NATURAL LANGUAGE UNDERSTANDING FOR AFRICAN LANGUAGES

Pierrette Mahoro Mastel[1], Aime Munezero[1], Ester Namara[1], Richard Kagame[1], Zihan Wang[2], Allan Anzagira[3], Akshat Gupta[3], and Jema David Ndibwile[1]

[1]Carnegie Mellon University
[2]Columbia University
[3]JP Morgan Chase

ABSTRACT

Natural Language Understanding(NLU) is a fundamental building block of goal-oriented conversational AI. In NLU, the two key tasks are predicting the intent of the user's query and the corresponding slots. Most NLU resources available are for high-resource languages like English. In this paper, we address the limited availability of NLU resources for African languages, most of which are considered Low Resource Languages(LRLs), by presenting the first extension of one the most widely used NLU dataset, the Airline Travel Information Systems (ATIS) dataset to Swahili, Kinyarwanda. We perform baseline experiments using BERT,mBERT, RoBERTa, XLM-RoBERTa under zero-shot settings and achieve promising results. We release the datasets and the annotation tool used for the utterance slot labeling to the community to further NLU research on NLU for African Languages.

1 INTRODUCTION

Natural language understanding (NLU) is a fundamental building block of goal-oriented dialogue systems such as Amazon's Alexa, Apple's Siri, Google Assistant, and Microsoft's Cortana (Krishnan et al., 2021b). Two key problems in Natural Language Understanding are predicting what the user intends to do (the intent) as well as the arguments of the intent (the slots) from a user's query (Tur et al., 2010; Rastogi et al., 2017; Jung, 2019). For example, a user might want to 'find a restaurant' given the slot labels 'type_of_food' and 'location' or they may want to 'book a flight' given the slot labels 'airport' and 'location'. The problem from this task can be understood as a joint sentence classification (for intent classification) and sequence labeling (for slot detection) task.

Several models can be trained to achieve high accuracy on both tasks but they require large amounts of labeled data in the target language (Wang et al., 2018b; Wu et al., 2020; Devlin et al., 2018). Such a requirement sets a limitation for low-resource languages(LRLs). LRLs are understood to have scarce resources hence being less studied and under-represented in the majority of Natural Language Processing(NLP) systems. African languages make up more than 2000 of the world's spoken languages (Campbell, 2008) but are barely represented in the 20 languages covered in the majority of NLP research (Magueresse et al., 2020) hence accounting for a big part of LRLs.

In this paper, we take the first step towards advancing the representation of African languages in goal-oriented dialogue systems by creating NLU resources, using the ATIS dataset Price, 1990. To achieve this, we extend the ATIS dataset by creating translations of both the train and test sets from English to Swahili,Kinyarwanda and Luganda. To facilitate the annotation process, we develop an annotation tool that follows the BIO format for slot filling and can be used for any language. We then only create slots for the test sets of Swahili and Kinyarwanda, and subsequently conduct baseline experiments BERT, RoBERTa, mBERT, and their multilingual versions XLM-RoBERTa, under zero-shot and code switch experiments.

The rest of this paper is organized as follows. Section 2 discusses NLU related works, section 3 gives an overview of the African languages (Swahili and Kinyarwanda) and details the steps we took to expand the ATIS dataset to these languages. Sections4 provide a description of the models tested, the different experiments done, and the model performance. Section 5 concludes the paper.

2 Related work

NLU datasets: As a crucial part of goal oriented dialogue systems, NLU has been a center of several research studies since the 1990 creation of the ATIS dataset (Price, 1990). Over the years several more NLU datasets have been released (Coucke et al., 2018; Schuster et al., 2019) as well as multilingual extensions of the ATIS datasets (Upadhyay t al., 2018; Xu,Hauder,Mansour, 2020). But to the best of our knowledge there are no open source NLU datasets that include any African Language.

