Linguistically-Motivated Yorùbá-English Machine Translation

Ife Adebara[1,2] **Muhammad Abdul Mageed**[1,2] **Miikka Silfverberg**[1]
Department of Linguistics[1]
Deep Learning and Natural Language Processing Group[2]
The University of British Columbia
{ife.adebara, muhammad.mageed, miikka.silfverberg}@ubc.ca

Abstract

Translating between languages where certain features are marked morphologically in one but absent or marked contextually in the other is an important test case for machine translation. When translating into English which marks (in)definiteness morphologically, from Yorùbá which uses bare nouns but marks these features contextually, ambiguities arise. In this work, we perform fine-grained analysis on how an SMT system compares with two NMT systems (BiLSTM and Transformer) when translating *bare nouns* in Yorùbá into English. We investigate how and to what extent the systems identify BNs, correctly translate them, and compare with human translation patterns. We also analyze the type of errors each model makes and provide a linguistic description of these errors. We glean insights for evaluating model performance in low-resource settings. In translating bare nouns, our results show the transformer model outperfoms the SMT and BiLSTM models for 4 categories, the BiLSTM outperforms the SMT model for 3 categories while the SMT outperforms the NMT models for 1 category.

1 Introduction

Languages differ with regard to how grammatical information such as “case" and “number" are expressed. In some languages, this information is overtly marked using morphological or syntactic means, whereas in others it has to be inferred from context. This asymmetry of information representation poses an important problem for machine translation (Mitkov, 1999; Hardmeier, 2012). Several phenomena where asymmetry arises have been identified as challenging problems for machine translation. These include: pronoun translation and coreference (Guillou et al., 2019), politeness (Sennrich et al., 2016), lexical cohesion (Carpuat, 2009), and lexical disambiguation (Gonzales et al., 2017).

Asymmetry of information representation presents an interesting test case for MT systems because it can shed light on their true linguistic ability (Voita et al., 2018; Bawden et al., 2017). It is, therefore, important to use evaluation measures which can capture this aspect of the translation task. However, the most popular evaluation metric for machine translation, the BLEU score (Papineni et al., 2002a), is a coarse metric which can often hide these fine-grained morphological and semantic distinctions. In fact, a high BLEU score is no guarantee of improved translation quality and BLEU, being based on precision on short ngrams, may be poorly suited for measuring the coherence and grammaticality of a sentence.

In this paper we investigate the performance of Yorùbá [1] to English Machine Translation. We specifically evaluate performance on translating *Bare nouns* (BNs). BNs (Cheng and Sybesma, 1999; Krifka, 2003; Chierchia, 1998; Larson, 1985; Carlson, 1989) are nouns without an overt determiner or quantifier. For instance "houses" in *Houses are expensive in New York* is a BN. Whereas English accounts for only plural BNs, BNs in Yorùbá are number neutral and can also be definite or indefinite depending on the context. Consider the following example:

(1) *Bàbá ra ìṣu*
FATHER BUY YAM

a. *‘Father bought a yam.’* **(Indefinite Singular)**
b. *‘Father bought some yams.’* **(Indefinite Plural)**
c. *‘Father bought the yam.’* **(Definite Singular)**
d. *‘Father bought the yams.’* **(Definite Plural)**

In Example (1), the BN *ìṣu yam* can be translated into English in four ways as: the indefinite singular *a yam*, an indefinite plural *some yams*, a

[1] Yorùbá is a tone language that belongs to the Yoruboid group of the Kwa branch of the Niger-Congo language family, which is spoken by over 40 million in Nigeria. Yorùbá is spoken primarily in western Nigeria and eastern Benin, with communities in Sierra Leone and Liberia, and expatriate communities throughout Africa, Europe, and the Americas.

Proceedings of the 29th International Conference on Computational Linguistics, pages 5066–5075
October 12–17, 2022.

definite singular *the yam*, and a definite plural *the yams*. This poses a challenge akin to anaphora resolution, as the correct translation of Yorùbá BNs can only be determined by examining the context in which the BN occurs in the source text. The context can span one or more preceding or current words, phrases, clauses or sentences. It can also include world knowledge.

Our study provides a fine grained analysis that sheds light on issues in MT that are not often discussed in main stream research. We turn away from the current research trend in massively multilingual translation systems, to this largely underexplored fine-grained aspect of the MT problem. We investigate how SMT and NMT systems compare in general, but also specifically with respect to translation of BNs. Although NMT has recently been reported to outperform SMT even in low-resource settings, these findings were reported for systems translating between somewhat similar languages (e.g., languages that belong to the same language family, have overlapping vocabulary and similar script) (Junczys-Dowmunt et al., 2016; Conneau et al., 2017; Lample et al., 2017; Wang et al., 2019) such as English → French, German → French, German → Italian; languages for which large corpora exists. In this work, we collect a new dataset and use it to test whether the same NMT advantage persists by exploring two *typologically dissimilar* languages belonging to different language families: Yorùbá and English. This work bears significance not only to Yorùbá but to analytic languages, low-resource languages, and languages that have BNs. Our contributions are as follows: **(1)** We align a new dataset for the Yorùbá and English low-resource setting. **(2)** We use our dataset to develop statistical and NMT Yorùbá → English models. **(3)** We study the linguistic ability of our models in disambiguating BNs.

