Towards a Crowdsourcing Platform for Low Resource Languages – A Semi-Supervised Approach

Sarah Luger
Orange Silicon Valley
60 Spear Street, Suite 1100
San Francisco, CA 94105
sarah.luger@orange.com

Tapo Allahsera*
Christopher M. Homan
Marcos Zampieri
Rochester Institute of Technology
102 Lomb Memorial Drive
Rochester, NY 14623
{aat3261, cmhvcs, Marcos.Zampieri}@rit.edu

Michael Leventhal
Centre National Collaboratif de l'Education
en Robotique et en
Intelligence Artificielle (RobotsMali)
Kabala, Mli, Mali
textscience@gmail.com

This paper demonstrates how semi-supervised learning and human-in-the-loop crowdsourcing can address machine translation challenges common in low-resource languages.

According to a Facebook study, only 53% of the world's population have access to encyclopedic knowledge in their first language.[1] Just over half of all tweets are in English (Hong, Convertino, and Chi 2011). The implications of this gap in information by language are substantial, particularly when sources of factual information are limited. For instance the United Nations launched the 'Verified' initiative, to combat the adverse effects on public health of false information about COVID-19.[2]

Machine translation is a challenging problem in itself (Korošec 2011; Okpor 2014), and although there have been some great advances made in recent years, particularly in *neural machine translation* (Rapp, Sharoff, and Zweigenbaum 2016; Hirschberg and Manning 2015; Lample et al. 2017), few systems (Martinus and Abbott 2019; Leventhal et al. 2020; Dossou and Emezue 2020) have been trained to communicate in African languages (Heine and Nurse 2000).

This research focuses on Bambara, the most widely-spoken language in the Mande family (Vydrin 2018) in Western Sub-Saharan Africa that includes 60 to 75 languages spoken by 30 to 40 million people. Bambara, the vernacular language of Mali, has approximately 16 million L1 (primary language) and L2 (secondary language) speakers. We attempt to combine the recent success of neural machine translation (NMT), with semi-supervised approaches to generalized machine translation (Cheng 2019). Used alone, NMT appears to be unsuited for under-resourced languages such as Bambara due to the lack of the quantity of labeled data (parallel digital texts) needed. Incorporating semi-supervised learning allows human-computer real-time collaboration in disambiguating word translations.

We designed a crowdsourcing platform that requests the annotator to supply information when the MT model has decision confusion. We have found this situation arises frequently for under-resourced languages. We aim to improve MT models through real-time user feedback. In semi-supervised learning, the computer requests specific help from the crowd when ambiguities are encountered in translating text. Thus, NMT combined with human-in-the-loop crowdsourcing will generate the volume and quality of data over NMT-only approaches. The project tests the hypothesis that a crowdsourcing interface, coupled with semi-supervised human-computer interaction, enables non-professional annotators to contribute data suitable for training an NMT system. Machine translation has the potential to enable Bambara speakers to understand information in their native language, and increase literacy, which, in turn will foster social and economic development. This project will not attempt to prove this hypothesis, but it will contribute results that are needed in order to test it.

Machine translation driven by humans-in-the-loop (González-Rubio, Ortiz-Martínez, and Casacuberta 2012; Carl, Gutermuth, and Hansen-Schirra 2015; Peris, Domingo, and Casacuberta 2017; Alabau et al. 2012) has recently shown great promise (Figure 1). In addition to providing needed data, crowdsourcing is a means to engage the Malian public in a technology project capable of furthering the social and economic development of their country and the development of post-colonial identity built around Malian languages and culture (Mohamed, Png, and Isaac 2020; Tomašev et al. 2020; Orife et al. 2020).

Human computation has been applied only sparingly to low-resourced languages since under-resourced languages bring with them a number of specific challenges, such as non-standard spellings, lack of precise vocabulary for many concepts, lack of a tradition of written, as opposed to oral, expression and the concomitant norms of written expression. Consequentially, there are fewer people with the skills to perform language-based crowdsourcing tasks. In the environments where low-resourced languages are used, crowdsourcing interfaces may be limited to smartphones or other small-screen platforms, as use of computers and high-

*Fullbright Program

[1] http://fbnewsroomus.files.wordpress.com/2015/02/state-of-connectivity1.pdf

[2] https://www.un.org/en/coronavirus/\%E2\%80\%98verified\%E2\%80\%99-initiative-aims-flood-digital-space-facts-amid-covid-19-crisis

bandwidth fixed-line internet connections may be rare.

