Boosting Zero-shot Cross-lingual Retrieval by Training on Artificially Code-Switched Data

CIS

nlp

Robert Litschko, Ekaterina Artemova, Barbara Plank

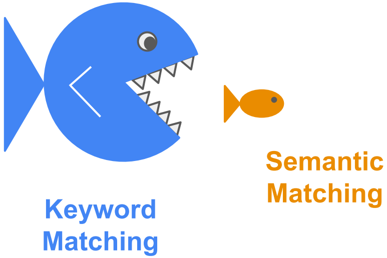
61 ACL 2023

Challenge: Monolingual Overfitting

Training

what is a death roll in crocodiles

The death roll performs a number of functions for the Saltwater Crocodile. When it grabs very large prey the crocodile has to drag it into the water and drown it so the crocodile [...] to roll over and over again to drown it's prey.



We use the mMARCO dataset (Bonifacio et al., 2021). See also Litschko et al. (2022) on monolingual overfitting.

Challenge: Monolingual Overfitting

Training	what is a death roll in crocodiles	The death roll performs a number of functions for the Saltwater Crocodile. When it grabs very large prey the crocodile has to drag it into the water and drown it so the crocodile [...] to roll over and over again to drown it's prey.	
Monolingual IR	Symptome von Fieber (symptoms of fever)	Die Liste der Anzeichen und Symptome, die in verschiedenen Quellen für Fieber erwähnt werden, umfasst die 8 unten aufgeführten Symptome: Schwitzen. Temperatur. Strenge. Brechreiz. Erbrechen. Durchfall. Lethargie.	✓

We use the mMARCO dataset (Bonifacio et al., 2021). See also Litschko et al. (2022) on monolingual overfitting.

Challenge: Monolingual Overfitting

Training	what is a death roll in crocodiles	The death roll performs a number of functions for the Saltwater Crocodile. When it grabs very large prey the crocodile has to drag it into the water and drown it so the crocodile [...] to roll over and over again to drown it's prey.	
Monolingual IR	Symptome von Fieber (symptoms of fever)	Die Liste der Anzeichen und Symptome, die in verschiedenen Quellen für Fieber erwähnt werden, umfasst die 8 unten aufgeführten Symptome: Schwitzen. Temperatur. Strenge. Brechreiz. Erbrechen. Durchfall. Lethargie.	✓
Cross-lingual IR	симптомы лихорадки (symptoms of fever)	Die Liste der Anzeichen und Symptome, die in verschiedenen Quellen für Fieber erwähnt werden, umfasst die 8 unten aufgeführten Symptome: Schwitzen. Temperatur. Strenge. Brechreiz. Erbrechen. Durchfall. Lethargie.	✗

We use the mMARCO dataset (Bonifacio et al., 2021). See also Litschko et al. (2022) on monolingual overfitting.

Method: Artificial Code-Switching

Zero-Shot Transfer

🇺🇸

Query: what is a death roll in crocodiles

Passage: the death roll performs a number of functions for the Saltwater...

Translate Train (Fine-tuning)

💸

Query: что такое список крокодилов

Passage: Die Todesrolle erfüllt für das Salzwasserkrokodil eine Reihe von Funktionen...

Method: Artificial Code-Switching

Zero-Shot Transfer

🇺🇸

Query: what is a death roll in crocodiles

Passage: the death roll performs a number of functions for the Saltwater...

Translate Train (Fine-tuning)

💸

Query: что такое список крокодилов

Passage: Die Todesrolle erfüllt für das Salzwasserkrokodil eine Reihe von Funktionen...

Bilingual Code-Switching (CS)*

Cross-lingual Word Embedding Space (Lample et al., 2018)

🇷🇺

🔻

🇺🇸

Query: что is a death roll in крокодилы

Passage: The death roll выполняет a число of функции for в Saltwater...

*Code-Switching similar to Tan and Joty (2021).

Method: Artificial Code-Switching

Zero-Shot Transfer

Query: what is a death roll in crocodiles

Passage: the death roll performs a number of functions for the Saltwater...

Translate Train (Fine-tuning)

Query: что такое список крокодилов

Passage: Die Todesrolle erfüllt für das Salzwasserkrokodil eine Reihe von Funktionen...

Bilingual Code-Switching (CS)*

Cross-lingual Word Embedding Space (Lample et al., 2018)

Query: что is a death roll in крокодилы

Passage: The death roll выполняет a число of функции for в Saltwater...

