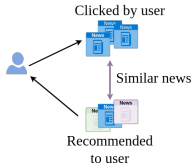
MIND Your Language: A Multilingual Dataset for Cross-lingual News Recommendation

Andreea Iana[1], Goran Glavaš[2], Heiko Paulheim[1]

[1]Data and Web Science Group, University of Mannheim, Germany

[2]Center for Artificial Intelligence and Data Science, University of Würzburg, Germany

Multilinguality in News & Recommendation

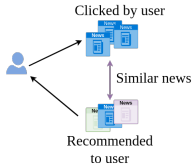


Personalized news recommendation shapes readers' views

IN

An increasingly **language-diverse** & **polyglot** online user community

Multilinguality in News & Recommendation

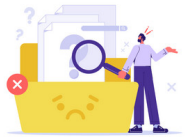


Personalized news recommendation shapes readers' views

IN

An increasingly **language-diverse** & **polyglot** online user community

BUT

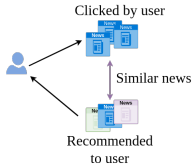


Scarcity of **publicly available**, **diverse**, **multilingual** news recommendation datasets

AND

Focus on **monolingual news consumption & high-resource languages**

Multilinguality in News & Recommendation

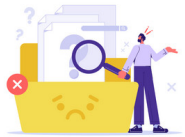


Personalized news recommendation shapes readers' views

IN

An increasingly **language-diverse** & **polyglot** online user community

BUT



Scarcity of **publicly available**, **diverse**, **multilingual** news recommendation datasets

AND

Focus on **monolingual news consumption & high-resource languages**

Less relevant & less balanced & less diverse

recommendations for polyglots and readers of low-resource languages

xMIND: A Multilingual News Dataset

Considerations

- Diversity (linguistic, geographic, digital footprint size)
- Multi-parallel data
- Open source

xMIND: A Multilingual News Dataset

Considerations

- Diversity (linguistic, geographic, digital footprint size)
- Multi-parallel data
- Open source

MIND: most used news recommendation benchmark

1 million users

130,379 unique articles

> 24 million clicks

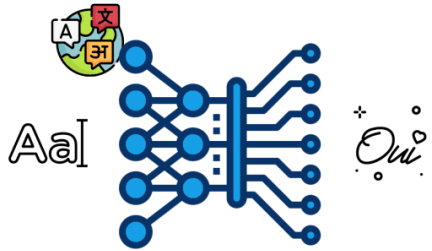
xMIND: A Multilingual News Dataset

Considerations

- Diversity (linguistic, geographic, digital footprint size)
- Multi-parallel data
- Open source

MIND: most used news recommendation benchmark

- 1 million users
- 130,379 unique articles
- > 24 million clicks



Neural Machine Translation (MT, i.e., NLLB 3.3B [1])

xMIND

Main hyperparameters chosen on subset of Global Voices[2] dataset (best averaged over all language pairs overlapping with xMIND)

xMIND: A Multilingual News Dataset

Code	Language	Script	Macro-area	Family	Genus	Total Speakers (M)	Res.
SWH	Swahili	Latin	Africa	Niger-Congo	Bantu	71.6	high
SOM	Somali	Latin	Africa	Afro-Asiatic	Lowland East Cushitic	22.0	low
CMN	Mandarin Chinese	Han	Eurasia	Sino-Tibetan	Sinitic	1,138.2	high
JPN	Japanese	Japanese	Eurasia	Japonic	Japanesic	1,234.5	high
TUR	Turkish	Latin	Eurasia	Altaic	Turkic	90.0	high
TAM	Tamil	Tamil	Eurasia	Dravidian	Dravidian	86.6	low
VIE	Vietnamese	Latin	Eurasia	Austro-Asiatic	Vietic	85.8	high
THA	Thai	Thai	Eurasia	Tai-Kadai	Kam-Tai	60.8	high
RON	Romanian	Latin	Eurasia	Indo-European	Romance	24.5	high
FIN	Finnish	Latin	Eurasia	Uralic	Finnic	5.6	high
KAT	Georgian	Georgian	Eurasia	Kartvelian	Georgian-Zan	3.9	low
HAT	Haitian Creole	Latin	North-America	Indo-European	Creoles and Pidgins	13.0	low
IND	Indonesian	Latin	Papunesia	Austronesian	Malayo-Sumbawan	199.1	high
GRN	Guarani	Latin	South America	Tupian	Maweti-Guarani	(L1 only) 6.7	low

✓ 14 languages from 13 families

✓ 5 / 6 macro-areas

✗ US-centric content

✓ 5 low-resource languages

✓ 6 scripts from 3 families

✗ Cultural nuances ignored

xMIND: A Multilingual News Dataset

Diversity Indices [3]

Typology
- quantifies the presence/absence of a linguistic property in a language based on predefined typological binary features

Family
- number of distinct language families in the sample size

Geography
- entropy of the distribution of languages in the sample over 6 geographic macro-areas of the world

xMIND: A Multilingual News Dataset

Diversity Indices [3]

Typology
- quantifies the presence/absence of a linguistic property in a language based on predefined typological binary features

Family
- number of distinct language families in the sample size

Geography
- entropy of the distribution of languages in the sample over 6 geographic macro-areas of the world

