Knowledge Based Template Machine Translation In Low-Resource Setting

Anonymous ACL submission

Abstract

Incorporating tagging into neural machine translation (NMT) systems has shown promising results in helping translate rare words such as named entities (NE). However, translating NE in low-resource setting remains a challenge. In this work, we investigate the effect of using tags and NE hypernyms from knowledge graphs (KGs) in parallel corpus in different level of resource conditions. We find the tag-and-copy mechanism (tag the NEs in the source sentence and copy them to the target sentence) improves translation in high-resource settings only. Introducing copying also results in polarizing effects in translating different parts-of-speech (POS). Interestingly, we find that copy accuracy for hypernyms is consistently higher than that of entities. As a way of avoiding "hard" copying and utilizing hypernym in bootstrapping rare entities, we introduced a "soft" tagging mechanism and found consistent improvement in high and low-resource setting.

1 Introduction

NMT methods usually require significant training data. For low-resource languages, NMT models generally do not work as well, especially when translating NEs. With low occurrences and large variations, NEs often remain unseen until inference time. In this paper, we investigate the usefulness of using template tagging methods and hypernyms to generalize NMT under low-resource settings.

Template Machine Translation Template NMT usually involves tagging the input sentences such that the templates simplify the task for the model during translation. One of the first works addressing rare entities in translation uses multiple numbered unknown (*unks*) tokens to link up source and target sentences (Luong et al., 2015). With the introduction of such copy mechanism, models only need to copy (instead of translate) the unknown token from source to target sentence, and (if needed) perform post-processing to replace the copied-over tags. Li et al. (2018a) replaces named entities with their type symbols (i.e. LOC, ORG) on both source and target side, and trains a character-level sequence to sequence model for NE translation.Crego et al. (2016) and Wang et al. (2017) use similar tagging mechanism, with the latter using a dictionary to translate tagged NE. Wang et al. (2019) and Li et al. (2018b) use a few tagging methods from code-switching, boundary tags (i.e. *<ORG>*, *<\ORG>*), to extra embedding to tag NE on both source and target side. Others have explored encouraging copying through constrained decoding (Hokamp and Liu, 2017, Post and Vilar, 2018), or modifying architecture or input format (Gu et al., 2018, Pham et al., 2018 Dinu et al., 2019).

Knowledge Augmented Translation In addition to tagging boundaries of NEs from previous section, a few methods also use POS and other linguistic features to improve NMT (Sennrich and Haddow, 2016, Modrzejewski et al., 2020, Hämäläinen and Alnajjar, 2019). Anwarus Salam et al. (2017) uses hypernyms in a statistical machine translation system for low-resource translation. Meanwhile, many have used KGs to improve NMT systems. Some use KGs for data augmentation(Zhao et al., 2021), while others combine NMT with knowledge graph embedding to improve translation quality (Lu et al., 2018, Zhao et al., 2020, Moussallem et al. (2019).

While our goal resembles similar efforts in template machine translation, we extend the tag types to a much wider range using hypernyms obtained through KGs. In addition, we perform extensive analysis to understand the pros and cons of copy mechanism under different resource conditions. Our paper provides 3 key insights:

- Copy mechanism improves translation only in high-resource setting.
- Copy models translate syntactic POS better

and semantic POS worse, yielding translation with similar sentence structures as the source.
- Appending hypernyms to NEs can improve translation accuracy in low-resource settings.

2 Methods

We first use statistical alignment (FastAlign, Dyer et al., 2013) to build a phrase translation table. We then use DBpedia Spotlight entity linking system (Mendes et al., 2011)[1]) to find NEs within sentences that connects to English DBpedia[2], as well as the translation of the NEs on target side through translation alignment. We substitute the NEs with corresponding templates. After model translation, we remove the tag[3], either keep the translation already in the tag or use the phrase translation table to translate copied entities. The system is modular and all code can be found in our repo[4].

