain between English and Swahili in both directions. We have also described the crawling of a new parallel corpus from the Internet which we have made publicly available.

We performed an automatic evaluation of both systems. According to it, the en→sw NMT system performs better than *Google Translate*, whereas the sw→en systems performs on par with it. In addition, the sw→en NMT system was manually evaluated to ascertain it usefulness for gisting purposes, and the en→sw NMT system as regards

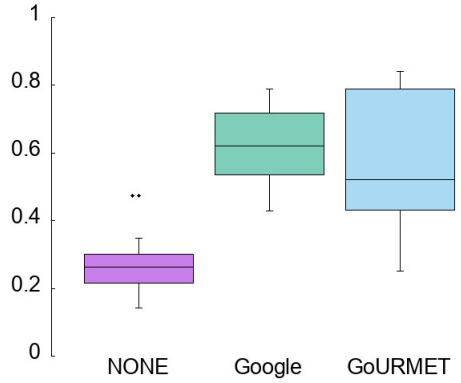

Figure 4: Gap-filling success rate distribution across evaluators for each hint type.

	Q1	Q2
Evaluator 1	77.65 ± 3.97	70.80 ± 4.30
Evaluator 2	47.30 ± 6.20	52.94 ± 5.97
Evaluator 3	48.42 ± 3.28	60.20 ± 3.75
Evaluator 4	54.40 ± 3.92	56.53 ± 4.02

Table 7: Average score and confidence intervals (estimated via standard significance testing) for questions Q1 and Q2 in the direct assessment evaluation.

its fluency and adequacy. The preliminary results of both evaluations show that the sw→en system performs similarly to *Google Translate* (which is consistent with the automatic evaluation), and that the en→sw system needs to be further evaluated because evaluators provided quite different scores.

As future work, and in view of the scarcity of bilingual resources available, we plan to try approaches based on monolingual corpora (Artetxe et al., 2018). We also plan to study if a correct segmentation of verbs, which are very rich and complex (see Table 3), as a pre-processing step helps improve performance.

Acknowledgements: Work funded by the European Union's Horizon 2020 research and innovation programme under grant agreement number 825299, project Global Under-Resourced Media Translation (GoURMET). We thank the editors of the SAWA corpus for letting us use it for training. We also thank Wycliffe Muia (BBC) for help with Swahili examples and DW for helping in the manual evaluation.

References

Agić, Ž. and I. Vulić. 2019. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Florence, Italy, July.

Artetxe, M., G. Labaka, and E. Agirre. 2018. Unsupervised statistical machine translation. *arXiv preprint arXiv:1809.01272*.

Bahdanau, D., K. Cho, and Y. Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Bojar, O., C. Federmann, M. Fishel, Y. Graham, B. Haddow, P. Koehn, and C. Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels, October. Association for Computational Linguistics.

De Pauw, G., P.W. Wagacha, and G.-M. De Schryver. 2011. Exploring the SAWA corpus: collection and deployment of a parallel corpus English–Swahili. *Language resources and evaluation*, 45(3):331.

Esplà-Gomis, M., M.L. Forcada, G. Ramírez-Sánchez, and H. Hoang. 2019. ParaCrawl: Web-scale parallel corpora for the languages of the EU. In *Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks*, pages 118–119, Dublin, Ireland, August.

Esplà-Gomis, M and M.L. Forcada. 2010. Combining content-based and url-based heuristics to harvest aligned bitexts from multilingual sites with bitextor. *The Prague Bulletin of Mathematical Linguistics*, 93:77–86.

Forcada, M.L., C. Scarton, L. Specia, B. Haddow, and A. Birch. 2018. Exploring gap filling as a cheaper alternative to reading comprehension questionnaires when evaluating machine translation for gisting. *CoRR*, abs/1809.00315.

Gao, Q. and S. Vogel. 2008. Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57, Columbus, Ohio, USA, June.