Other Multilingual Datasets

NeMig [4]
- Languages: ENG, DEU
- Focus on one topic (refugee migration)
- Open-source dataset

Wu et al. [5]
- Languages: ENG, DEU, FRA, ITA, JPN, SPA, KOR
- Proprietary dataset

xMIND: A Multilingual News Dataset

Diversity Indices [3]

Typology
- quantifies the presence/absence of a linguistic property in a language based on predefined typological binary features

Family
- number of distinct language families in the sample size

Geography
- entropy of the distribution of languages in the sample over 6 geographic macro-areas of the world

Other Multilingual Datasets

NeMig [4]
- Languages: ENG, DEU
- Focus on one topic (refugee migration)
- Open-source dataset

Wu et al. [5]
- Languages: ENG, DEU, FRA, ITA, JPN, SPA, KOR
- Proprietary dataset

	Range	xMIND	NeMig	Wu et al.
Typology	[0, 1]	0.42	0.05	0.31
Family	[0, 1]	0.93	0.50	0.43
Geography	[0, ln 6]	1.13	0.00	0.00

xMIND: Translation Quality

xMIND: Translation Quality

xMIND: Translation Quality

xMIND

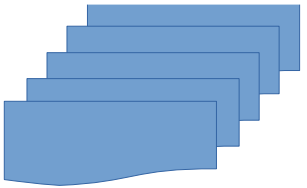
1. Manual quality estimation of translations through annotations with native speakers

&

2. Robustness of neural news recommenders trained & evaluated on different translations

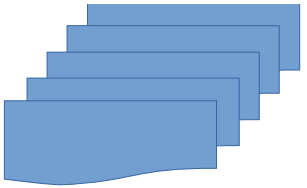
Evaluation Setup w/ xMIND (small)

Open-source MT



versus

Commercial MT



Manual Quality Estimation of Translation

Data: 50 news from xMIND used for testing, sampled according to (i) categorical & (ii) length distribution

Annotators: 2 per target language, native speakers of target language & fluent in English

Setup: randomized source of translations during annotation to avoid position bias

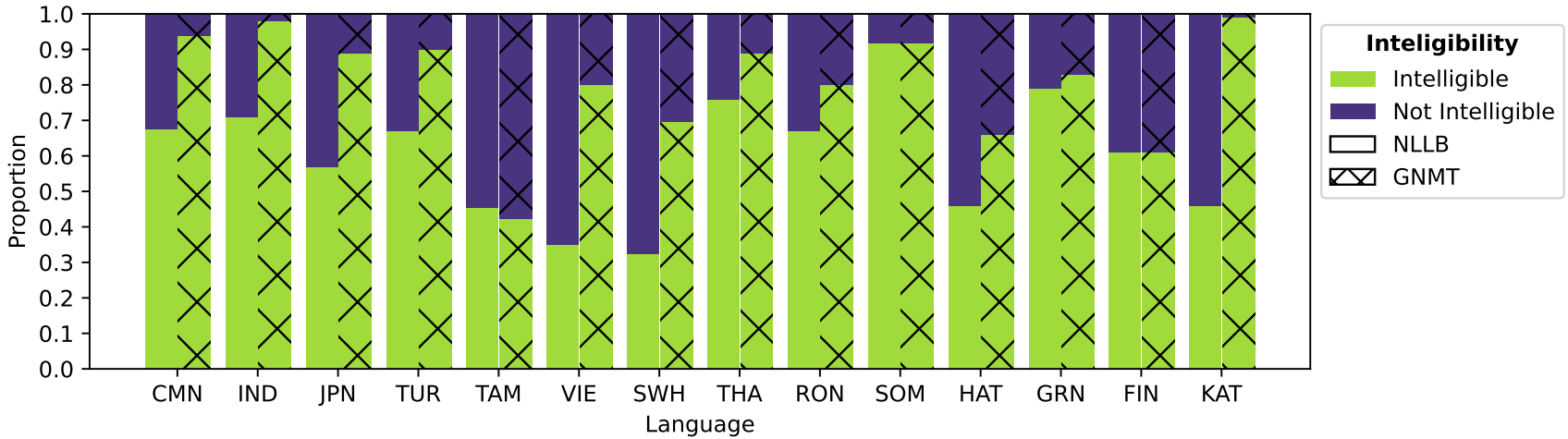
Manual Quality Estimation of Translation

Data: 50 news from xMIND used for testing, sampled according to (i) categorical & (ii) length distribution

Annotators: 2 per target language, native speakers of target language & fluent in English

Setup: randomized source of translations during annotation to avoid position bias

Intelligibility (how acceptable is the translation)