Tagging Methods We use the following templates in our experiments (Table 1): **Tag** and **Trans** are similar to previous works shown to improve translation adequacy (Wang et al., 2019, Li et al., 2018b). In addition to the two methods, we also experiment with adding entity's hypernym provided by DBpedia. Since hypernym is a more generalized term for the entity with higher term frequency, we expect the model to use it as context when translating the sentence in addition to using it to copy. **Add** adds hypernym after entity tag, **TransA** adds hypernym after tag and translation, while **TransR** replaces original entity with hypernym and adds translation. For the target sentences, we replace the NE translations with the same templates as the source sentences.

In addition to enforcing a "hard" copying mechanism using tagging templates, we also include a "soft" signal by simply adding the hypernym after the entity (**HypA**). On the target side, we append the translated hypernym if possible (from phrase translation table) otherwise we use the source language hypernym for **HypA**. Without direct signal for copying, we expect the model to rely on the hypernyms as context when translating NEs.

In our experiments, we ensure the same NEs are tagged across templates, with about 25% of all sentences tagged in each dataset (Appendix Table 6).

[1]https://www.dbpedia-spotlight.org/

[2]https://www.dbpedia.org/

[3]Our soft tagging approach, **HypA**, does not contain explicit tag and requires no removal post translation

[4]Anonymized. Our code is included in a zip file as software component in the submission

2.1 NMT Model

For NMT model, we used XLM introduced by Conneau et al. (2020)[5]. We use the same transformer architecture as Wang et al. (2019): 512 embedding size, 6 encoder and decoder layer, 8 multi-attention heads. Refer to Appendix Section A.5 for more details. We train on both source → target and target → source direction.

3 Experiments

In order to evaluate our results in different resource amount settings, we test our methods in English-Chinese as well as English-Hausa. For English-Chinese, we randomly select 3 million pairs of sentences from MultiUN (Ziemski et al., 2016) as training dataset in high-resource setting. To evaluate English-Chinese translation, we use WMT newstest datasets from 2017-2020. For English-Hausa, we combine available parallel corpus on WMT-21 website[6] including ParaCrawl (Bañón et al., 2020), Wikititles, Khamenei corpus, and English-Hausa Opus corpus (Tiedemann, 2012), in total of 740K parallel sentences. For simulated low-resource condition, we randomly sample 6K sentences from English-Hausa training set. We evaluate English-Hausa translation on newsdev2021 and newstest2021. We treat the WMT newstest as the out-of-domain datasets, and randomly select 5K valid and 5K test sentences as in-domain evaluation sets from each training dataset.

Other than evaluating translation results with multi-BLEU metric, we also investigate the accuracy of the copy mechanism. We report the copy accuracy for hypernym, entity, and entity translation whenever possible. Additionally, we calculate the word translation accuracy by POS occurring before and after the tagged entity to observe the effect of copying on the rest of the sentence. We use *en_core_web_sm* and *zh_core_web_sm* in SpaCy library for POS tagging. For Hausa, since there is not an available POS tagger, we use alignment file from FastAlign and project English POS to corresponding words in Hausa sentence, following Rasooli et al. (2021).

4 Results

4.1 English-Chinese (High-Resource)

Tagging Improves Adequacy and Accuracy We can see a clear improvement of around 1-4

[5]https://github.com/facebookresearch/xlm

[6]https://www.statmt.org/wmt21/translation-task.html

Base.	myanmar was a highly civilized country.
Tag	<start> myanmar <end> was a highly civilized country.
Add	<start> myanmar <mid> state <end> was a highly civilized country.
Trans	<start> myanmar <mid> 缅甸 <end> was a highly civilized country.
TransA	<start>myanmar <mid1>缅甸<mid2>state <end> was a highly civilized country.
TransR	<start> state <mid> 缅甸 <end> was a highly civilized country.
HypA	myanmar state was a highly civilized country.

Table 1: Tagging Templates for English-Chinese source sentence. NE (in red) are replaced with templates (underlined), NE hypernyms are in blue and NE translations are in green. Best viewed in color.