NLU models: Studies have been conducted on various natural language understanding (NLU) models, including Bert (Devlin et al., 2018), HNN (He et al., 2019), DSSM (Wang et al., 2019), RoBERTa (Liu et al., 2019b), and they have been evaluated on different datasets, including MultiATIS++ (Xu et al., 2020a), GLUE (Wang et al., 2018a)), and LexGLUE (Chalkidis et al., 2021). These studies have been carried out in a range of settings, including multitasking (Liu et al., 2019a), zero-shot learning (Yang et al., 2022), cross-lingual (Xu et al., 2020b; Gupta et al., 2021a), semi-supervised (Zhu et al., 2020), (Gupta et al., 2021b), and unsupervised learning (Arava et al., 2021). Zero-shot learning (ZSL) (Yang et al., 2022) and cross-lingual (Xu et al., 2020b; Gupta et al., 2021c) approaches have gained recognition as effective strategies for natural language understanding in low-resource languages. The scarcity of labeled data is a common issue in these types of NLU tasks, but ZSL (Yang et al., 2022) provides a solution by allowing models to generalize to unseen classes using knowledge transferred from seen classes, potentially reducing the need for extensive labeled data. Additionally, cross-lingual methods allow leveraging resources from high-resource languages to improve performance in low-resource languages, such as transferring learning from models trained on high-resource languages to low-resource languages.

3 Data

3.1 African Languages of Interest

Swahili and Kinyarwanda were selected for annotation of intent and slot labels selected based available resources. These languages are spoken by over 100 million in East Africa. Due to time and resource constraints, we only annotate the test sets of Swahili and Kinyarwanda for NLU tasks.

Kinyarwanda is a tonal languages that is the official and native language of Rwanda. Kinyarwanda alphabet comprises of the 24 out of the 26 Latin characters. It excludes q and x. Additionally there are what are termed as complex consonants which are a combination of consonants like kw, mbw and nyw [1]. There are more than 12M Kinyarwanda speakers.

Swahili is one the most widely spoken languages in Africa with over 90M (Campbell (2008)). It is one of the official languages of the East African Community and the African Union. The Swahili alphabet comprises of 30 letters, including 24 Latin characters (it also excludes q and x) and six additional consonants (ch, dh, gh, ng', sh, th).

3.2 Creating African-ATIS Dataset

The ATIS collection was developed to support the research and development of organized understanding systems. It consists of audio recordings of human speakers interacting with either partially or fully automated ATIS systems. The audio is transcribed and annotated with information such as the intent of the speaker and entities mentioned in the utterance (e.g. flight numbers, departure, and arrival cities). We extended this dataset to Swahili and Kinyarwanda[2]. We achieved this by translating the train and test English utterances into the three languages using Google Cloud Translation API. Initially, we had the expectation that using Google API would result in accurate translations,

[1] https://bit.ly/3wYpdka

[2] https://github.com/hannawong/ZeroShot_CrossLingual/tree/master/data

thereby saving us time. However, we discovered that this was not the case, particularly when working with low-resource languages that lack sufficient training data. As a result, a large proportion of the translations were found to be inaccurate as seen in table 1. Consequently, we were compelled to engage the services of native translators to re-translate the sentences that were not accurately translated as well as correct the errors.

With the translations and correction of errors, we usually observed a one-to-many mapping between English and some of the target languages. For example - ”9 am” translates to ”saa tatu zamugitondo” in Kinyarwanda. This, therefore, leads to a change in slot labels in the target language and requires the annotation of the slot labels to be adjusted to fit the translated utterance for better results. To simplify the annotation process, an open-source annotation tool was created, which can be accessed through the website [3]. The tool was used to annotate the test sets for Swahili and Kinyarwanda languages using BIO tagging.This tool facilitates the annotation process by allowing the entire dataset to be uploaded, and subsequently, the data is automatically mapped and restructured. This eliminates the need for researchers to perform these tasks manually, thereby saving them time and effort. Additionally, the annotation tool can be used to annotate any target language, with English being the original language. A Figure showing the annotation tool is illustrated in Appendix A, while Table 2 provides a sample of English slot labels and their corresponding slot labels in Kinyarwanda and Swahili.

