	4K	8K	14K
Tokens	84,764	170,493	300,648
Types	14,554	23,134	33,288
OOV tokens	4,465 (12.7%)	3,509 (10.0%)	2,965 (8.4%)
OOV types	3,610 (50.3%)	2,950 (41.1%)	2,523 (35.1%)

Table 1: Statistics of the Swahili–English corpora and source-side OOV for 4K, 8K, 14K parallel training sentences.

	4K	8K	14K
Tokens	35,978	71,584	121,718
Types	7,210	11,144	15,112
OOV tokens	3,268 (16.6%)	2,585 (13.1%)	2,177 (11.1%)
OOV types	2,382 (55.0%)	1,922 (44.4%)	1,649 (38.1%)

Table 2: Statistics of the Romanian–English corpora and source-side OOV for 4K, 8K, 14K parallel training sentences.

4 Results

Translation results are shown in tables 6 and 7. We evaluate separately the contribution of the integrated OOV translations, and the same translations annotated with phonetic and semantic features. We also provide upper bound scores for integrated loanword dictionaries as well as for recovering all OOVs.

	4K	8K	14K
Baseline	13.2	15.1	17.1
+ Translit. OOVs	13.4	15.3	17.2
+ Loan OOVs	14.3	15.7	18.2
+ Features	**14.8**	**16.4**	**18.4**
Upper bound loan	18.9	19.1	20.7
Upper bound all OOVs	19.2	20.4	21.1

Table 6: Swahili–English MT experiments.

	4K	8K	14K
Baseline	15.8	18.5	20.7
+ Translit. OOVs	15.8	**18.7**	**20.8**
+ Loan OOVs	**16.0**	**18.7**	20.7
+ Features	**16.0**	18.6	20.6
Upper bound loan	16.6	19.4	20.9
Upper bound all OOVs	28.0	28.8	30.4

Table 7: Romanian–English MT experiments.

Swahili–English MT performance is improved by up to $+1.6$ BLEU when we augment it with translated OOV loanwords leveraged from the Arabic–Swahili borrowing and then Arabic–English MT. The contribution of the borrowing dictionaries is $+0.6$–1.1 BLEU, and phonetic and semantic features contribute additional half BLEU. More importantly, upper bound results show that the system can be improved more substantially with better dictionaries of OOV loanwords. This result confirms that OOV borrowed words is an important type of OOVs, and with proper modeling it has the potential to improve translation by a large margin. Romanian–English systems obtain only small (but significant for 4K and 8K, $p < .01$) improvement. However, this is expected as the rate of borrowing from French into Romanian is smaller, and, as the result, the integrated loanword dictionaries are small. Transliteration baseline, conversely, is more effective in Romanian–French language pair, as two languages are related typologically, and have common cognates in addition to loanwords. Still, even with these dictionaries the translations with pivoting via borrowing/transliteration improve, and even almost approach the upper bounds results.

5 Conclusion

This paper focuses on fully- and partially-assimilated foreign words in the source lexicon—borrowed words—and a method for obtaining their translations. Our results substantially improve translation and confirm that OOV loanwords are important and merit further investigation. In addition, we propose a simple technique to calculate an upper bound of improvements that can be obtained from integrating OOV translations in SMT.

Acknowledgments

This work was supported by the U.S. Army Research Laboratory and the U.S. Army Research Office under contract/grant number W911NF-10-1-0533. Computational resources were provided by Google Cloud Computing grant. We are grateful to Waleed Ammar for his help with transliteration, and to the anonymous reviewers.