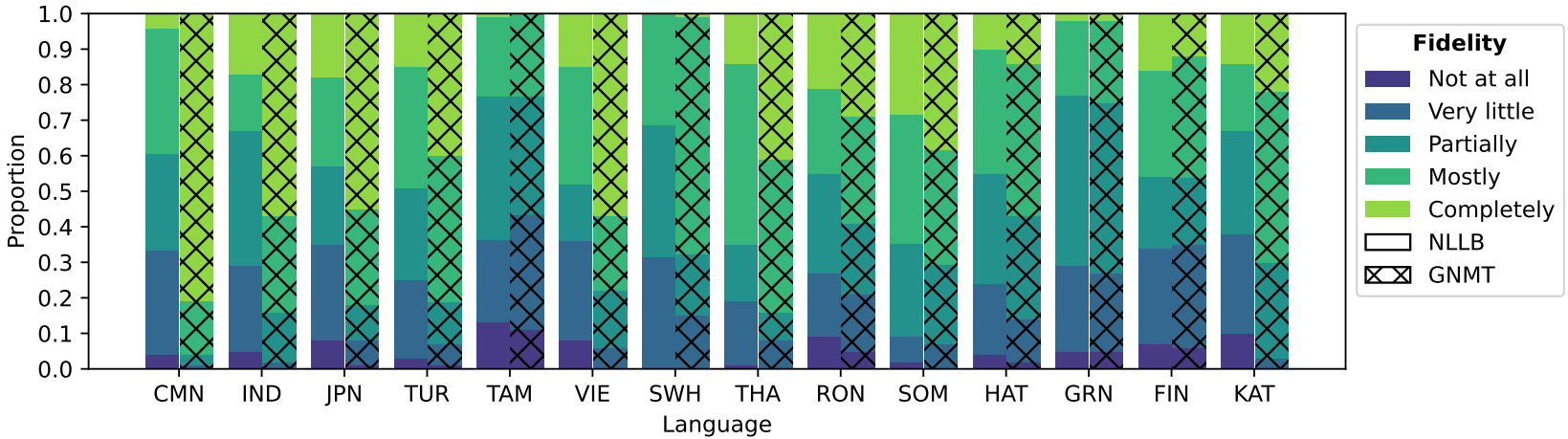
Manual Quality Estimation of Translation

Data: 50 news from xMIND used for testing, sampled according to (i) categorical & (ii) length distribution

Annotators: 2 per target language, native speakers of target language & fluent in English

Setup: randomized source of translations during annotation to avoid position bias

Fidelity (the extent to which the original information is retained in the translation)



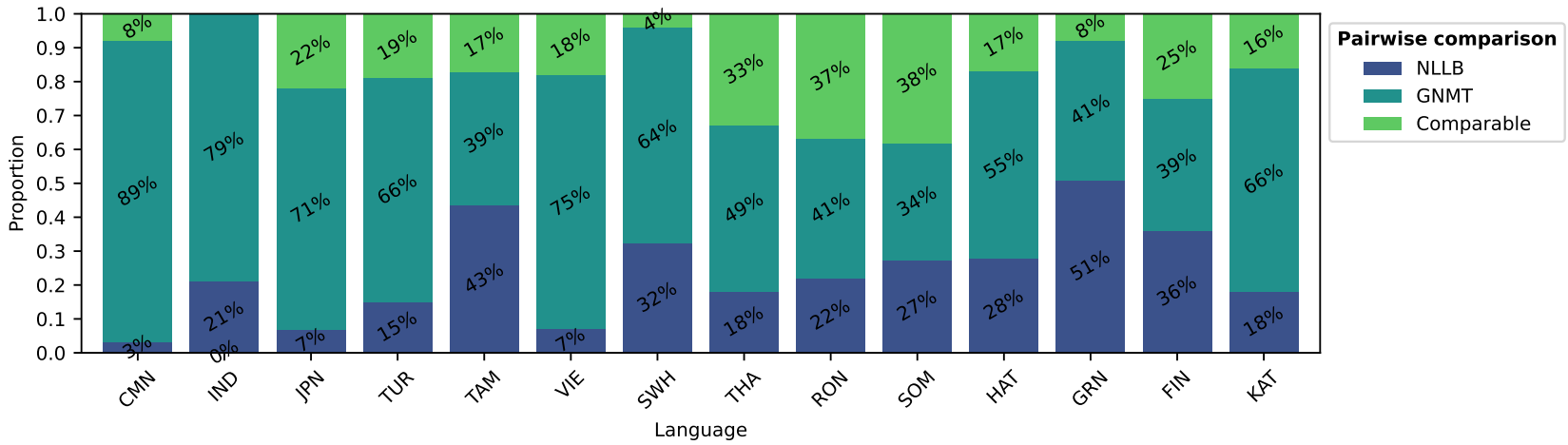
Manual Quality Estimation of Translation

Data: 50 news from xMIND used for testing, sampled according to (i) categorical & (ii) length distribution

Annotators: 2 per target language, native speakers of target language & fluent in English

Setup: randomized source of translations during annotation to avoid position bias

Pairwise comparison (which translation is better)



Cross-Lingual News Recommendation

Evaluation Setup

Zero-shot Cross-lingual Transfer (ZS-XLT)

Few-shot Cross-lingual Transfer (FS-XLT)

News Consumption Patterns

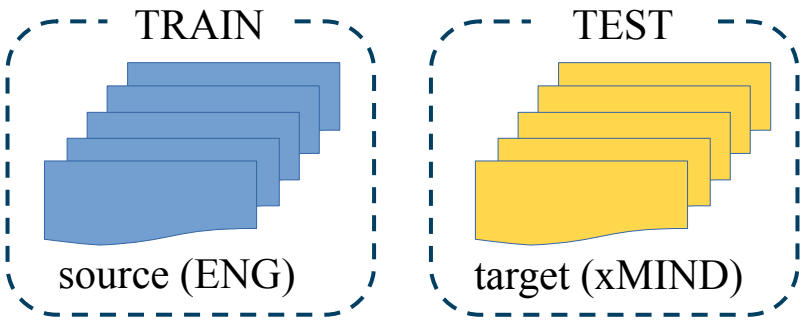
Monolingual consumption

Simulated **bilingual consumption**

Cross-Lingual News Recommendation

Evaluation Setup

Zero-shot Cross-lingual Transfer (ZS-XLT)