Table 1: *Examples of translated sentences using Google Cloud Translation API in Kinyarwanda and Swahili and their corrected versions by native translators*

ENGLISH	
Utterance	which flights on us air go from orlando to cleveland
KINYARWANDA	
API translation	izo ndege kuri twe ikirere ziva muri orlando zerekeza muri cleveland
Native translators	Ni izihe ndege za US AIR ziva Orlando zerekeza cleveland
SWAHILI	
API translation	ambayo ndege juu yetu huenda kutoka orlando hadi cleveland
Native translatorsn	Kuna safari gani za za shirika la ndege la US Air kutoka Orlando to cleveland

Table 2: *Example of English slot labels and their equivalent slot labels in Kinyarwanda and Swahili*

english							
Words	find	a	flight	from	memphis	to	tacoma
Slots	O	O	O	O	B-fromloc.city_name	O	B toloc.city_name
kinyarwanda							
Words	shaka	indege	iva		memphis	yerekeza	tacoma
Slots	O	O	O		B-fromloc.city_name	O	B-toloc.city_name
swahili							
Words	tafuta	ndege	kutoka		memphis	hadi	tacoma
Slots	O	O	O		B-fromloc.city_name	O	B-toloc.city_name

The ATIS English dataset originally consisted of 4978 training and 893 test samples. We utilized Google Translation API to translate the dataset into Swahili and Kinyarwanda which were later corrected by human experts. However, only the test set was annotated for intents and slots in Kinyarwanda and Swahili. The statistics of the annotated test set for intent and slot labels are displayed in the table 3.

[3] kwandasl.site

Table 3: *Statistics of annotated test datasets for both Kinyarwanda and Swahili, including the number of sentences, independently labeled slots, and intents. We primarily focus on Semi and non-supervised learning, hence only annotate the test datasets.*

	Annotated Results	
	Swahili	Kinyarwanda
Sentences	893	893
Labelled slots	59	66
Intent	20	20

4 Experimental Setup & Results

4.1 Baseline Models

To assess baseline performance on our dataset, we conduct experiments using four well-known transformer models: BERT, mBERT, RoBERTa, and XLM-RoBERTa.

- **BERT** (Bidirectional Encoder Representations from Transformers) is an NLP model pre-training architecture based on transformer architecture. It was introduced in (Devlin et al., 2018). The model is trained using unannotated text data, with the objective of masked language modeling. BERT employs a bidirectional training strategy, considering both the left and right context for each token in the input sequence, resulting in a more robust model with enhanced contextual word understanding abilities.
- **mBERT (Multilingual BERT)** is a variant of BERT (Xu et al., 2021) designed to support multiple languages. It is trained on 104 languages, including Swahili and Yorùbá, two African languages. mBERT utilizes this multilingual training data to cross-share knowledge between languages, thereby improving its performance on low-resource languages.
- **RoBERTa** (Robustly Optimized BERT Pre-training Approach) is a modification of BERT aimed at addressing some of its limitations. RoBERTa, introduced in (Liu et al., 2019b), is trained on a larger text corpus and employs a dynamic masking strategy during pre-training to produce more varied predictions. It also utilizes larger batch sizes and longer training cycles, making it a more robust method for pre-training NLP models.
- **XLM-RoBERTa** is a multilingual version of RoBERTa (Conneau et al., 2019) that was trained on 100 languages, including Swahili. It uses a similar method to mBERT to provide support for many languages. As a result of its training on a large multilingual corpus, it produces cross-lingual representations that can be used for various NLP tasks across multiple languages, with improved accuracy thanks to its strong pre-training process.

4.2 Results

4.2.1 Zero shot Experiments

These involve training a model on data from one domain and then evaluating the model's performance on data from a completely distinct domain, without any target-domain specific labelled data. The objective of zero-shot learning is to train a model that can generalize to the novel, unseen data, without requiring labeled examples from that domain. In our study, we trained our models on the English ATIS dataset, then evaluated their performance on the low-resourced languages Swahili and Kinyarwanda, and obtained the following results presented in Table 4

In our zero-shot experiments, we observe that mBERT;(Xu et al., 2021), the multilingual language model trained over 104 languages, including one of the African languages being studied, achieves a remarkable performance, likely due to its comprehensive understanding of various languages (Libovický et al., 2020) in comparison to BERT (Devlin et al. (2018)) and RoBERTa (Liu et al., 2019b), which are not multilingual.

