



Robert Litschko, Ekaterina Artemova, Barbara Plank

MaiNLP, Center for Information and Language Processing (CIS), LMU Munich, Germany

{robert.litschko, ekaterina.artemova, b.plank}@lmu.de

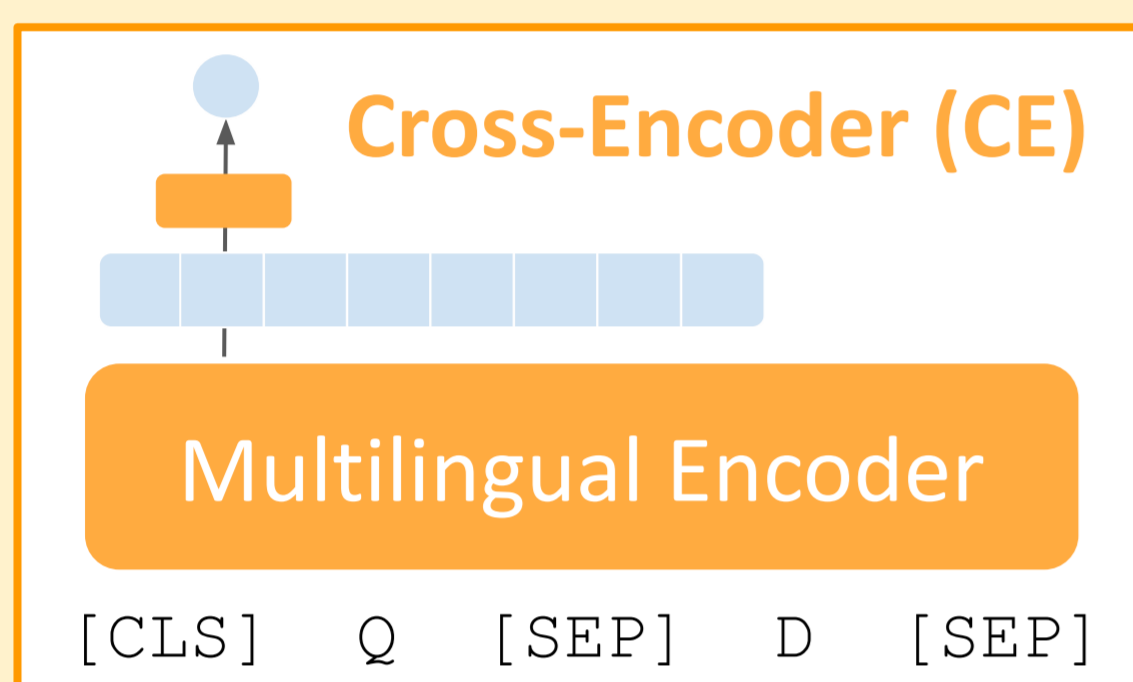
Problem Statement

- 1) Keyword overlap is a strong relevance signal and works well in **monolingual IR (MoIR)**,...
- 2) ...but falls short in **cross-lingual IR (CLIR)**.
- 3) Training zero-shot rankers on monolingual data (EN-EN) biases rankers towards learning features that cannot be exploited at test time (CLIR)
→ **Monolingual overfitting** (Litschko et al. 2022)

| | | | |
|-------|---|--|---|
| EN-EN |  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. | |
| MoIR |  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. | ✓ |
| CLIR |  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. | ✗ |

Methodology

- 1) Reranking with Cross-Encoders (Nogueira et al., 2019).



- 2) We use Code Switching (CS) to **reduce the importance of keyword matching**, we randomly replace tokens with their translation (Tan and Joty, 2021).
- 3) For this, we induce bilingual dictionaries from **Cross-lingual Word Embedding Spaces** (Lample et al., 2018).

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)

Query: что is a death roll in крокодилы
Passage: The death roll выполняет a число of функции for в Saltwater...

Multilingual Code-Switching (CS)

Query: cosa is a موت rollen in крокодилы
Passage: Der death rotolo performs a число of المهام for в Saltwater...