Few-shot Cross-lingual Transfer (FS-XLT)

News Consumption Patterns

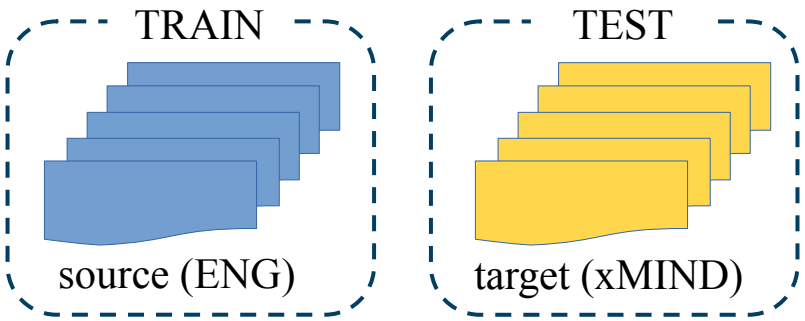
Monolingual consumption

Simulated **bilingual consumption**

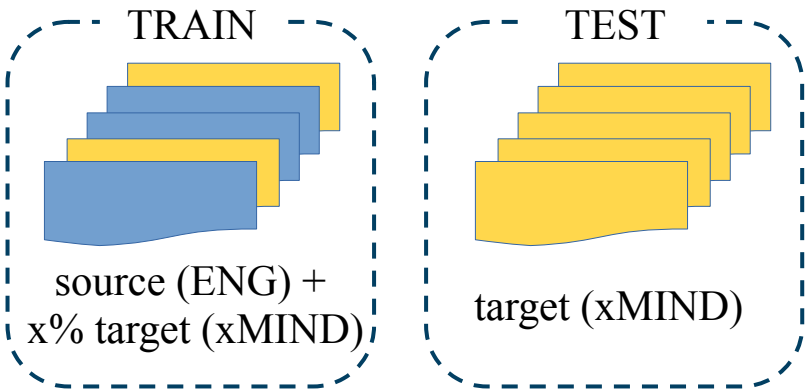
Cross-Lingual News Recommendation

Evaluation Setup

Zero-shot Cross-lingual Transfer (ZS-XLT)



Few-shot Cross-lingual Transfer (FS-XLT)



News Consumption Patterns

Monolingual consumption

Simulated bilingual consumption

Cross-Lingual News Recommendation

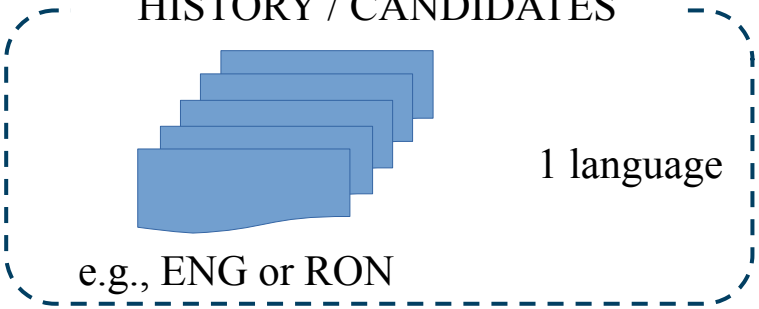
Evaluation Setup

Zero-shot Cross-lingual Transfer (ZS-XLT)

Few-shot Cross-lingual Transfer (FS-XLT)

News Consumption Patterns

Monolingual consumption



***Simulated* bilingual consumption**

Cross-Lingual News Recommendation

Evaluation Setup

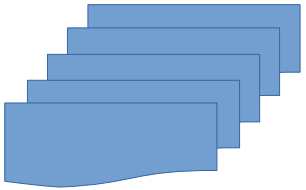
Zero-shot Cross-lingual Transfer (ZS-XLT)

Few-shot Cross-lingual Transfer (FS-XLT)