Table 4: *A comparison of benchmark models performance on the Swahili and Kinyarwanda test datasets in the Multi Atis format, showing the F1-score for slot predictions and the accuracy for intent classification after 25 training iterations for zero-shot experiments*

Models	Swahili		Kinyarwanda	
	Intent Accuracy	Slot F1-Score	Intent Accuracy	Slot F1-Score
BERT	0.1780	0.5757	0.0470	0.5281
mBERT	0.7997	0.8844	0.2004	0.7997
RoBERTa	0.1578	0.7117	0.1478	0.7384
XLM-RoBERTa	0.6730	0.8762	0.6976	0.6176

4.2.2 Code-Switching Experiments

The practice of switching between two or more languages, or language varieties, in the same sentence is known as code-switching (Sitaram et al., 2019) . Code-switched text has been used to enhance the performance of pretrained language models in tasks including sentiment classification (Prabhu et al., 2016), part-of-speech tagging (Soto & Hirschberg, 2018), dialogue understanding (Krishnan et al., 2021a), etc.

In our experiment, we use code-switching to augment each monolingual training sample in downstream task by generating artificially code-switched data at a chunk-level. More specifically, we select each word chunk bounded by the BIO-tagged slot labels, and then translate each chunk into a random language selected from 107 languages with google translate api [4]. After translation, the BIO-tagged slot labels are recreated for the translated word chunks. We augment each sentence in English training set for $k = 5$ times. Experiment results demonstrate that adding code-switching samples encourages the fine-tuned model to generalize well on new languages that do not appear in the training set.

Table 5: *A comparison of benchmark models performance on the Swahili and Kinyarwanda test datasets in the Multi Atis format, showing the F1-score for slot predictions and the accuracy for intent classification after 25 training iterations for code-switch experiments*

Models	Swahili		Kinyarwanda	
	Intent Accuracy	Slot F1-Score	Intent Accuracy	Slot F1-Score
BERT	0.9372	0.8867	0.7099	0.8056
mBERT	0.9395	0.9117	0.4524	0.8093
RoBERTa	0.9260	0.8426	0.7256	0.8173
XLM-RoBERTa	0.9260	0.9049	0.6842	0.7510

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

5 Conclusion

In this study, we release human translations for Swahili, Kinyarwanda and Luganda of the ATIS dataset and slot filling and intent classification for Swahili and Kinyarwanda. We performed baseline experiments in the zero-shot setting using pre-trained models for two of the datasets (Swahili and Kinyarwanda). In Zero shot settings, XLM-RoBERTa gives the best results for both Swahili and Kinyarwanda while XLM-RoBERTa and RoBERTa provide the best results for Swahili and Kinyarwanda in chunk-wise code switched experiments respectively. The experiments give promising results and we outsource these resource to encourage the research community to use semi-supervised and unsupervised experiments in the case of low-resourced African languages and evaluate their models on our human annotated test sets. In the future, we want to extend this study to several other African languages in order to avail more resources for low-resourced languages.

REFERENCES

Radhika Arava, Matthew Trager, Boya Yu, and Mohamed AbdelHady. Unsupervised testing of nlu models with multiple views. In *EMNLP 2021 Workshop on Evaluations and Assessments of Neural Conversation Systems (EANCS)*, 2021. URL https://www.amazon.science/publications/unsupervised-testing-of-nlu-models-with-multiple-views.

Lyle Campbell. Ethnologue: Languages of the world, 2008.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael J. Bommarito II, Ion Androutsopoulos, Daniel Martin Katz, and Nikolaos Aletras. Lexglue: A benchmark dataset for legal language understanding in english. *CoRR*, abs/2110.00976, 2021. URL https://arxiv.org/abs/2110.00976.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116, 2019. URL http://arxiv.org/abs/1911.02116.

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

Akshat Gupta, Xinjian Li, Sai Krishna Rallabandi, and Alan W Black. Acoustics based intent recognition using discovered phonetic units for low resource languages. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7453–7457. IEEE, 2021a.

Akshat Gupta, Sargam Menghani, Sai Krishna Rallabandi, and Alan W Black. Unsupervised self-training for sentiment analysis of code-switched data. *arXiv preprint arXiv:2103.14797*, 2021b.