Method	In-Domain	Out-of-Domain
Baseline (all)	33.30 ± 0.63	11.09 ± 0.78
- (tag-only)	34.64 ± 2.1	12.21 ± 0.81
Tag (all)	33.77 ± 0.24	11.26 ± 0.91
- (tag-only)	36.07 ± 0.28	12.89 ± 1.34
Add (all)	33.69 ± 0.21	11.29 ± 0.81
- (tag-only)	35.77 ± 0.36	12.89 ± 1.11
Trans (all)	33.77 ± 0.04	11.25 ± 0.90
- (tag-only)	35.80 ± 0.48	12.97 ± 1.00
TransA (all)	33.35 ± 0.28	11.32 ± 0.83
- (tag-only)	35.37 ± 0.65	13.03 ± 0.98
TransR (all)	33.84 ± 0.29	11.18 ± 0.87
- (tag-only)	35.73 ± 0.61	12.75 ± 0.88
HypA (all)	34.39 ± 0.14	11.48 ± 0.87
- (tag-only)	**37.54** ± 0.07	**13.69** ± 0.95

Table 2: Average and standard deviation of BLEU scores across evaluation sets for all tagging methods in English-Chinese. Evaluation is performed on whole dataset (**all**) and on tagged sentences only (**tag-only**). Best performances in tag-only subsets are in bold. Best performances in all datasets are underscored. See results for individual datasets in Appendix Table 7

BLEU point on average (Table 2). The improvements are much larger when we evaluate it on tag-only subsets. **HypA** outperforms other methods consistently. Similar trend is observed in Chinese-English Translation (see Appendix Table 8).

When looking at translation accuracy (Table 3) of the tagged NEs, we see about 35 points improvement in translation accuracy. This is expected because copying is much easier than translating. **HypA** method, while performing better in BLEU, does not improve NE translation accuracy as much because it does not enforce "hard" copying. **Tag** method performs best in translating NEs with 91.92% accuracy (assuming perfect phrase translation table).

Method	Entity	Translation	Hypernym
Baseline	-	55.38	-
Tag	91.92	-	-
Add	91.02	-	92.04
Trans	**92.12**	90.99	-
TransA	91.83	**91.27**	**92.97**
TransR	-	89.12	91.66
HypA	-	55.76	58.69

Table 3: Copy accuracy (percentage) mean for different parts of the tag in English-Chinese across evaluation sets. We equate correct NE translation in baseline to correct translation copy. The hypernym translation accuracy for **HypA** is approximated with the word translation accuracy after the entity.

Effects of Copy Mechanism on Translation As seen in Figure 1, copying provides benefits and downfalls. It improves translation accuracy for POSs which serve as structural syntactic signals in sentences such as conjunctions, particles, punctuation while decreasing accuracy for POSs containing more semantic information that require more context to translate (verb, adjective, adverb). Qualitatively, this is equivalent to producing translations with similar sentence structures to source sentence (Appendix Table 10). Since copying is a direct signal for models to ignore context and translate word by word for the entity, it is not surprising to see such polarizing effects on the rest of the sentences. Unexpectedly, despite being a "soft" copy signal, **HypA** shows similar effects. We suspect that the repeating semantic of appending hypernyms after NEs yields similar signal for models to follow word-by-word order sensitive translation.

Similarly in Table 2, we do not see significant BLEU improvement of tagging methods that contain hypernym (**Add**, **TransA**, **TransR**) over those that do not (**Tag**, **Trans**). We believe, by the same mechanism described above, the copy mechanism shifts models' priority from using the semantics of the hypernym to simply copying the word.

4.2 English-Hausa (Medium-Resource)

Full English-Hausa yields similar results as English-Chinese, except that the improvements in BLEU from tagged models over baseline become marginal (Appendix Table 11). **HypA** and **Tag** performs best in-domain while baseline performs best out-of-domain. Additionally, copy accuracy decreases from 90% to 80%, but remains 20% higher over baseline accuracy (Appendix Table 12). **Tag** still outperforms other methods in copy accuracy.