News Consumption Patterns

Monolingual consumption

HISTORY / CANDIDATES

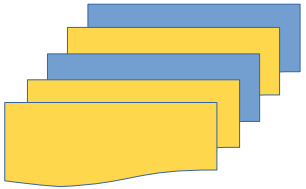


1 language

e.g., ENG or RON

***Simulated* bilingual consumption**

HISTORY / CANDIDATES



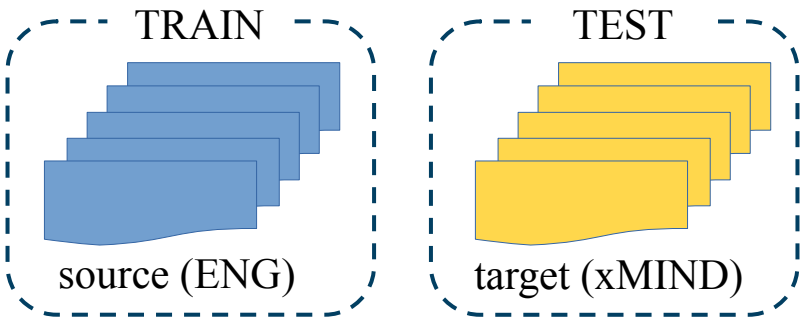
2 languages

i.e., randomly replace a portion of ENG news w/ news in target language

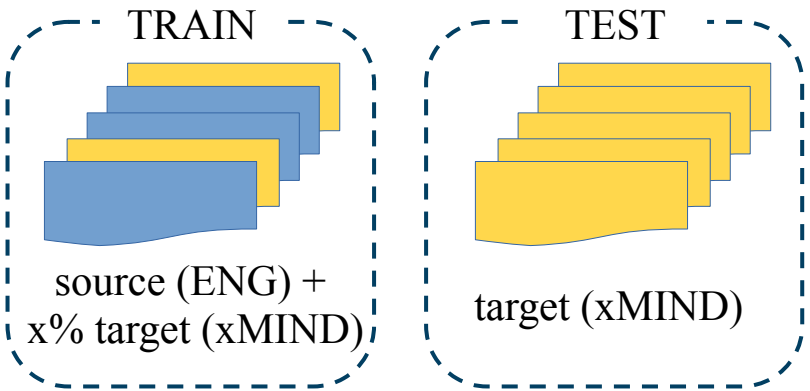
Cross-Lingual News Recommendation

Evaluation Setup

Zero-shot Cross-lingual Transfer (ZS-XLT)

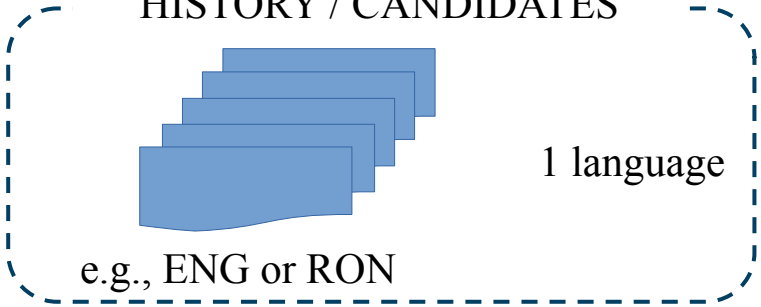


Few-shot Cross-lingual Transfer (FS-XLT)