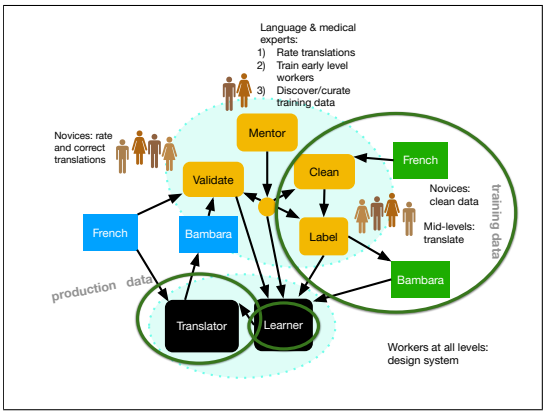

Figure 1: Integration diagram of the proposed system. This work focuses on the green circles.

Our work consists of neural machine translation and crowdsourcing modules, based on JoeyNMT (Kreutzer, Bastings, and Riezler 2019) a light-weight machine translation framework and a crowdsourcing framework of Flask (Ronacher and others 2018) and SQLite (Owens 2006) inspired from previous work of (Kreutzer, Berger, and Riezler 2020) respectively. Figure 2 shows the annotation system.

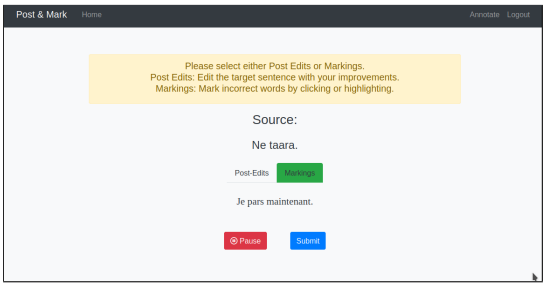

Figure 2: The marking errors interface in the target language (e.g., French). The interface also highlights "Post-Edits".

Our study, Assessing Human Translations from French to Bambara for Machine Learning: a Pilot Study (Leventhal et al. 2020) was presented at the AfricaNLP workshop co-hosted with the International Conference on Learning Representations. The study guided our data collection decisions.

Table 1 shows the results for the Malian news broadcast and the Wikipedia article. The differences in average scores between the news broadcast and the Wikipedia article, aside from the small sampling, most probably reflect the different challenges of the texts. The news broadcast is essentially an oral text and it is easier to reproduce the exact meaning with a more colloquial style. The Wikipedia article has long and complex sentences, making it easier to miss details and nuances while the translator hews closer to the French source and falls back on more formal Bambara.

The translations of the news broadcast showed a limited difference in meaning and use of standard Bambara between written and oral translations, but significant difference in their literalness. The relatively large standard deviations shown in Table 2 indicate a wide range of quality between translators and translations, suggesting that screening translations based on basic quality metrics may be effective.

Malian news	Exact meaning	Literal-ness	Standard Bambara	Highest BLEU Pair
Written	0.83	0.73	0.74	0.4
Oral	0.77	0.53	0.79	0.363
Overall	0.84	0.64	0.76	0.408
Wikipedia				
Written	0.87	0.83	0.83	0.645
Oral	0.53	0.58	0.85	0.377
Overall	0.73	0.76	0.83	0.645

Table 1: Malian news broadcast and Wikipedia article translation ratings and BLEU scores.

Malian news	Exact meaning	Literalness	Standard Bambara
Written	0.181	0.13	0.234
Oral	0.208	0.298	0.178
Overall	0.197	0.243	0.207
Wikipedia			
Written	0.206	0.238	0.171
Oral	0.186	0.211	0.068
Overall	0.22	0.244	0.144

Table 2: Score variance standard deviation

As an example, consider the translations of a French text:

French Objectif réfléchir à de nouvelles stratégies de lutte contre le terrorisme qui continue de faire des victimes dans le sahel.

English Objective to reflect on new strategies to fight terrorism which continues to claim victims in the Sahel.

Written Lajɛ ni kun tun ye ka hakili jakabɔ kɛ fɛɛrɛ kuraw la banbaanciw juguya la miniw bɛ ka ciyɛnni kɛ Saheli kɔnɔnɔnɔna la

Oral A kun tun ye ka miriya kuraw ta ka banbaanciw kɛlɛli sira kan o mun bi ka kɛ sababu ye ka fagali caman kɛ saheli kɔnɔnana

The highest scoring BLEU pairs in all but one of the aligned translations from the news source were between oral and written translation methods. In the one remaining case written-written and written-oral pairs had approximately the same high BLEU scores, the scores being the highest from all the news source translations.