Multilingual Code-Switching (CS)*

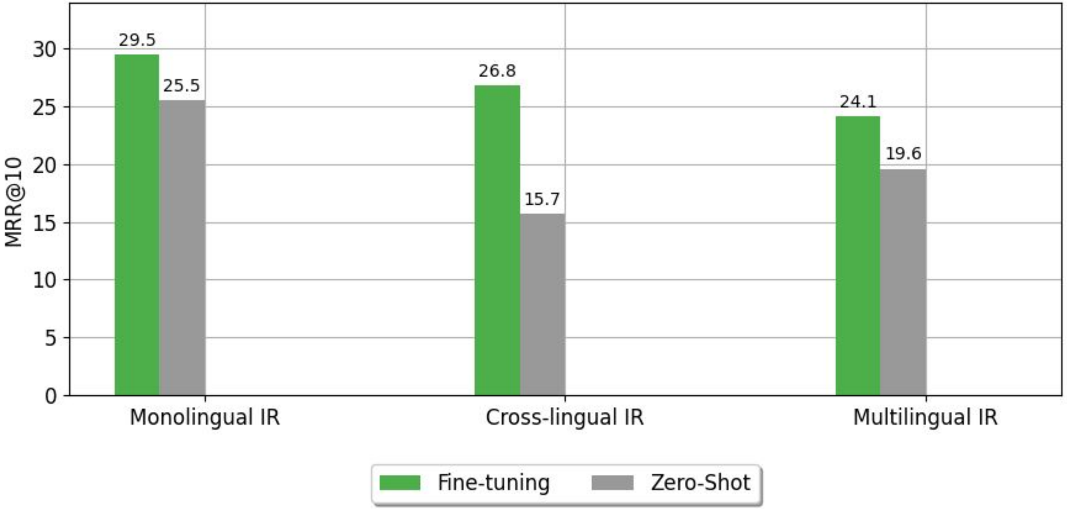
Multilingual Word Embedding Space (Lample et al., 2018)

Query: cosa is a موت rollen in крокодилы

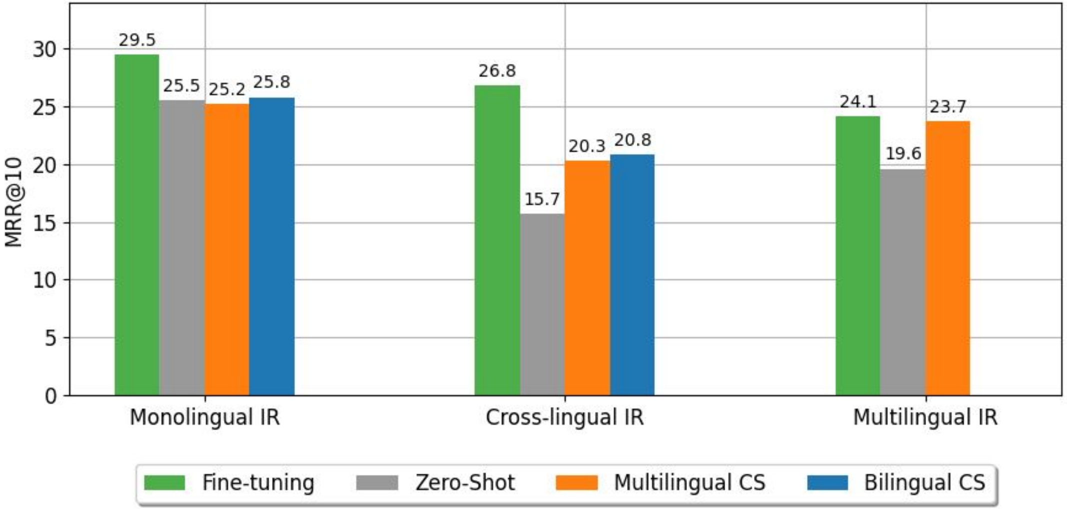
Passage: Der death rotolo performs a число of المهام for в Saltwater...

*Code-Switching similar to Tan and Joty (2021).