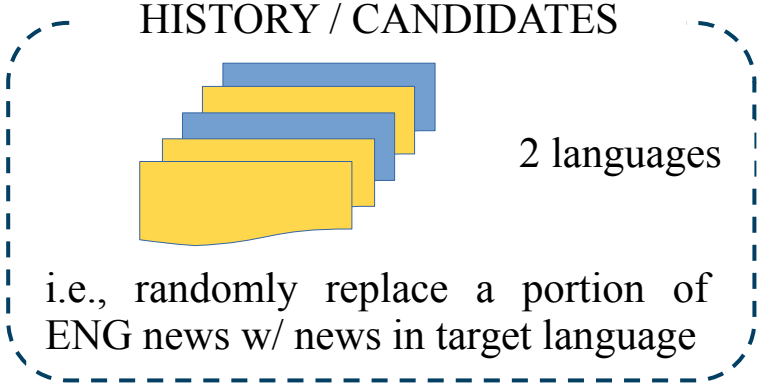


News Consumption Patterns

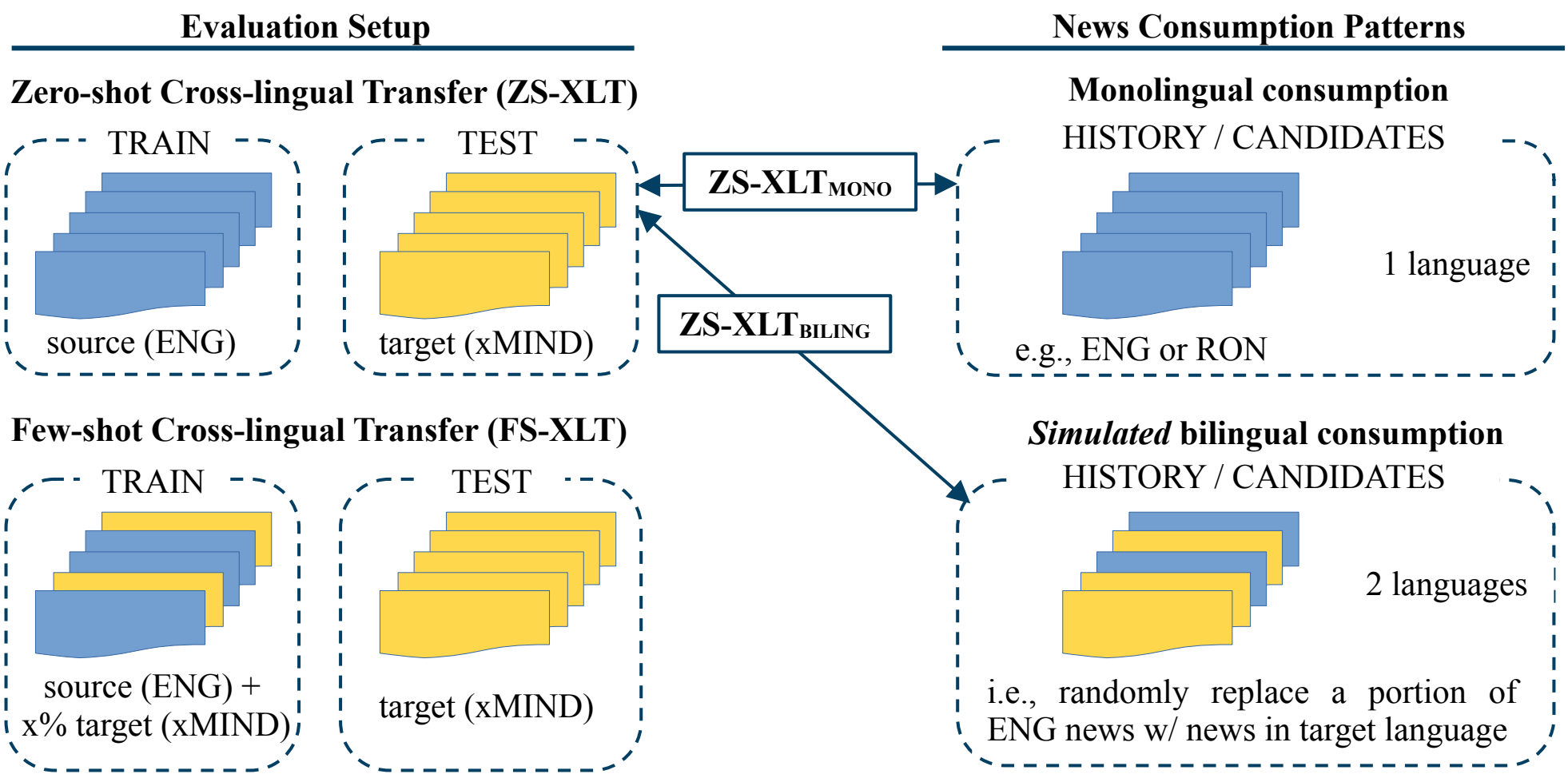
Monolingual consumption



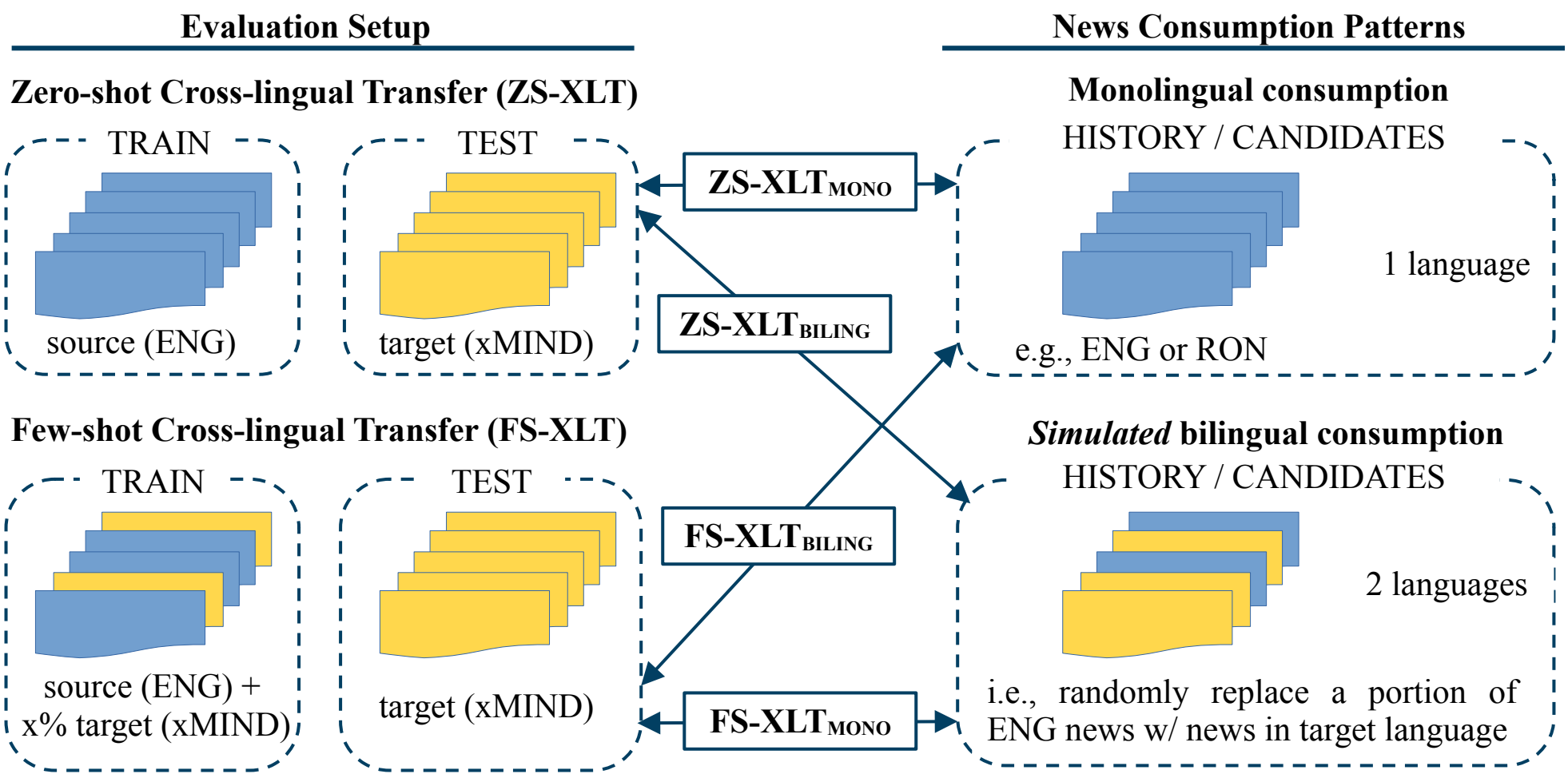
Simulated bilingual consumption



Cross-Lingual News Recommendation



Cross-Lingual News Recommendation



Zero-Shot Cross-Lingual News Recommendation

Monolingual News Consumption

Average across all xMIND languages

Language & content-agnostic

Model	AUC			nDCG@10		
	ENG	AVG	%Δ	ENG	AVG	%Δ
NAML_{CAT}	55.46		0.0	35.81		0.0
CAUM-PLM	57.82	55.90	-3.32	37.49	35.96	-4.08
LSTUR-PLM	56.80	56.28	-0.92	37.45	36.03	**-3.78**