Akshat Gupta, Sai Krishna Rallabandi, and Alan W Black. Task-specific pre-training and cross lingual transfer for sentiment analysis in dravidian code-switched languages. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pp. 73–79, 2021c.

Pengcheng He, Xiaodong Liu, Weizhu Chen, and Jianfeng Gao. A hybrid neural network model for commonsense reasoning. *CoRR*, abs/1907.11983, 2019. URL http://arxiv.org/abs/1907.11983.

Sangkeun Jung. Semantic vector learning for natural language understanding. *Computer Speech Language*, 56: 130–145, 2019. ISSN 0885-2308. doi: https://doi.org/10.1016/j.csl.2018.12.008. URL https://www.sciencedirect.com/science/article/pii/S0885230817303595.

Jitin Krishnan, Antonios Anastasopoulos, Hemant Purohit, and Huzefa Rangwala. Multilingual code-switching for zero-shot cross-lingual intent prediction and slot filling. 03 2021a.

Jitin Krishnan, Antonios Anastasopoulos, Hemant Purohit, and Huzefa Rangwala. Multilingual code-switching for zero-shot cross-lingual intent prediction and slot filling. *arXiv preprint arXiv:2103.07792*, 2021b.

Jindrich Libovický, Rudolf Rosa, and Alexander Fraser. On the language neutrality of pre-trained multilingual representations. *CoRR*, abs/2004.05160, 2020. URL https://arxiv.org/abs/2004.05160.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4487–4496, Florence, Italy, July 2019a. Association for Computational Linguistics. doi: 10.18653/v1/P19-1441. URL https://aclanthology.org/P19-1441.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019b. URL http://arxiv.org/abs/1907.11692.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. Low-resource languages: A review of past work and future challenges. *arXiv preprint arXiv:2006.07264*, 2020.

Ameya Prabhu, Aditya Joshi, Manish Shrivastava, and Vasudeva Varma. Towards sub-word level compositions for sentiment analysis of hindi-english code mixed text. 11 2016.

Patti Price. Evaluation of spoken language systems: The atis domain. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*, 1990.

Abhinav Rastogi, Dilek Hakkani-Tur, and Larry Heck. Scalable multi-domain dialogue state tracking. In *Proceedings of IEEE ASRU*, 2017.

Sunayana Sitaram, Khyathi Raghavi Chandu, Sai Krishna Rallabandi, and Alan W Black. A survey of code-switched speech and language processing. *arXiv preprint arXiv:1904.00784*, 2019.

Victor Soto and Julia Hirschberg. Joint part-of-speech and language id tagging for code-switched data. pp. 1–10, 01 2018. doi: 10.18653/v1/W18-3201.

Gokhan Tur, Dilek Hakkani-Tür, and Larry Heck. What is left to be understood in atis? In *2010 IEEE Spoken Language Technology Workshop*, pp. 19–24, 2010. doi: 10.1109/SLT.2010.5700816.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *CoRR*, abs/1804.07461, 2018a. URL http://arxiv.org/abs/1804.07461.

Shuohang Wang, Sheng Zhang, Yelong Shen, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, and Jing Jiang. Unsupervised deep structured semantic models for commonsense reasoning. *CoRR*, abs/1904.01938, 2019. URL http://arxiv.org/abs/1904.01938.

Yu Wang, Yilin Shen, and Hongxia Jin. A bi-model based rnn semantic frame parsing model for intent detection and slot filling. *arXiv preprint arXiv:1812.10235*, 2018b.

Di Wu, Liang Ding, Fan Lu, and Jian Xie. Slotrefine: A fast non-autoregressive model for joint intent detection and slot filling. *arXiv preprint arXiv:2010.02693*, 2020.

Haoran Xu, Benjamin Van Durme, and Kenton W. Murray. Bert, mbert, or bibert? A study on contextualized embeddings for neural machine translation. *CoRR*, abs/2109.04588, 2021. URL https://arxiv.org/abs/2109.04588.