Main Results

Results on the mMARCO (Bonifacio et al. 2021):

- 1) CS is **effective**: gains of up to +5.1 MRR@10 in CLIR (Figure 1).
- 2) CS mitigates **monolingual overfitting**, largest gains for
 - a) queries with some token overlap and no token overlap with their relevant documents (Figure 2),
 - b) typologically distant languages with gains up to 2x in absolute performance (see paper).
- 3) CS is **robust**: gains obtainable with different translation probs. (Figure 3).

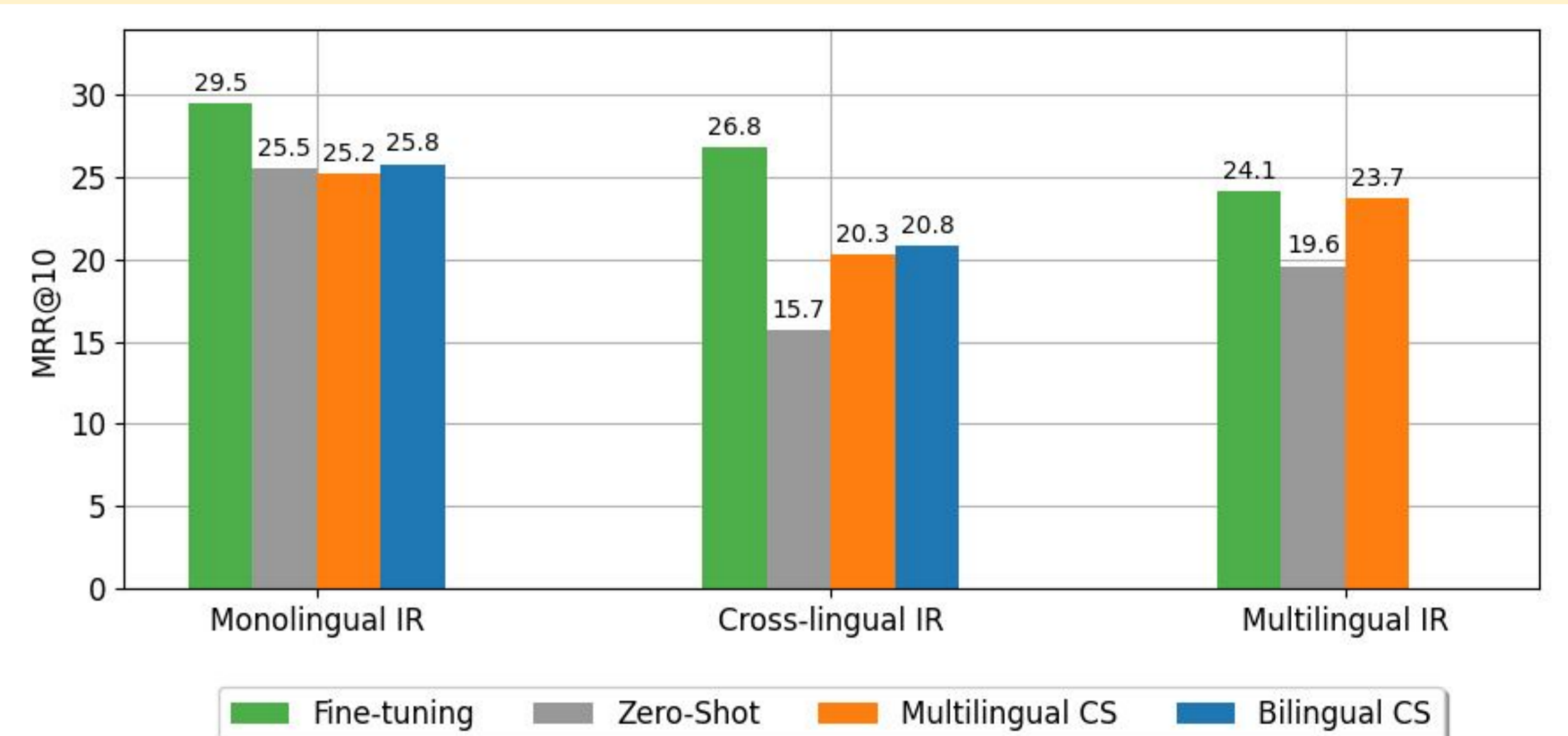


Figure 1: Results averaged over 5 (MoIR), 9 (CLIR) and 3 language pairs (MLIR).