MANNeR	50.00	50.00	**0.00**	40.17	37.64	-6.28
MINER	57.73	55.81	-3.32	36.45	35.02	-3.90
MINS-PLM	**59.89**	**56.94**	-4.93	39.35	37.64	-4.35
NAML-PLM	52.85	52.49	-0.68	**40.43**	**38.38**	-5.06
TANR-PLM	54.18	53.27	-1.68	40.03	36.78	-8.11

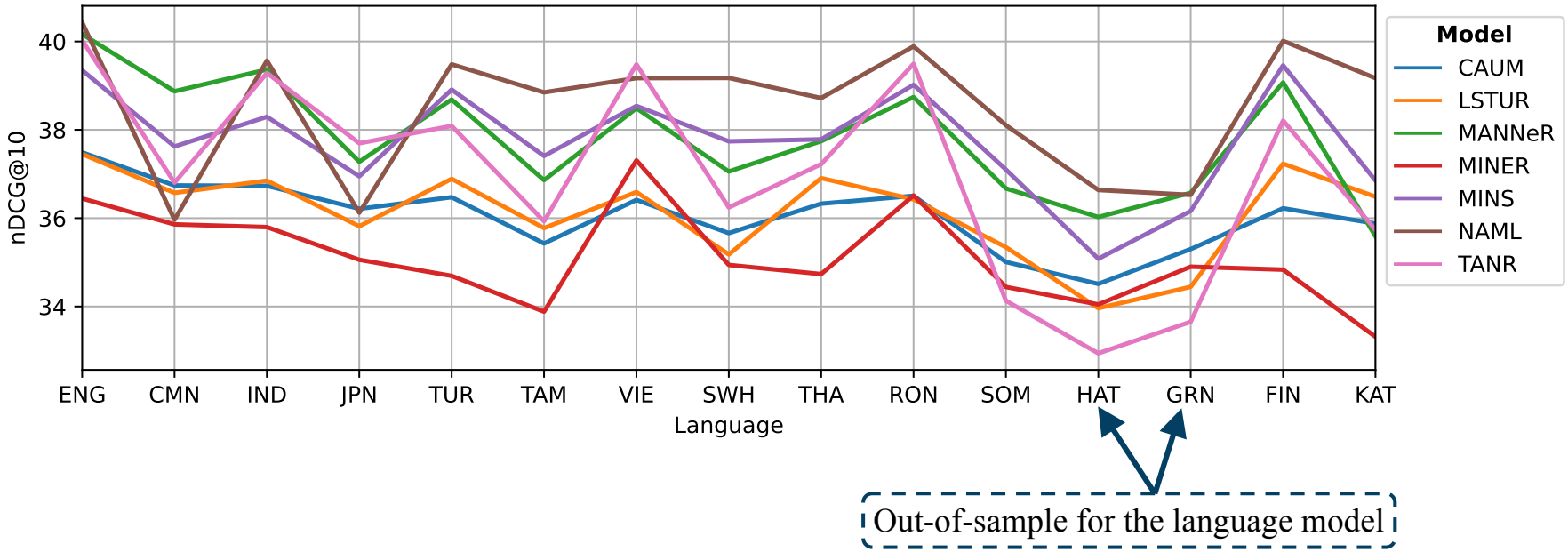
XLM-RoBERTa [6]

Relative performance w.r.t. ENG

News recommenders suffer **substantial performance losses** under zero-shot cross-lingual transfer.

Zero-Shot Cross-Lingual News Recommendation

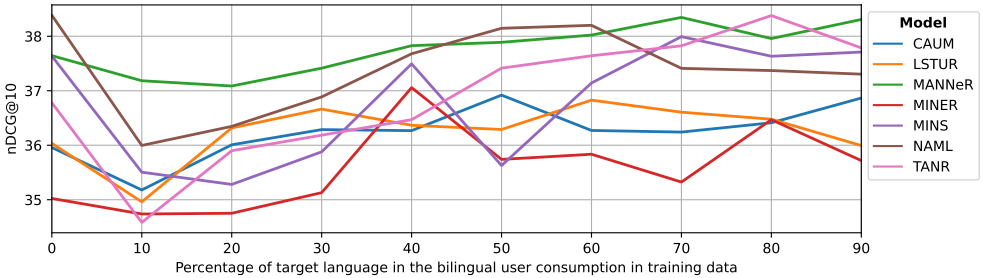
Monolingual News Consumption



Lowest performance for low-resource and unseen languages.