The rest of the paper is organized as follows: Section 2 is a description of disambiguation patterns of Yorùbá BNs. We discuss related work in Section 3. Section 4 presents the datasets, data collection process, and preprocessing. In Section 5, we present our methods. We present results of our models and an evaluation of BN disambiguation in Section 6. We conclude in Section 7.

2 Disambiguating Yorùbá BNs

A number of contextual variables are important when disambiguating BNs in Yorùbá. These include the so-called *familiarity* and *uniqueness* conditions (Roberts, 2003; Russell, 1905; Abbott, 2006), as well as the *category of the verb* that occurs in the environment of the BN. Familiarity refers to information which is already known from the previous textual context either explicitly or through inference. In Yorùbá, a definite interpretation is permitted when the existence of the entity referred to has been established in the discourse, an indefinite occurs otherwise. In example (2), the first mention of book is indefinite, the second is definite.

(2) *Mo ra* ***ìwé*** *fún Kọ́lá. L'ọ́jọ́ kejì,*
1SG BUY BOOK FOR KỌLÁ. IN DAY SECOND
Kọ́lá ti so ***ìwé*** *nù*
KỌLÁ PERF THROW BOOK AWAY

'I bought a book for KỌLÁ. By the second day, Kọ́lá had lost the book.'

The uniqueness condition claims that there exists one and only one entity that meets the descriptive content of the BN as in example (3). This entails that the BN will be interpreted as definite and singular.

(3) ***Oòrùn*** *máa ń ràn ní* ***ọ̀sán***
SUN HAB PROG SHINE IN AFTERNOON

'The sun shines in the afternoon.'

The category of verb is also an important contextual element in correctly translating BNs. *Stative* verbs (verbs that describe the state of being or situation such as *to own* and *to feel*) introduce the *Generic* description where a noun phrase is used to refer to a whole class as in example (4). *Eventive* verbs (verbs that describe events such as *to break* and *to appear*) introduce all other disambiguation patterns as in example (1).

(4) *Ọmọ fẹ́rán* ***ajá*** ***(generic) (STATIVE)***
CHILD LOVE DOG

'Children love dogs'

3 Related work

Traditional evaluation methods like BLEU are inadequate for evaluation of fine-grained discourse phenomena in MT, and various approaches have been explored to alleviate this problem. Guillou and Hardmeier (2016) present the PROTEST pronoun translation test suite. The test suite contains

examples of pronoun translation which are known to be challenging for MT systems. Isabelle et al. (2017) present another challenge set for MT. This set consists of a number of human annotated sentences which are designed to probe a system's capacity to handle various linguistic phenomena (like EXAMPLES). Using the challenge set, Isabelle et al. (2017) present a comparison of SMT and NMT systems for English-French MT which provides a fine-grained exploration of the strengths of NMT, as well as insight into linguistic phenomena that typically present difficulties for NMT models. Tests suites that evaluate more than 100 linguistic phenomena was developed for English to German and German to English machine translation (Macketanz et al., 2021). The test sentences are given as input to the MT systems. The MT outputs are then evaluated by the set of rules which determine whether the output was correctly translated or not.

Sennrich (2016) assess the grammaticality of the output of a character-level NMT system by evaluating the MT model's capacity to correctly rank contrastive pairs of pre-existing translations, one of which is correct and the other one incorrect. This approach has also been applied to lexical disambiguation of English-German MT (Gonzales et al., 2017). Bentivogli et al. (2016) explore automatic detection and classification of translation errors based on manual post-edits of MT output both for SMT and NMT systems. They use a classification that evaluated outputs for morphological, lexical, and word order errors which was a simplification of those used in Hjerson (Popović, 2011). Hjerson detects word level error classes: morphological errors, re-ordering errors, missing words, extra words and lexical errors.

Neural models have also been evaluated for syntactic competence. For instance, Linzen et al. (2016) probe the ability of LSTM models to learn English subject-verb agreement. When provided explicit supervision, LSTMs were able to learn to perform the verb number agreement task in most cases, although their error rate increased on particularly difficult sentences. NMT systems have also been evaluated for morphological competence while translating from English to a morphologically rich language (Burlot and Yvon, 2017). Certain linguistic phenomena have also been tested across language families. For instance, for four language families: Slavic, Germanic, Finno-Ugric and Romance, the best NMT system outperformed the best phrase-based SMT system for all language directions to English (Toral and Sánchez-Cartagena, 2017). The NMT systems produced fluent and more accurate inflections and word order but performed poorly when translating long sentences (Vanmassenhove et al., 2019; Bentivogli et al., 2016). Morphologically rich languages have also been shown to have more fluent outputs (Klubička et al., 2017; Toral and Sánchez-Cartagena, 2017; Popović, 2018). These studies have used metrics such as BLEU (Papineni et al., 2002b), HTER (Snover et al., 2006), TTR (DARLEY, 1959), YULE (Yule, 2014) to evaluate the output of MT models for fluency and adequacy. MT was also evaluated for discuss phenomena where the model's capacity to exploit linguistic context for co-reference and coherence was evaluated (Bawden et al., 2018).