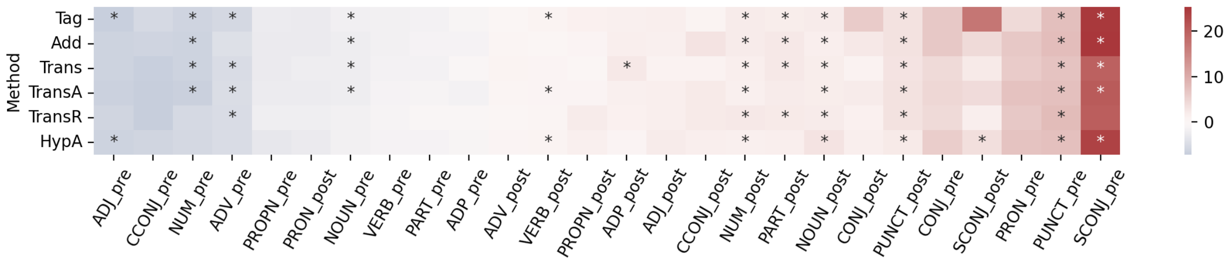

Figure 1: POS translation accuracy (percentage) difference against baseline before (**_pre**) and after (**_post**) the tagged entity in English-Chinese. * indicates a statistical significant difference against baseline with p-value < 0.05

4.3 6K English-Hausa (Low-Resource)

Method	In-Domain	Out-of-Domain
Baseline	7.61 $\pm$ 0.21	3.80 $\pm$ 3.37
- (tag-only)	7.21 $\pm$ 0.85	3.40 $\pm$ 2.87
Tag (all)	7.39 $\pm$ 0.14	3.67 $\pm$ 3.12
- (tag-only)	6.69 $\pm$ 0.79	3.39 $\pm$ 3.13
Trans (all)	7.45 $\pm$ 0.08	3.91 $\pm$ 3.44
- (tag-only)	6.99 $\pm$ 0.92	**3.60** $\pm$ 3.44
HypA (all)	7.53 $\pm$ 0.25	3.52 $\pm$ 2.88
- (tag-only)	**7.82** $\pm$ 1.40	2.55 $\pm$ 1.89

Table 4: BLEU scores for 6K English-Hausa data. Only top performing methods are included.

Method	entity	translation	hypernym
Baseline	-	42.44	-
Tag	30.72↓	-	-
Add	34.48↓	-	55.66
Trans	37.81	35.69↓	-
TransA	**39.01**	37.53	**55.91**
TransR	-	30.61↓	55.39
HypA	-	**44.77**↑	48.32

Table 5: Copy accuracy (mean) in 6K English-Hausa dataset models across evaluation sets. Arrows indicate statistical difference from baseline with p-value < 0.05.

In low-resource setting, tagging does not improve performance (Table 4). The NE copy accuracy drops below baseline. Interestingly, hypernyms are more consistently copied to the target side (Table 5). We believe this is due to hypernyms having higher term frequency in the training data. Compared to baseline, only **HypA** method is able to improve NE translation accuracy and obtain higher BLEU for tag-only subsets in-domain (Table 4). Despite not having as high of hypernym copy accuracy, the model is able to use hypernym as context, and improve NE translation.

5 Discussion

Copy mechanism in low-resource? As results show, copy mechanism is able to increase NE translation accuracy in both high and medium-resource but not in low-resource condition. Learning to copy requires significant amount of data. Once tags are recognized, the semantics of the content within are ignored. Translations become structurally similar to source sentence, while focusing less on semantics of words that depend on the context. Without enough data, "softer" methods of augmentation (**HypA** or extra embedding used by (Moussallem et al., 2019)) that incorporates hypernym in translation is a better choice. Work by (Currey et al., 2017), which copies target sentences to source side to create additional bitext, might be interesting alternatives to encourage copying.

Low-Resource translation affected by term frequency. As suggested by Table 5, before copy mechanism generalizes, models are more likely to copy words that occur more frequently (hypernyms). This points to potential directions in low-resource NLP in using hypernyms to bootstrap performance of other words or sentences. Data augmentation techniques like randomly inserting/replacing NEs with hypernyms could be potential ways of adding data points in low-resource settings and better generalize embedding space.

6 Conclusion

In our paper, we analyzed the tag-and-copy mechanism under different resource conditions. We found that learning to copy requires significant amount of resource which is often not achievable in low-resource languages. Additionally, we found that copying can induce polarizing effects on translating different POSs. It discouraged models from using contextual information, but provide "structural supervision". In low-resource setting, we found correlation between term frequency and copying accuracy. Our proposed method of appending hypernym after NEs was able to encourage better translation in both low and high-resource setting.