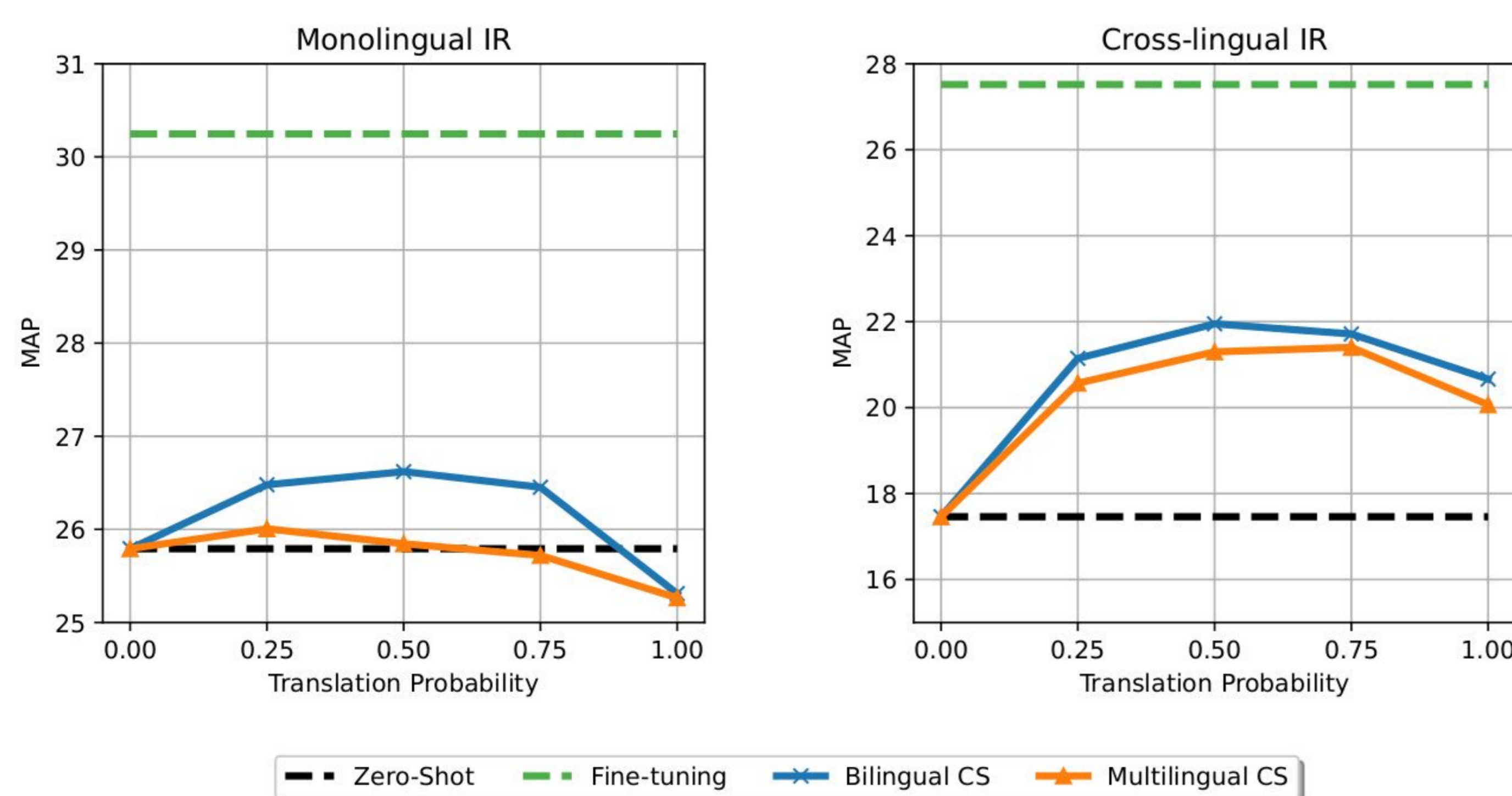


Figure 3: Retrieval performance for different translation probabilities.