Evaluation frameworks like HOPE have also been used for task-oriented and human-centric evaluation for machine translation outputs (Gladkoff and Han, 2021). This framework uses professional post-editing annotations and contains commonly occuring error types and error penalty points. The error penalty points use geometric progression to reflect the severity level of errors for each translation unit.

4 Dataset

We use the Yorùbá Bible *Bíbélì Mímó ní Èdè Yorùbá Òde-Òní* (BMEYO) and the New International Version (NIV) English Bible.

4.1 Bible Data

We crawled the BMEYO Yorùbá Bible from the public website Biblica.[3] This version is the modern translation of the Yorùbá Bible and is the closest equivalent to the English NIV Bible, according to Biblica. We thus use the NIV English translation with the Bible to create our parallel data. We organize the Bible according to their verses. In Table 1, we show an example verse from our Bible dataset.

4.2 Data Preprocessing

The Yorùbá Bible is based on old Bible manuscripts. A total of 16 verses which were included in early English Bible versions like the King James Bible but which were omitted from later versions are still part of the Yorùbá Bible.

[3] https://www.Bible.com/versions/911-ycb-bibeli-mim-ni-ede-yoruba-ode-oni.

Data	Scripture
BMEYO	Ìgbà láti pa àti ìgbà láti mú láradá ìgbà àti wó lulẹ̀ àti ìgbà láti kọ̀
NIV	a time to kill and a time to heal, a time to tear down and a time to build.

Table 1: A description of Ecclesiastes Ch. 3, Verse 3 for the BMEYO. **(NIV English Bible translation)**

Data	#TOK	#SENT	#TTR
BMEYO	793,870	38,149	25.5

Table 2: A statistical description of the Yorùbá (source) data. #TOK refers to number of tokens, #SENT is number of sentences, and TTR is type token-ratio.

Therefore, we start data preprocessing by adding these missing verses into our English NIV Bible to make it equivalent with the Yorùbá text. Those verses were footnotes in the English NIV. In addition, the book of Third John has 15 verses in the Yorùbá Bible but 14 verses in the English NIV. The 15^{th} verse in the Yorùbá is a part of 14^{th} verse in the English Bible. As a result, we combine verse 15 into verse 14, as it is in the English NIV. The aligned dataset can be found on GitHub at **https://github.com/UBC-NLP/COLING2022**

Next, we tokenized the English data using `SpaCy`[4]. `SpaCy` currently does not provide a tokenization package for Yorùbá, so we used the whitespace tokenizer for all the Yorùbá data. We use python scripts to ensure the punctuations is appropriately tokenized. Next, we convert all words to lowercase in order to alleviate data sparsity.

We also split words using Byte Pair Encoding (BPE) (Sennrich et al., 2015). In low-resource settings, large vocabularies result in the representation of low-frequency (sub)words as BPE units which affects the ability to learn good high-dimensional representations (Barone et al., 2017). Thus, we choose smaller merge operations and varied the number of merge operations from $10,000$ to $30,000$. Finally, we split the dataset into training, validation, and test sets using an 80%-10%-10% standard split.

5 Methods

We train 3 sentence level models: SMT, BiLSTM, and Transformer models to translate from Yorùbá to English. We choose English as our target language because we are interested in analyzing how ambiguous BNs in Yorùbá are translated into English where (i.e., in English) nouns are typically marked both for number and determinacy. The hyperparameters and training procedure are described in the next subsections.

5.1 SMT

We use the `Moses`[5] statistical translation system for our SMT model. We apply tokenization, true casing, and perform word alignment on the parallel data using GIZA++ (Och and Ney, 2003). The word alignments were used to extract phrase-paired translations and calculate probability estimates (Koehn et al., 2007). We used KENLM (Heafield, 2011) to train and query a LM for English. KENLM is a library implemented for efficient language model queries, reducing both time and memory costs, and is integrated into `Moses`. The decoder uses this LM to ensure a fluent output of the target language, in our case, English. We used the validation sets of our parallel data for the final tuning process just before we perform the blind testing.

5.2 BiLSTM

We use a Sequence to Sequence (Seq2Seq) BiLSTM with attention model. Our best BiLSTM model has an embedding layer with $1,024$ dimensions, and 2 encoder and decoder layers each.[6] We use the Adam optimizer with a learning rate of $5e{-}4$ and a batch size of 32. For regularization, we use a dropout of 0.2.