Acknowledgements

References

Khan Md Anwarus Salam, Setsuo Yamada, and Nishino Tetsuro. 2017. Improve example-based machine translation quality for low-resource language using ontology. *International Journal of Networked and Distributed Computing*, 5:176.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *ACL*.

Josep Crego, Jungi Kim, Guillaume Klein, Anabel Rebollo, Kathy Yang, Jean Senellart, Egor Akhanov, Patrice Brunelle, Aurelien Coquard, Yongchao Deng, et al. 2016. Systran's pure neural machine translation systems. *arXiv preprint arXiv:1610.05540*.

Anna Currey, Antonio Valerio Miceli-Barone, and Kenneth Heafield. 2017. Copied monolingual data improves low-resource neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 148–156.

Georgiana Dinu, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. 2019. Training neural machine translation to apply terminology constraints. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3063–3068.

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648.

Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor OK Li. 2018. Search engine guided neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Mika Hämäläinen and Khalid Alnajjar. 2019. A template based approach for training nmt for low-resource uralic languages-a pilot with finnish. In *Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence*, pages 520–525.

Chris Hokamp and Qun Liu. 2017. Lexically constrained decoding for sequence generation using grid beam search. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1535–1546.

Belinda Z. Li, Sewon Min, Srinivasan Iyer, Yashar Mehdad, and Wen-tau Yih. 2020. Efficient one-pass end-to-end entity linking for questions. In *EMNLP*.

Xiaoqing Li, Jinghui Yan, Jiajun Zhang, and Chengqing Zong. 2018a. Neural name translation improves neural machine translation. In *China Workshop on Machine Translation*, pages 93–100. Springer.

Zhongwei Li, Xuancong Wang, AiTi Aw, Eng Siong Chng, and Haizhou Li. 2018b. Named-entity tagging and domain adaptation for better customized translation. In *Proceedings of the Seventh Named Entities Workshop*, pages 41–46.

Yu Lu, Jiajun Zhang, and Chengqing Zong. 2018. Exploiting knowledge graph in neural machine translation. In *China Workshop on Machine Translation*, pages 27–38. Springer.

Minh-Thang Luong, Ilya Sutskever, Quoc Le, Oriol Vinyals, and Wojciech Zaremba. 2015. Addressing the rare word problem in neural machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 11–19.

Pablo N. Mendes, Max Jakob, Andres Garcia-Silva, and Christian Bizer. 2011. Dbpedia spotlight: Shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems (I-Semantics)*.

Maciej Modrzejewski, Miriam Exel, Bianka Buschbeck, Thanh-Le Ha, and Alex Waibel. 2020. Incorporating external annotation to improve named entity translation in nmt. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 45–51.

Diego Moussallem, Mihael Arčan, Axel-Cyrille Ngonga Ngomo, and Paul Buitelaar. 2019. Augmenting neural machine translation with knowledge graphs. *arXiv preprint arXiv:1902.08816*.

Ngoc-Quan Pham, Jan Niehues, and Alex Waibel. 2018. Towards one-shot learning for rare-word translation with external experts. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 100–109.

Matt Post and David Vilar. 2018. Fast lexically constrained decoding with dynamic beam allocation for 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 1314–1324.

Mohammad Sadegh Rasooli, Chris Callison-Burch, and Derry Tanti Wijaya. 2021. "wikily" supervised neural translation tailored to cross-lingual tasks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1655–1670.

Rico Sennrich and Barry Haddow. 2016. Linguistic input features improve neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers*, pages 83–91.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Tao Wang, Shaohui Kuang, Deyi Xiong, and António Branco. 2019. Merging external bilingual pairs into neural machine translation. *arXiv preprint arXiv:1912.00567*.

Yuguang Wang, Shanbo Cheng, Liyang Jiang, Jiajun Yang, Wei Chen, Muze Li, Lin Shi, Yanfeng Wang, and Hongtao Yang. 2017. Sogou neural machine translation systems for wmt17. In *Proceedings of the Second Conference on Machine Translation*, pages 410–415.