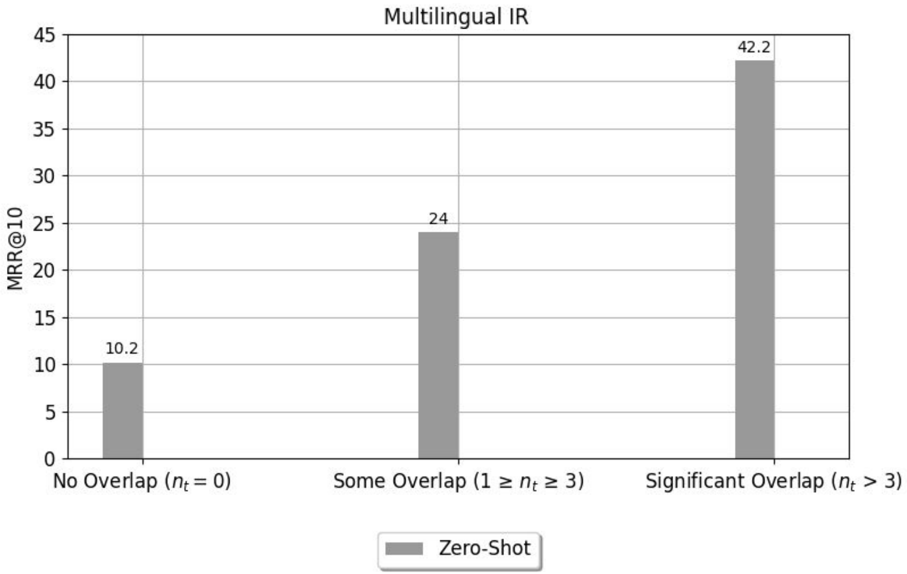
Code-Switching is **Effective**



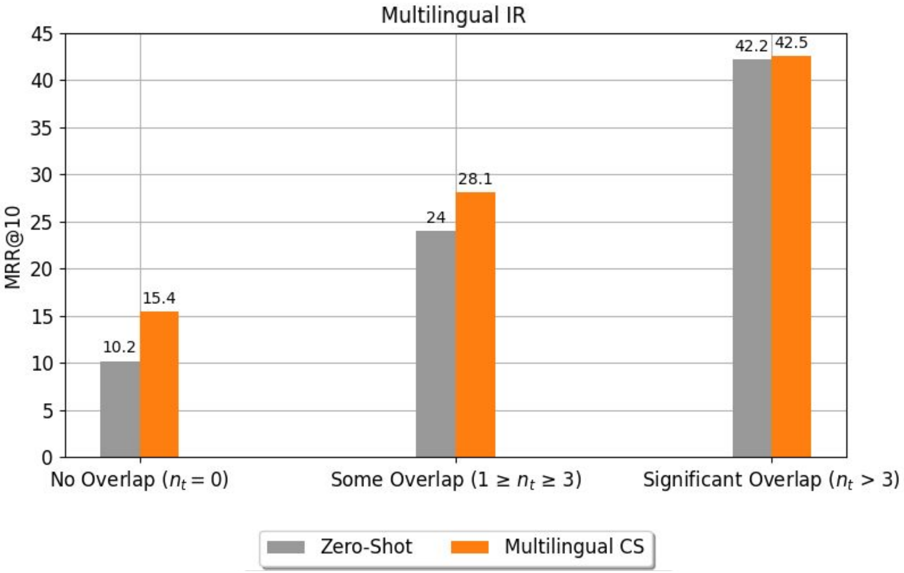
Code-Switching is **Effective**



CS mitigates **Monolingual Overfitting**



CS mitigates **Monolingual Overfitting**



More Results in our Paper

Robustness w.r.t. Translation Probability

Code-Switching with Parallel Wikipedia Titles

Eight Unseen Languages

- ID
- FR
- PT
- HI
- VT
- ZH
- ES
- JA

twitter.com/MaiNLPlab

github.com/MaiNLP/CodeSwitchCLIR

References

Luiz Bonifacio, Vitor Jeronymo, Hugo Queiroz Abonizio, Israel Campiotti, Marzieh Fadaee, Roberto Lotufo, and Rodrigo Nogueira. 2021. mmarco: A multilingual version of the ms marco passage ranking dataset. arXiv preprint arXiv:2108.13897.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In International Conference on Learning Representations.

Robert Litschko, Ivan Vulić, Simone Paolo Ponzetto, and Goran Glavaš. 2022. On cross-lingual retrieval with multilingual text encoders. Information Retrieval Journal, 25(2):149–183.

Samson Tan and Shafiq Joty. 2021. Code-mixing on sesame street: Dawn of the adversarial polyglots. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 3596–3616, Online. Association for Computational Linguistics.

Emojis designed by OpenMoji - the open-source emoji and icon project. License: CC BY-SA 4.0

Wikipedia logo by Nohat (concept by Paullusmagnus); Wikimedia Commons. License: CC BY-SA 3.0