Hyperparameter	Values
encoder layers	2, 3, 4, 6, 8
decoder layers	2, 3, 4, 6, 8
attention heads	4, 8, 16
embedding dimension	256, 512, 1024.
batch sizes	32, 64, 128
number of tokens	4000, 4096
dropout	0.2, 0.3, 0.4, 0.6, 0.8

Table 3: Hyperparameter settings for tuning BiLSTM and Transformer models.

5.3 Transformer

For the transformer model, we use 5 layers with 8 attention heads in both encoder and decoder. We

[4] https://spacy.io/

[5] https://github.com/moses-smt

[6] Model architecture and hyperparameter values are identified on validation data using values listed in Table 3.

use embedding dimension with $1,024$ units. We express our batch size in number of tokens, and set it to $4,096$. Detailed hyperparameter settings is available in Table 4

Hyperparameter	Values
adam-betas	(0.9,0.98)
clip-norm	0.0
learning rate	5e-4
learning rate scheduler	inverse square root
warmup-updates	4000
dropout	0.3
weight-decay	0.0001
criterion	label smoothed cross entropy
label-smoothing	0.5
encoder-layerdrop	0.2
decoder-layerdrop	0.2

Table 4: Hyperparameters for Transformer model

5.4 Hyperparameters, Vocab. & Training

Hyperparameters. We experimented with different hyperparameter values to ensure optimization of our models. Since the size of data is small, we used fewer layers, and smaller batch sizes as referenced in literature (Sennrich and Zhang, 2019; Nguyen and Chiang, 2017). A full range of hyperparameters and values are in Table 3.

Vocabulary Size. We varied the number of BPE merge operations from $10,000$ to $30,000$. Our optimal models used 10K merge operations.

Training. We train the model with fairseq toolkit for 7 days, on 1 GPU, and choose best epoch on our development set, reporting performance on TEST.

6 Evaluation

In this section, we first evaluate using BLEU. We provide details of this part of the evaluation in Section 6.1. Next, we provide details on our approach to evaluating BNs in Section 6.2.

6.1 Model Performance

We evaluate the output of our models using BLEU (Papineni et al., 2002b). We use BLEU score because it is the most commonly used metric for MT evaluation. Table 5 shows our n-gram precision figures from 1 to 4-grams and our overall BLEU score. Our transformer model outperforms the BiLSTM and SMT models respectively. We give examples of the output from each model and the gold data in Table 6.

SMT

1-gram	2-gram	3-gram	4-gram	BLEU
62.00	47.97	38.68	31.92	33.01[7]

BiLSTM

1-gram	2-gram	3-gram	4-gram	BLEU
62.97	49.21	40.08	33.43	33.44[8]

Transformer

1-gram	2-gram	3-gram	4-gram	BLEU
66.27	**53.42**	**44.45**	**37.68**	**37.93**[9]

Table 5: Model performance at 1-4 grams with final column showing BLEU score as given by the sacrebleu (Post, 2018).

Gold Data	SMT	BiLSTM	Transformer
as in the days when you came out of egypt, i will show them my wonders.	like the day she came up out of egypt, i will show wonders known him.	as soon as he came out of egypt, i will show him all the wonders he has done.	like the day he came up out of egypt, i will show him the wonders.
rescue me from the mouth of the lions; save me from the horns of the wild oxen.	the lord . deliver my life from the lions; rescue me from the horns of the wild ox.	deliver my life from the lion's hands ; deliver me from the hand of the wild ox.	rescue my life from the lion; rescue me from the horns of the wild ox.
so now the lord has put a lying spirit in the mouths of all these prophets of yours. the lord has decreed disaster for you.	so the lord has put a lying spirit in the mouths of all your prophets. the lord has decreed disaster for you.	so the lord has put a lying spirit in the mouths of all these prophets of yours. the lord has decreed disaster for you.	so the lord has put a lying spirit in the mouths of all these prophets of yours. the lord has decreed disaster for you.

Table 6: A comparison of the gold data with the output of our models.

6.2 Preparing a Gold Standard for Bare Noun Disambiguation

We group English translations of Yorùbá BNs into 5 categories described in Table 7: *generic*, *indefinite singular*, *indefinite plural*, *definite singular* and *definite plural*. We do this both for the gold standard target data and for the output of our MT systems, and compare the distribution of categories in the MT output to the category distribution in

the English gold standard data. The annotation was performed by a linguist who is also a native speaker of Yorùbá.

Category	Description
Generic	noun refers to a kind or class of individuals.
Indefinite Singular	noun refers to a single non-specific object
Indefinite Plural	noun refers to multiple non-specific objects
Definite Singular	noun refers to a single specific object
Definite Plural	noun refers to multiple specific objects

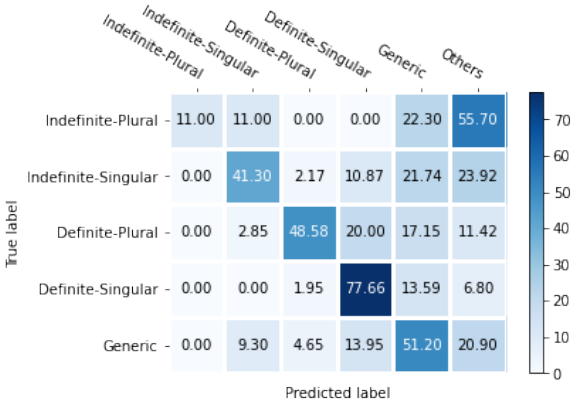
Table 7: Noun Categories when translating from Yorùbá to English