Yang Zhao, Lu Xiang, Junnan Zhu, Jiajun Zhang, Yu Zhou, and Chengqing Zong. 2020. Knowledge graph enhanced neural machine translation via multi-task learning on sub-entity granularity. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4495–4505.

Yang Zhao, Jiajun Zhang, Yu Zhou, and Chengqing Zong. 2021. Knowledge graphs enhanced neural machine translation. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 4039–4045.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534, Portorož, Slovenia. European Language Resources Association (ELRA).

A Appendix

A.1 Text Preprocessing

We follow default preprocessing steps in XLM repo. For English and Hausa, we use Moses *tokenizer.perl* script, after which we lower-case letters and remove accents. For Chinese, we use Moses *tokenizer_PTB.perl* script.

A.2 Special Tags in XLM Model

During tagging, in order to prevent creating additional vocabulary, we use four of the special tokens (i.e. <special2>, <special3>, <special4>, <special5>), that already exist in pretrained XLM-R model vocab, instead of actual <start>, <end>, etc.

A.3 Tagging Statistics

Language Pair	Train Size	Tag Size
English-Hausa	6 K	1.5 K (25.6%)
English-Hausa	746 K	191 K (25.6%)
English-Chinese	2,990 K	816 K (27.3%)

Table 6: Tagging Statistics in Training Sets

A.4 Entity Linking

During experimentation, we have also tried more recent Entity linking systems such as BLINK (Li et al., 2020) [7]. In reality, we find BLINK tagging less entities as well as taking a longer time. We presume this is because BLINK expects normally-cased sentences while our entity linking occurs after input sentences are lower-cased.

A.5 Model Training Details

In all of our experiments, we use the pretrained XLM-R BPE vocab with 200,000 tokens, trained on 100 lanugages [8]. We use Adam optimizer, learning rate 0.0001, epoch size 300000, dropout rate of 0.1. We fix number of tokens in a batch to be around 2000. To increase batch size with GPU memory constraint, we use gradient accumulation for every four batches to increase effective batch size. For low-resource condition with 6K training sentences (see Section 3), we change epoch size to 120,000, dropout of 0.2, and enforce minimum sentence length to 10 words. All models are trained on NVIDIA V100 GPUs. Each English-Chinese model takes about 5 days to train (1 GPU time). Each English-Hausa model takes about 3 days and each English-Hausa 6K model takes about 15 hours.

Method	subset	valid	test	nd2017	nt2017	nt2018	nt2019	nt2020	ntB2020
Baseline	all	32.85	33.75	11.23	10.77	11.02	10.20	12.54	10.78
Baseline	tag-only	33.15	36.12	13.22	12.69	12.13	11.30	12.69	11.20
Tag	all	33.59	33.94	11.20	11.38	<u>11.34</u>	10.14	12.85	10.66
Tag	tag-only	35.86	36.27	13.72	14.20	13.18	11.16	13.85	11.25
Add	all	33.53	33.84	11.15	<u>11.58</u>	11.19	10.36	12.71	10.72
Add	tag-only	35.51	36.03	13.25	14.48	12.88	12.17	13.33	11.20
Trans	all	33.74	33.80	11.23	11.10	10.72	<u>10.73</u>	13.04	10.68
Trans	tag-only	35.45	36.14	13.46	13.97	12.40	12.34	**14.04**	11.59
TransA	all	33.14	33.55	11.10	11.33	11.28	10.47	12.89	10.85
TransA	tag-only	34.90	35.83	13.50	13.72	**13.54**	12.02	13.84	11.53
TransR	all	33.63	34.05	11.10	11.08	11.18	10.31	12.82	10.61
TransR	tag-only	35.29	36.16	13.32	13.65	12.63	11.85	13.46	11.56
HypA	all	<u>34.29</u>	<u>34.39</u>	<u>11.31</u>	11.51	11.17	<u>10.73</u>	<u>13.18</u>	<u>10.99</u>
HypA	tag-only	**37.49**	**37.59**	**14.67**	**14.73**	13.49	**13.28**	13.76	**12.18**