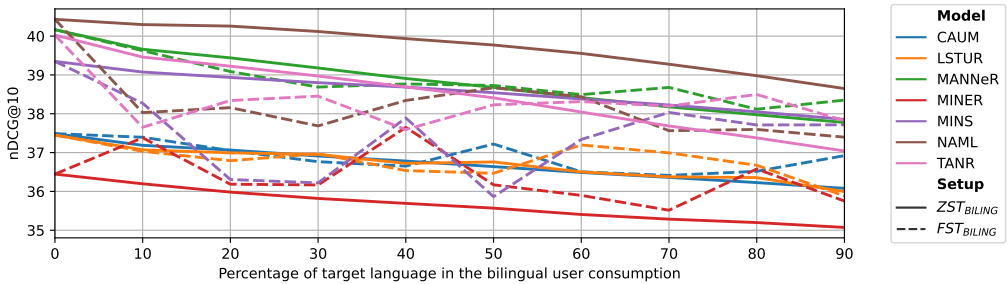
Few-Shot Cross-Lingual News Recommendation

Monolingual News Consumption



- Target language injection ameliorates performance losses from ZS-XLT$_{MONO}$
- But: over-representing one language hurts performance

Bilingual News Consumption

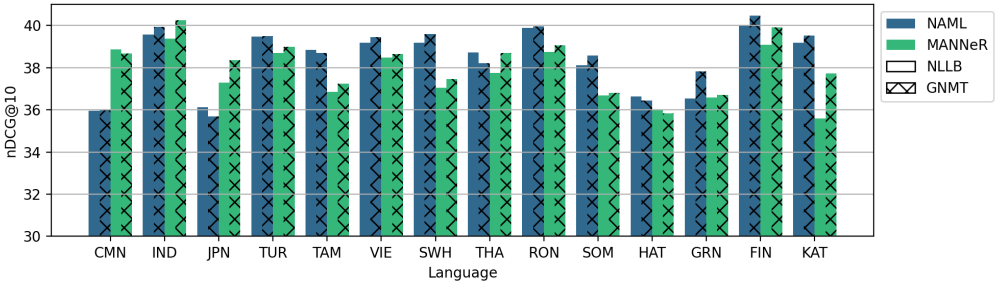


- Target language injection is beneficial primarily for languages w/ the highest losses under ZS-XLT$_{BILING}$
- Not all recommenders benefit from few-shot transfer

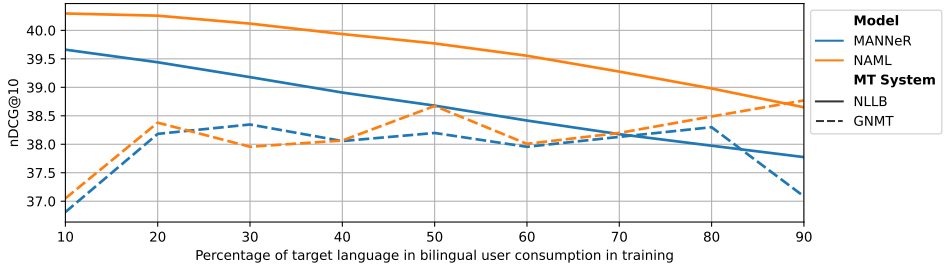
Few-shot **target-language injection** during training shows **limited benefits.**

Recommenders' Robustness to Translations

Setup: Same experiments using data translated w/ open-source vs. commercial MT



ZS-XLT$_{MONO}$ ranking performance, w.r.t. MT system, for NAML-PLM & MANNeR



FS-XLT$_{BILING}$ ranking performance, w.r.t. MT system, for NAML-PLM & MANNeR

- The quality of translations with the open-source MT is on par with those generated by SOTA commercial MT
- Translation quality has no significant effects on the recommenders' performance

Conclusion

➔ News recommendation needs (more) diverse multilingual datasets

➔ xMIND: **open multi-parallel** multilingual news recommendation dataset w/ 14 linguistically and geographically **diverse** languages, derived from the English MIND dataset using machine translation

➔ Current recommenders suffer substantial performance losses under ZS-XLT

➔ Few-shot target language injection during training brings limited gains

➔ ***More research needed on multilingual and cross-lingual news recommendation***

xMIND

xMIND @ GitHub

xMIND @ HuggingFace

Contact

References

[1] Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672* (2022).

[2] Jörg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)* (23-25), Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association (ELRA), Istanbul, Turkey

[3] Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A Multilingual Dataset for Causal Commonsense Reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2362–2376.

[4] Andreea Iana, Mehwish Alam, Alexander Grote, Katharina Luwig, Philipp Müller, Christof Weinhardt, and Heiko Paulheim. 2023. NeMig-A Bilingual News Collection and Knowledge Graph about Migration. In *Proceedings of the Workshop on News Recommendation and Analytics co-located with RecSys 2023*.

[5] Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. 2021. *Empowering news recommendation with pre-trained language models. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1652–1656.

[6] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 8440–8451