Determining the correct category for the translation of a BN requires us to control for the following contextual information: **type of verb** (STATIVE versus EVENTIVE); **the discourse context** (e.g. the preceding sentence, the familiarity constraint); and **real-world knowledge** (which may trigger the uniqueness constraint). E.g. Yorùbá words (ọba *king*, ọlọ́run *god*) should always be translated into definite singulars because Yorùbá speakers commonly use them to refer to specific entities.

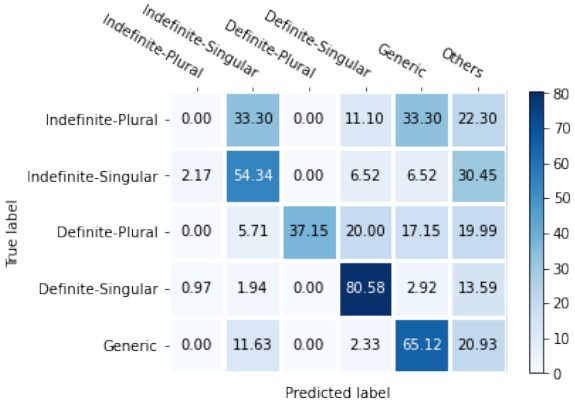
6.3 Evaluation of Bare Noun Disambiguation

To measure how well our models translate BNs, we arbitrarily select 100 sentences and evaluate the percentage of correct BN translations for each model. The 100 sentences selected contained 236 occurrences of BNs. There were 9 indefinite plural, 46 indefinite singular, 35 definite plural, 103 definite singular and 43 generic examples in this randomly selected set. We perform a detailed error analysis on the 100 sentences for each case of incorrect BN translation.

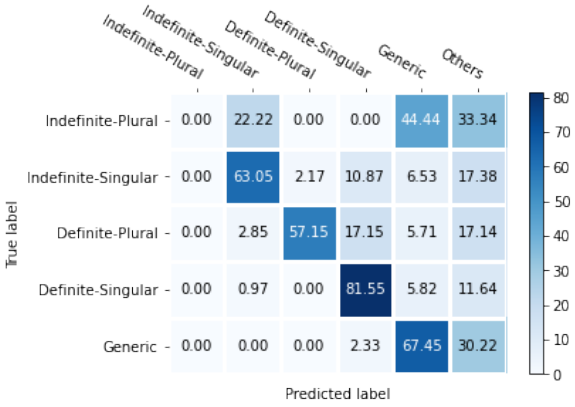
We assemble the type of errors found in the translation output into a confusion matrix that captures the behaviour of the 3 models in Figure 1. Each row in the confusion matrix represents one of our 5 noun categories and each row adds up to 100%.The dark blue coloured boxes represent the correctly translated values in percentages. The light blue boxes are the incorrectly translated categories that fell within one of the 5% categories. The column "others" contain incorrectly translated choices outside the 5 categories. We discuss these type of errors in detail in 6.4.



(a) SMT



(b) BiLSTM



(c) Transformer

Figure 1: Confusion matrices showing the disambiguation patterns of the SMT, BiLSTM, and Transformer Models.

The results show that the transformer model outperforms the SMT and BILSTM models for all categories except the indefinite plural category where the SMT outperforms the Transformer model. The BiLSTM also outperforms the SMT for 3 categories.

How the models handle indefinites. In the context of indefinites, we found that all 3 models achieved low accuracy scores. We assume this

to be the case because fewer cases of indefinites are often reported in languages like Yorùbá that lack overt definite and indefinite markers and this will be represented in the data we used for training. Languages like Yorùbá that do not have overt definite and indefinite determiners are ambiguous between definite and indefinite readings. Indefinites are blocked when the common ground establishes the uniqueness, or familiarity of the set denoted by the noun (Dayal and Sağ, 2019). It is expected that indefinites can occur everywhere else. However, in texts such as these, context introduces either familiarity or uniqueness therefore reducing the occurrence of indefinites.

How the models handle definites. There were 138 BN instances that translated as definites in the set we evaluated. This is due to the aforementioned system of disambiguation that assigns BNs to the definite class if they occur in an environment of uniqueness or familiarity. There are also certain words that have an inherent unique meaning due to cultural beliefs of Yorùbá people. Titles such as ọba *king*, olúwa *lord* and many more often have a definite translation because it is culturally believed for instance that only one king can rule a domain. The models show an improvement in translating definites when compared to indefinites.

How the models handle generics. The Transformer model outperforms the BiLSTM and SMT models for this category while the BiLSTM outperforms the SMT.