Table 7: BLEU scores across evaluation sets for all tagging methods in English-Chinese. Evaluation is performed on whole dataset and on tagged sentences only. Best performances in tagged subset are in bold. Best performances in all datasets are underscored. Each point represents a single data point. (nd2017=newsdev2017, nt2017=newstest2017, etc)

A.6 English-Chinese Full Results

A.7 Chinese-English Translation Results

A.8 Copy Efficiency In / Out of Domain

In English-Chinese translation results, we can observe that the copy accuracy for the tags is similar across different set regardless of the domain (Table 9), which is a good sign considering the drop in BLEU across the out-of-domain datasets. This indicate copy mechanism is a valuable method in translation avenues where entity translation accuracy is more valuable than adequacy (i.e. medical, scientific domain), confirming with results in Pham et al. (2018) and Dinu et al. (2019).

A.9 English-Hausa POS Accuracy Qualitative Analysis

A.10 English-Hausa 6K Translation Results

[7] https://github.com/facebookresearch/BLINK

[8] See https://github.com/facebookresearch/XLMthe-17-and-100-languages for language details

Method	subset	valid	test	nd2017	nt2017	nt2018	nt2019	nt2020	ntB2020
Baseline	all	38.46	42.33	12.06	12.74	13	10.37	12.13	11.65
Baseline	tag-only	43.28	44.87	13.01	13.81	14.16	11	12.88	12.47
Tag	all	41.47	42.56	12.53	12.76	13.06	10.55	12.48	<u>11.84</u>
Tag	tag-only	44.01	45.13	14.51	13.87	14.57	11.94	13.43	13.17
Add	all	41.42	42.37	12.76	13.14	12.74	10.38	12.46	11.83
Add	tag-only	43.82	44.86	14.73	14.11	14.33	11.54	**13.67**	**13.26**
Trans	all	41.31	42.42	12.35	13	13.17	10.42	12.21	11.61
Trans	tag-only	43.4	44.8	13.84	14.26	14.96	12.14	13.26	13.02
TransA	all	41.1	42.17	12.76	<u>13.21</u>	13.13	10.66	12.07	11.52
TransA	tag-only	42.99	44.39	14.3	**14.69**	14.84	12.24	13.12	12.72
TransR	all	41.21	42.28	<u>12.8</u>	13.03	12.88	<u>10.75</u>	<u>12.52</u>	11.81
TransR	tag-only	43.49	44.75	**15.03**	14.26	14.69	12.26	13.39	12.82
HypA	all	<u>41.84</u>	<u>42.99</u>	12.47	12.98	<u>13.29</u>	10.48	12.2	11.68
HypA	tag-only	**45.32**	**46.08**	14.76	14.55	**15.07**	**12.62**	13.18	13.23

Table 8: BLEU scores across evaluation sets for all tagging methods in Chinese-English. There is a consistent 0.5-2 point improvement with tagged methods over baseline. Each point represents a single data point.

	Valid	Test	nd2017	nt2017
H	91.98	90.92	94.88	97.19
E	91.84	90.5	92.79	91.8
T	91.91	90.15	93.17	93.91
	nt2018	nt2019	nt2020	ntB2020
H	94.45	94.16	88.48	91.73
E	93.76	93.67	88.02	92.27
T	92.37	92.94	86.41	89.33

Table 9: Copy Accuracy of TransA model across different in and out-of-domain evaluation datasets. Each point represents a single data point. H=Hypernym, E=Entity, T=Entity translation