The Wikipedia article translations show that the meaning of the text was captured much better in the written-to-written translations. Except once, the highest scoring BLEU pairs were the written-to-written translations. This suggests that written-to-written translation may be best for more complex texts while oral translations works well on simple texts. Experiment run-time was under three hours; our best performing results were the BPE model with a BLEU score of 17.5.

References

Alabau, V.; Leiva, L. A.; Ortiz-Martínez, D.; and Casacuberta, F. 2012. User evaluation of interactive machine translation systems. In *Proc. EAMT*, 20–23.

Carl, M.; Gutermuth, S.; and Hansen-Schirra, S. 2015. Post-editing machine translation. *Psycholinguistic and cognitive inquiries into translation and interpreting* 115:145.

Cheng, Y. 2019. Semi-supervised learning for neural machine translation. In *Joint Training for Neural Machine Translation*. Springer. 25–40.

Dossou, B. F. P., and Emezue, C. C. 2020. Ffr v1.0: Fon-french neural machine translation.

González-Rubio, J.; Ortiz-Martínez, D.; and Casacuberta, F. 2012. Active learning for interactive machine translation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 245–254.

Heine, B., and Nurse, D. 2000. *African languages: An introduction*. Cambridge University Press.

Hirschberg, J., and Manning, C. D. 2015. Advances in natural language processing. *Science* 349(6245):261–266.

Hong, L.; Convertino, G.; and Chi, E. H. 2011. Language matters in twitter: A large scale study. In *Fifth international AAAI conference on weblogs and social media*. Citeseer.

Korošec, M. K. 2011. Applicability and challenges of using machine translation in translator training. *ELOPE: English Language Overseas Perspectives and Enquiries* 8(2):7–18.

Kreutzer, J.; Bastings, J.; and Riezler, S. 2019. Joey NMT: A minimalist NMT toolkit for novices. *EMNLP-IJCNLP 2019: System Demonstrations*.

Kreutzer, J.; Berger, N.; and Riezler, S. 2020. Correct me if you can: Learning from error corrections and markings.

Lample, G.; Conneau, A.; Denoyer, L.; and Ranzato, M. 2017. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*.

Leventhal, M.; Tapo, A.; Luger, S.; Zampieri, M.; and Homan, C. M. 2020. Assessing human translations from french to bambara for machine learning: a pilot study.

Martinus, L., and Abbott, J. Z. 2019. A focus on neural machine translation for african languages. *CoRR* abs/1906.05685.

Mohamed, S.; Png, M.-T.; and Isaac, W. 2020. Decolonial ai: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology* 1–26.

Okpor, M. 2014. Machine translation approaches: issues and challenges. *International Journal of Computer Science Issues (IJCSI)* 11(5):159.

Orife, I.; Kreutzer, J.; Sibanda, B.; Whitenack, D.; Siminyu, K.; Martinus, L.; Ali, J. T.; Abbott, J.; Marivate, V.; Kabongo, S.; Meressa, M.; Murhabazi, E.; Ahia, O.; van Biljon, E.; Ramkilowan, A.; Akinfaderin, A.; Öktem, A.; Akin, W.; Kioko, G.; Degila, K.; Kamper, H.; Dossou, B.; Emezue, C.; Ogueji, K.; and Bashir, A. 2020. Masakhane – machine translation for africa.

Owens, M. 2006. *The definitive guide to SQLite*. Apress.

Peris, Á.; Domingo, M.; and Casacuberta, F. 2017. Interactive neural machine translation. *Computer Speech & Language* 45:201–220.

Rapp, R.; Sharoff, S.; and Zweigenbaum, P. 2016. Recent advances in machine translation using comparable corpora. *Natural Language Engineering* 22(4):501–516.

Ronacher, A., et al. 2018. Flask (a python microframework). *Dosegljivo: http://flask. pocoo. org.[Dostopano: 20. 7. 2018]*.

Tomašev, N.; Cornebise, J.; Hutter, F.; Mohamed, S.; Picciariello, A.; Connelly, B.; Belgrave, D. C.; Ezer, D.; van der Haert, F. C.; Mugisha, F.; et al. 2020. Ai for social good: unlocking the opportunity for positive impact. *Nature Communications* 11(1):1–6.

Vydrin, V. 2018. Mande languages. *Oxford Research Encyclopedia of Linguistics*.