6.4 Error Analysis

We focus on other errors, that is those errors presented in the "others" category/column in Figure 1. For this class of errors, the wrongly translated BN did not translate into one of definite singular or plural, indefinite singular or plural, generic or BN category. We therefore evaluate the errors found in the sentences. For the sentences evaluated, we focus only on errors that involve nouns and determiners and ignore errors relating to other classes of words. This means that if a sentence had errors with, for example, verbs or adjectives, we ignored these errors. We categorize the errors we found and provide examples errors for each category in the SMT and NMT models in Table 8. We **bold** face relevant words and phrases occurring in the gold data that had errors in the model output.

Missing word. Outputs with missing nouns or determiners are classified under this category.

Wrong word or spelling. Wrong use of determiners or incorrect nouns belong to this category. We also classify wrong spellings, poorly inflected forms of the noun and unknown words under this category.

Grammaticality. We categorize both syntactic and semantic errors here. Wrong tense or aspect, incorrect number, poor punctuation, and lack of coherence are categorized under this class.

Wrong word order or category. If the order of the noun and determiner is wrong, even if the determiner and nouns are correct, we classify this as an error. We also include instances where the order of the noun and determiner occurs inappropriately, either before the verb or any other category. We found that some instances require both word level and phrase level order errors. In the case of word level order errors, we can generate a correct sentence by moving individual words, independently of each other, whereas for a phrase level order error, blocks of consecutive words should be moved together to form a right translation. In addition, this category includes instances in which a different type is used instead of a noun or determiner.

In evaluating each category, we do not consider correct synonyms as errors.

6.5 Word-Level BN Disambiguation

We randomly select 10 nouns that occur as BNs in the test data and check how well the 3 models disambiguate these words. We use the following words: ìlú *town*, ìwé *book*, ilé *house*, ọkùnrin *male*, obìnrin *female*, aṣọ *clothing*, ilẹ̀ *land*, ọba *king*, àlùfáà *priest* and baba *father*.

We then calculate the percentage of correct translations of these BNs in the test set. We use the English gold data to determine the correct disambiguation and compare each instance of the BNs with the corresponding occurrence in the gold data.

Our analysis in Figure 2 shows that the transformer model performs better in disambiguating the BNs selected. The transformer model achieves **67.85%** accuracy while the BiLSTM model achieves an accuracy of **63.83%**. The SMT model, on the other hand, achieves an accuracy of **59.12%**.

Error	Model	Gold	Model Output
Missing word	SMT	with each bull prepare a grain offering of three-tenths of **an ephah** of fine flour who shut up the sea behind doors when it burst forth from **the womb**,	with each bull prepare a drink offering with three-tenths of fine flour or who shut the doors of the sea, when he flow back as if he had the,
	BiLSTM	as in **the days** when you came out of egypt, i will show them my wonders. and he inserted the poles into **the rings** on the sides of the ark to carry it.	as soon as he came out of egypt , i will show him all the wonders he has done . then he put the poles on each side of the ark to put it on the chest .
	Transf	now hiram had sent to **the king** 120 talents of gold. each day one ox, six choice sheep and **some poultry** were prepared for me,...	and hiram sent him 120 talents of gold each day an ox, a choice sheep and six days they provide for me...
Wrong word / spelling	SMT	everyone who quotes **proverbs** will quote this **proverb** about you: like mother, like daughter ...he who had received **the promises** was about to sacrifice his one and only son,	all those who were powe, it will this powe you: as mothers, so his son of woman. ...he who receive **the promised** almost ready to take your son into one sacrifice.
	BiLSTM	the lot settles **disputes** and keeps strong opponents apart. he spoke, and there came swarms of **flies**, and gnats throughout their country.	lot lays up the battle and makes up two unchange as from each other . he spoke, and the kind of reitution came and became gnats in their land
	Transf	the lot settles **disputes** and keeps strong opponents apart. with **an opening** in **the center** of **the robe**...	the snow finish quarreling and two oppose each other with **the holes** among **the belt**...
Grammaticality	SMT	this is **a decree** for israel, an ordinance of the god of jacob. **the simple** inherit folly, but **the prudent** are crowned with knowledge.	this is a lasting ordinance for israel, and the law of the god of jacob. a simple inherit folly but to be wise in the crown of knowledge .
	BiLSTM	as for **the donkeys** you lost three days ago, do not worry about them; they have been found... and we know that in all things god works for the good of those who love him ...	for **a donkeys** he was three days ago, not terrified by them; they were found to be found We know that everything is in good deeds for those who love god,...
	Transf	there will be **a highway** for the remnant of his people ... overlay the frames with gold and make gold rings to hold **the crossbars.**	good **ways** will be for his people,... overlay the frames with gold and make gold rings so they can be **crossbars**...
Wrong word order	SMT	by faith abraham, when god tested him, offered **isaac** as a sacrifice. ...the next day the south wind came up,	by faith of abraham, when he was tempted to, isaac offer sacrifices, ...the next day, forth the south wind began to blow,...
	BiLSTM	...i will give my daughter acsah in **marriage** to the man who attacks and captures kiriath sepher. do not love sleep or you will grow poor; stay awake and you will have food to spare	...i'll give my daughter acsah to a man who struck down kiriath sepher and took him **wedding**. do not love a sleep, or you will be poor. do not sleep and have food to give you something to eat .
	Transf	about this time next year, elisha said, you will hold ***a son*** *in your arms.* ...israel served to get a wife, and to pay for her he tended sheep.	elisha said, this time is coming, you will take *your hand for your son.* ...israel worshiped as a wife and took care of the meat to pay **the bride** for money.