Label	in the gambia 's interim paper , it was noted that major factors in poverty among rural women include their predominance in subsistence agriculture , where they have less access than men to mechanized technologies , and the fact that , in addition to farming , they work longer hours than men carrying out household tasks .
Baseline	the interim document of the gambia indicated that rural women 's poverty was mainly due to their livelihood agriculture , which was less skilled than men ; and that they were more time spent than men to run their household than men , in addition to their work .
Tag	the <special2> gambia <special5> interim paper indicated that the main cause of poverty among rural women was their main livelihood agriculture , less access to mechanized technologies than men ; and that in addition to farming , they were more time-consuming than men .
Add	the <special2> gambia <special3> country <special5> 's interim paper noted that the main causes of poverty among rural women were their primary work in subsistence agriculture , more than men 's access to mechanical techniques , and that they would have more time than men to take their household roles in addition to their farm .
Trans	the <special2> gambia <special3> 冈比亚 <special5> 's provisional document noted that the main causes of poverty among rural women are their primary subsistence agriculture , less than men 's access to mechanized technologies , and that in addition to their farm , they are more time than men to operate household .
TransA	in the <special2> gambia <special3> 冈比亚 <special4> country <special5> 's interim paper , it was noted that major factors in poverty among rural women include their predominance in subsistence agriculture , where they have less access than men to mechanized technologies , and the fact that , in addition to farming , they work longer hours than men carrying out household tasks .
TransR	the provisional document of the <special2> country <special3> 冈比亚 <special5> indicates that the main causes of poverty among rural women are their predominance in livelihood agriculture , less access to mechanized technologies than men , and that they are more time than men to take up their housework in addition to their agricultural work .
HypA	the interim document of the gambia country indicated that the main reason for poverty among rural women was their predominant livelihood farming , less than the mechanized technique of access to men ; and that they were also taking more time than men to operate their household tasks .

Table 10: Translation example before post-translation tag removal. In Chinese-English translation setting, we compare all model translation results with ground truth English sentence. In all tagging methods, models tend to produce more similar sentence structures due to similar syntactic word choices. Given fixed sentence structures, there is less emphasis on translating the rest of the words that contain more semantic variations (verbs, adjectives, adverbs, etc.). NE (in red) are replaced with templates (underlined), NE hypernyms are in blue and NE translations are in green. Best viewed in color.

Method	valid	test	nd2021	nt2021
Base(all)	32.94	32.89	11.31	21.62
- (tag-only)	35.35	37.12	11.50	**23.18**
Tag(all)	33.17	32.99	10.77	21.84
- (tag)	**35.91**	37.28	11.86	23.13
Add (all)	32.25	32.62	11.16	21.42
- (tag-only)	34.58	36.44	12.07	22.54
Trans(all)	32.27	32.29	10.85	21.56
- (tag-only)	35.45	36.14	**12.01**	22.71
TransA	32.22	32.3	10.58	21.38
- (tag-only)	33.88	35.94	11.33	22.56
TransR	32.65	32.77	11.18	21.74
- (tag-only)	34.74	36.73	12.38	22.71
HypA(all)	33.02	33.00	9.59	20.24
- (tag-only)	35.89	**37.39**	8.12	15.42

Table 11: BLEU scores with English-Hausa full data. Each point represents a single data point.

Method	entity	translation	hypernym
Tag	81.93	-	-
Add	79.16	-	79.34
Trans	**82.10**	**81.30**	-
TransR	-	80.99	**81.86**
TransA	80.87	80.23	80.90
HypA	-	61.00	64.29
Baseline	-	59.56	-

Table 12: Copy accuracy mean with English-Hausa full data. Aggregated across all evaluation datasets.

Method	valid	test	nd2021	nt2021
Base (all)	7.75	7.46	1.41	6.18
- (tag-only)	6.61	7.81	**1.37**	5.43
Tag (all)	7.49	7.29	1.46	5.87
- (tag-only)	6.13	7.25	1.18	5.6
Add (all)	7.59	7.52	1.38	6.29
- (tag-only)	6.19	7.61	1.25	5.48
Trans (all)	7.51	7.39	1.48	6.34
- (tag-only)	6.34	7.64	1.16	6.03
TransA (all)	7.14	7.12	1.35	6.13
- (tag-only)	5.86	7.3	1.19	5.56
TransR (all)	7.35	7.32	1.4	6.5
- (tag-only)	5.73	7.22	1.03	5.74
HypA (all)	7.71	7.35	1.48	5.55
- (tag-only)	**6.83**	**8.18**	1.21	3.88

Table 13: BLEU scores in 6K English-Hausa data for all models across individual evaluation sets. Each point represents a single data point. nd2021=newsdev2021, nt2021=newstest2021