Graham, Y., T. Baldwin, M. Dowling, M. Eskevich, T. Lynn, and L. Tounsi. 2016a. Is all that glitters in machine translation quality estimation really gold? In *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3124–3134, Osaka, Japan.

Graham, Y., T. Baldwin, A. Moffat, and J. Zobel. 2016b. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, 23(1):3–30.

Heafield, K. 2011. KenLM: faster and smaller language model queries. In *Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland, UK, July.

Hoang, V.C.D., P. Koehn, G. Haffari, and T. Cohn. 2018. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24, Melbourne, Australia, July.

Junczys-Dowmunt, M., R. Grundkiewicz, T. Dwojak, H. Hoang, K. Heafield, T. Neckermann, F. Seide, U. Germann, A. Fikri Aji, N. Bogoychev, A.F.T. Martins, and A. Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia, July.

Koehn, P., H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 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, Prague, Czech Republic, June.

Lim, R.V., K. Heafield, H. Hoang, M. Briers, and A.D. Malony. 2018. Exploring hyper-parameter optimization for neural machine translation on GPU architectures. *CoRR*, abs/1805.02094.

Nadejde, M., S. Reddy, R. Sennrich, T. Dwojak, M. Junczys-Dowmunt, P. Koehn, and A. Birch. 2017. Predicting target language CCG supertags improves neural machine translation. In *Proceedings of the Second Conference on Machine Translation, Volume 1: Research Papers*, pages 68–79, Copenhagen, Denmark, September.

Ney, H., U. Essen, and R. Kneser. 1994. On structuring probabilistic dependences in stochastic language modelling. *Computer Speech & Language*, 8(1):1 – 38.

Papineni, K., S. Roukos, T. Ward, and W.-J. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, Philadelphia, PA, USA, July.

Perrott, D.V. 1965. *Teach Yourself Swahili*. English Universities Press.

Press, O. and L. Wolf. 2017. Using the output embedding to improve language models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 157–163, Valencia, Spain, April.

Qi, P., T. Dozat, Y. Zhang, and C.D. Manning. 2018. Universal dependency parsing from scratch. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 160–170, Brussels, Belgium, October.

Sánchez-Cartagena, V.M., M. Bañón, S. Ortiz-Rojas, and G. Ramírez-Sánchez. 2018. Prompsit's submission to wmt 2018 parallel corpus filtering shared task. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, Brussels, Belgium, October.

Schmid, H. 2013. Probabilistic part-of-speech tagging using decision trees. In *New methods in language processing*, page 154.

Sennrich, R. and M. Volk. 2010. MT-based sentence alignment for ocr-generated parallel texts. In *Proceedings of the Ninth Conference of the Association for Machine Translation in the Americas*, Denver, USA, October.

Sennrich, R., B. Haddow, and A. Birch. 2016a. Edinburgh Neural Machine Translation Systems for WMT 16. In *Proceedings of the First Conference on Machine Translation*, pages 371–376, Berlin, Germany, August.

Sennrich, R., B. Haddow, and A. Birch. 2016b. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany, August.

Sennrich, R., B. Haddow, and A. Birch. 2016c. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1715–1725, Berlin, Germany, August.

Sennrich, R., A. Birch, A. Currey, U. Germann, B. Haddow, K. Heafield, A.V. Miceli Barone, and P. Williams. 2017. The University of Edinburgh's Neural MT Systems for WMT17. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 389–399, Copenhagen, Denmark, September.

Silveira, N., T. Dozat, M.-C. de Marneffe, S. Bowman, M. Connor, J. Bauer, and C.D. Manning. 2014. A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, pages 2897–2904, Reykjavik, Iceland, May.

Varga, D., P. Halácsy, A. Kornai, V. Nagy, L. Németh, and V. Trón. 2007. Parallel corpora for medium density languages. *Amsterdam Studies In The Theory And History Of Linguistic Science Series 4*, 292:247.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, and I. Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008.