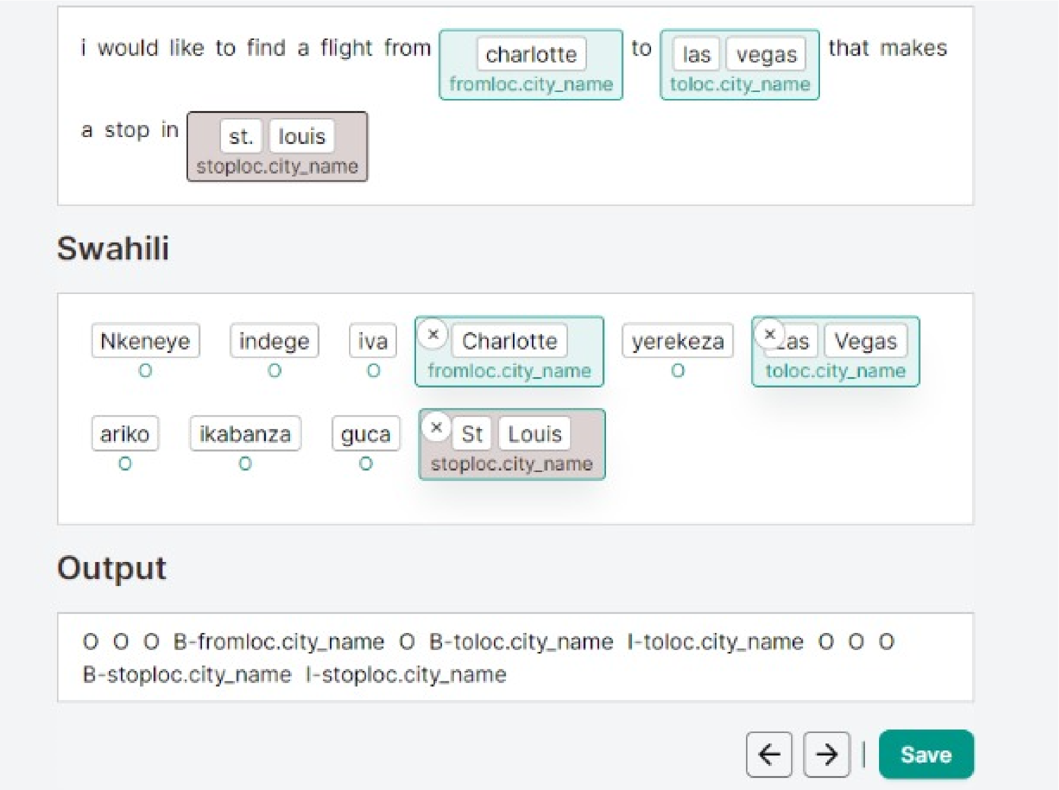
Weijia Xu, Batool Haider, and Saab Mansour. End-to-end slot alignment and recognition for cross-lingual NLU. *CoRR*, abs/2004.14353, 2020a. URL https://arxiv.org/abs/2004.14353.

Weijia Xu, Batool Haider, and Saab Mansour. End-to-end slot alignment and recognition for cross-lingual nlu. In *EMNLP 2020*, 2020b. URL https://www.amazon.science/publications/end-to-end-slot-alignment-and-recognition-for-cross-lingual-nlu.

Ping Yang, Junjie Wang, Ruyi Gan, Xinyu Zhu, Lin Zhang, Ziwei Wu, Xinyu Gao, Jiaxing Zhang, and Tetsuya Sakai. Zero-shot learners for natural language understanding via a unified multiple choice perspective, 2022. URL https://arxiv.org/abs/2210.08590.

Su Zhu, Ruisheng Cao, and Kai Yu. Dual learning for semi-supervised natural language understanding. *CoRR*, abs/2004.12299, 2020. URL https://arxiv.org/abs/2004.12299.

A ANNOTATION TOOL



i would like to find a flight from charlotte fromloc.city_name to las vegas toloc.city_name that makes a stop in st. louis stoploc.city_name

Swahili

Nkeneye O indege O iva O Charlotte fromloc.city_name yerekeza O Las Vegas toloc.city_name ariko O ikabanza O guca O St Louis stoploc.city_name

Output

```
O O O B-fromloc.city_name O B-toloc.city_name I-toloc.city_name O O O
B-stoploc.city_name I-stoploc.city_name
```

Figure 1: *An illustration of the annotation tool showing the procedure of annotating the target language (Swahili) using the original language (English)*