Table 8: Example of errors for the three models under each error category we described. The English Gold is the NIV translation for the Yorùbá Source

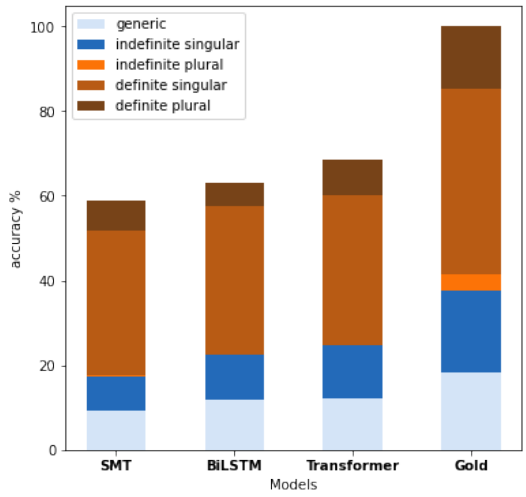


Figure 2: Distribution of disambiguation patterns in our models. Our gold data has 100% in disambiguation.

We also found the transformer outperforming the SMT and BiLSTM models in correctly translating BNs. We also found that all 3 models best performed in translating a BN in Yorùbá into an definite singular in English. This finding corroborates research that predicts that languages which lack overt definite and indefinite markers have larger cases of definites, and findings within the MT community that MT models improve with more data. We further analyzed the type of errors our systems produce. We identified cases of missing words, wrong word or spellings, grammaticality issues, and word-order errors. We found that even when certain BNs have been correctly categorized by the models, the models still had semantic, and or syntactic errors.

7 Conclusion

In this work, we showed how SMT, BiLSTM, and transformer models translate BNs in Yorùbá, a resource scarce language. We compared the ability of SMT and NMT models to correctly translate BNs into various categories referenced in the syntax literature of Yorùbá. We measured the performance of the MT models and the output using BLEU scores, and by counting the percentage of correctly disambiguated BNs compared with incorrectly disambiguated BNs. We found a positive correlation between disambiguation accuracy, and BLEU scores as well as a positive correlation between number of occurrences of a category and the accuracy in translation.

In order to further probe the capacity of SMT and NMT models in disambiguating BNs, further work can be carried out to improve the SMT and NMT models and perform human based evaluations on the entire quality of the MT output. We can improve the MT systems with back-translation, cross-lingual word embeddings, increasing the size of data used for training, transfer learning, among other approaches. Standard test sets can also be developed to aid automatic comparison of human evaluations and machine based evaluations. In addition, the Yorùbá data we used for this work is a translated document; a translation from English to Yorùbá and it will be interesting to use a text originally written in Yorùbá but translated to English for this experiment.

Acknowledgements

We gratefully acknowledge support from Canada Research Chairs (CRC), the Natural Sciences and Engineering Research Council of Canada (NSERC; RGPIN-2018-04267), the Social Sciences and Humanities Research Council of Canada (SSHRC; 435-2018-0576; 895-2020-1004; 895-2021-1008), Canadian Foundation for Innovation (CFI; 37771), Compute Canada (CC),[10] UBC ARC-Sockeye,[11], Advanced Micro Devices, Inc. (AMD), and Google. Any opinions, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of CRC, NSERC, SSHRC, CFI, CC, AMD, Google, or UBC ARC-Sockeye.

References

Barbara Abbott. 2006. Definite and indefinite. *Encyclopedia of language and linguistics*, 3(392):99.

Antonio Valerio Miceli Barone, Jindřich Helcl, Rico Sennrich, Barry Haddow, and Alexandra Birch. 2017. Deep architectures for neural machine translation. *arXiv preprint arXiv:1707.07631*.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2017. Evaluating discourse phenomena in neural machine translation. *arXiv preprint arXiv:1711.00513*.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313, New Orleans, Louisiana. Association for Computational Linguistics.

Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2016. Neural versus phrase-based machine translation quality: a case study. *arXiv preprint arXiv:1608.04631*.

Franck Burlot and François Yvon. 2017. Evaluating the morphological competence of machine translation systems. In *Proceedings of the Second Conference on Machine Translation*, pages 43–55.

Greg N Carlson. 1989. On the semantic composition of english generic sentences. In *Properties, types and meaning*, pages 167–192. Springer.

Marine Carpuat. 2009. One translation per discourse. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009)*, pages 19–27.