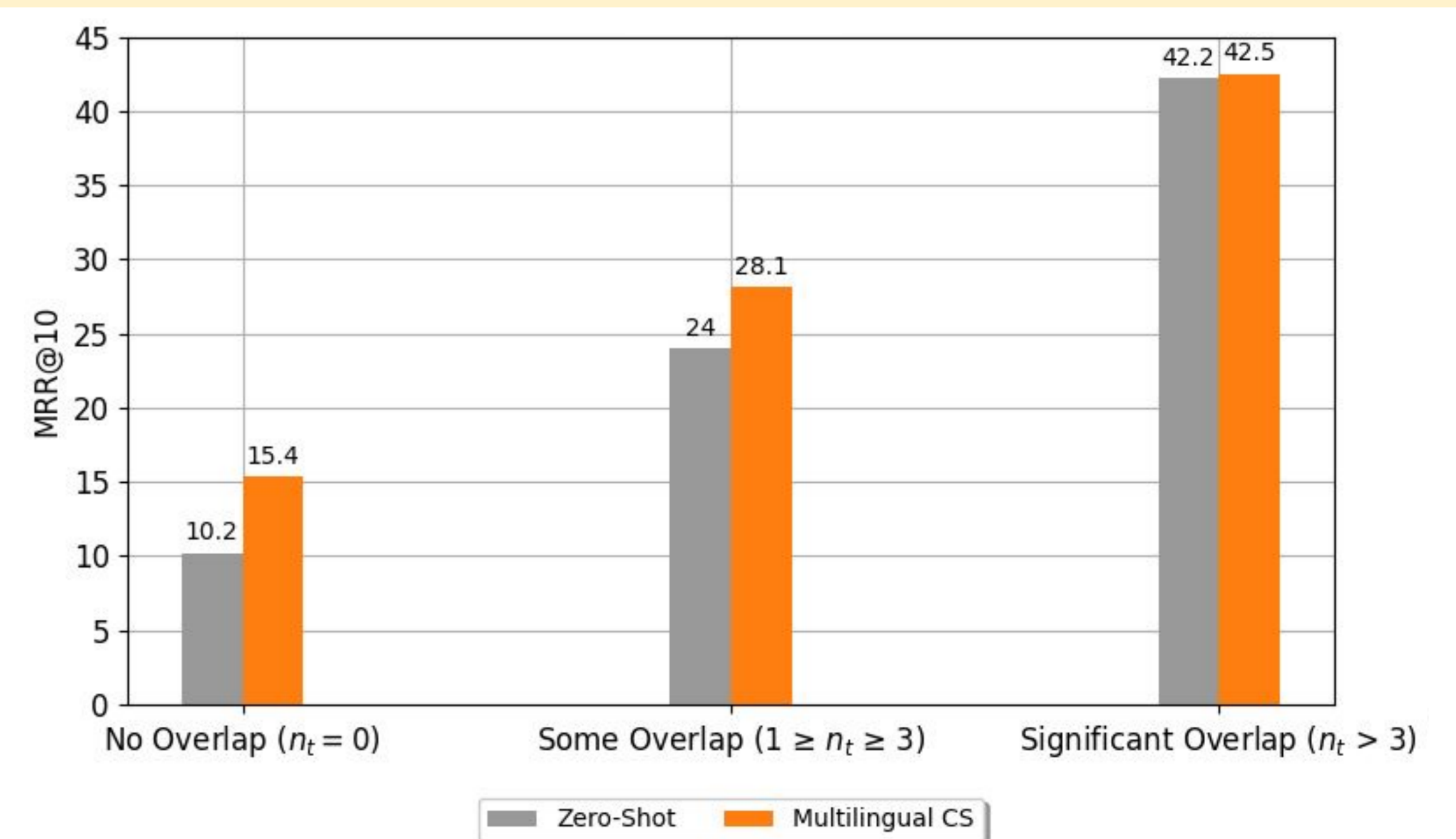


Figure 2: Multilingual IR results broken down by token overlap to relevant documents.

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 :2108.13897.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. [Word translation 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.

Rodrigo Nogueira and Kyunghyun Cho. 2019. [Passage re-ranking with bert](#). arXiv preprint arXiv:1901.04085.

Samson Tan and Shafiq Joty. 2021. [Code-mixing on sesame street: Dawn of the adversarial polyglots](#). In Proceedings of NAACL 2021, 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