[10] https://www.computecanada.ca

[11] https://arc.ubc.ca/ubc-arc-sockeye

Lisa Lai-Shen Cheng and Rint Sybesma. 1999. Bare and not-so-bare nouns and the structure of np. *Linguistic inquiry*, 30(4):509–542.

Gennaro Chierchia. 1998. Reference to kinds across language. *Natural language semantics*, 6(4):339–405.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

FREDERIC L. DARLEY. 1959. Certain language skills in children: Their development and interrelationships. *Pediatrics*, 23(4):819–819.

Veneeta Dayal and Yağmur Sağ. 2019. Determiners and bare nouns. *Annual Review of Linguistics*, 6.

Serge Gladkoff and Lifeng Han. 2021. HOPE: A task-oriented and human-centric evaluation framework using professional post-editing towards more effective MT evaluation. *CoRR*, abs/2112.13833.

Annette Rios Gonzales, Laura Mascarell, and Rico Sennrich. 2017. Improving word sense disambiguation in neural machine translation with sense embeddings. In *Proceedings of the Second Conference on Machine Translation*, pages 11–19.

Liane Guillou and Christian Hardmeier. 2016. Protest: A test suite for evaluating pronouns in machine translation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 636–643.

Liane Guillou, Christian Hardmeier, Preslav Nakov, Sara Stymne, Jörg Tiedemann, Yannick Versley, Mauro Cettolo, Bonnie Webber, and Andrei Popescu-Belis. 2019. Findings of the 2016 wmt shared task on cross-lingual pronoun prediction. *arXiv preprint arXiv:1911.12091*.

Christian Hardmeier. 2012. Discourse in statistical machine translation: A survey and a case study. *Discours*.

Kenneth Heafield. 2011. Kenlm: Faster and smaller language model queries. In *Proceedings of the sixth workshop on statistical machine translation*, pages 187–197.

Pierre Isabelle, Colin Cherry, and George Foster. 2017. A challenge set approach to evaluating machine translation. *arXiv preprint arXiv:1704.07431*.

Marcin Junczys-Dowmunt, Tomasz Dwojak, and Hieu Hoang. 2016. Is neural machine translation ready for deployment? a case study on 30 translation directions. *arXiv preprint arXiv:1610.01108*.

Filip Klubička, Antonio Toral, and Víctor M Sánchez-Cartagena. 2017. Fine-grained human evaluation of neural versus phrase-based machine translation. *The Prague Bulletin of Mathematical Linguistics*, 108(1):121–132.

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.

Manfred Krifka. 2003. Bare nps: kind-referring, indefinites, both, or neither? In *Semantics and linguistic theory*, volume 13, pages 180–203.

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

Richard K Larson. 1985. Bare-np adverbs. *Linguistic inquiry*, 16(4):595–621.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of lstms to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Vivien Macketanz, Eleftherios Avramidis, Shushen Manakhimova, and Sebastian Möller. 2021. Linguistic evaluation for the 2021 state-of-the-art machine translation systems for German to English and English to German. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1059–1073, Online. Association for Computational Linguistics.

Ruslan Mitkov. 1999. Introduction: special issue on anaphora resolution in machine translation and multilingual nlp. *Machine translation*, pages 159–161.

Toan Q Nguyen and David Chiang. 2017. Transfer learning across low-resource, related languages for neural machine translation. *arXiv preprint arXiv:1708.09803*.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1):19–51.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002a. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002b. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Maja Popović. 2011. Hjerson: An open source tool for automatic error classification of machine translation output. *The Prague Bulletin of Mathematical Linguistics*, 96(1):59–67.

Maja Popović. 2018. Language-related issues for nmt and pbmt for english–german and english–serbian. *Machine Translation*, 32(3):237–253.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Craige Roberts. 2003. Uniqueness in definite noun phrases. *Linguistics and philosophy*, 26(3):287–350.

Bertrand Russell. 1905. On denoting. *Mind*, 14(56):479–493.

Rico Sennrich. 2016. How grammatical is character-level neural machine translation? assessing mt quality with contrastive translation pairs. *arXiv preprint arXiv:1612.04629*.

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

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Controlling politeness in neural machine translation via side constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 35–40.

Rico Sennrich and Biao Zhang. 2019. Revisiting low-resource neural machine translation: A case study. *arXiv preprint arXiv:1905.11901*.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*, volume 200. Cambridge, MA.

Antonio Toral and Víctor M Sánchez-Cartagena. 2017. A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions. *arXiv preprint arXiv:1701.02901*.

Eva Vanmassenhove, Dimitar Shterionov, and Andy Way. 2019. Lost in translation: Loss and decay of linguistic richness in machine translation. *arXiv preprint arXiv:1906.12068*.

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. Context-aware neural machine translation learns anaphora resolution. *arXiv preprint arXiv:1805.10163*.

Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F Wong, and Lidia S Chao. 2019. Learning deep transformer models for machine translation. *arXiv preprint arXiv:1906.01787*.

C Udny Yule. 2014. *The statistical study of literary vocabulary*. Cambridge University Press.