

MELO: An Evaluation Benchmark for Multilingual Entity Linking of Occupations

Federico Retyk, Luis Gascó, Casimiro Pio Carrino, Daniel Deniz and Rabih Zbib

Avature Machine Learning

Abstract

We present the Multilingual Entity Linking of Occupations (MELO) Benchmark, a new collection of 48 datasets for evaluating the linking of entity mentions in 21 languages to the ESCO Occupations multilingual taxonomy. MELO was built using high-quality, pre-existent human annotations. We conduct experiments with simple lexical models and general-purpose sentence encoders, evaluated as bi-encoders in a zero-shot setup, to establish baselines for future research. The datasets and source code for standardized evaluation are publicly available at <https://github.com/Avature/melo-benchmark>.

Keywords

Entity Linking, Entity Normalization, Taxonomy Alignment, Cross-lingual, Multilingual

1. Introduction

The current trend in the digital transformation of human resources (HR) processes is the integration of artificial intelligence (AI) components that can improve automation and operational efficiency. These systems often need to process input data in the form of natural language text, which can be noisy and diverse in terms of language and other domain-specific aspects.

One common approach to deal with this challenge is the application of entity linking (EL) methods. EL helps normalizing input data into standardized entities within well-curated taxonomies. These taxonomies facilitate interoperability across different systems and, when multilingual, enable the integration of information across languages. In highly specialized domains like HR and recruiting, the development of EL methods faces significant challenges, particularly when training resources are scarce or nonexistent [1, 2]. These challenges are further amplified in multilingual environments [3, 4]. Therefore, achieving accurate entity resolution across languages is key to ensuring the consistency and effectiveness of digitalized HR systems in a global setting.

Previous research in the application of AI within the HR domain has made extensive use of taxonomies, such as occupation and skill classifications [5, 6, 7, 8, 9, 10]. These HR-specific taxonomies have been used for normalizing raw data [11, 12, 13, 14, 15, 16, 17], removing noise and enabling AI models to operate on standardized information, which in turn leads to more accurate and reliable outcomes. Substantial progress has been made, particularly in the normalization of occupational

data [18, 19, 20, 21, 22]. However, despite these advancements, there is still a surprising lack of high-quality public evaluation benchmarks for measuring progress consistently in this important area.

To address this gap, we propose the Multilingual Entity Linking of Occupations (MELO) Benchmark, a new collection of 48 datasets designed to evaluate multilingual EL tasks. This benchmark leverages pre-existing, high-quality human annotations and covers 21 languages. Furthermore, we present an experimental study using the new benchmark to evaluate the performance of both simple lexical baselines and existing deep learning models employed as zero-shot bi-encoders. Our goal is for MELO to serve as a valuable resource for advancing research and fostering innovation in this field.


The main contributions of this work are:

- We introduce the MELO Benchmark, a suite of 48 datasets involving monolingual, cross-lingual, and multilingual tasks in 21 languages. Each dataset corresponds to an entity linking task framed as a ranking problem, where queries and corpus elements are occupation names taken from a source and a target taxonomy, respectively, and binary-relevance annotations are derived from high-quality crosswalks between the taxonomies. Additionally, we release code for standardizing the evaluation of models on this benchmark.
- We provide experimental results for both simple lexical systems and state-of-the-art deep learning models evaluated as zero-shot bi-encoders on MELO, to serve as baselines for future research. We find that, while the lexical baselines perform fairly well, the semantic baselines generally achieve better results, particularly in cross-lingual tasks. However, there remains significant

RecSys in HR'24: The 4th Workshop on Recommender Systems for Human Resources, in conjunction with the 18th ACM Conference on Recommender Systems, October 14–18, 2024, Bari, Italy.

✉ machinelearning@avature.net (F. Retyk)

© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

room for improvement.

To the best of our knowledge, MELO is the first public evaluation benchmark to address the task of multilingual entity linking in the HR domain.

2. Background

In this Section, we introduce the context necessary for understanding the subsequent task definitions (§3), the methodology employed in constructing the benchmark (§4), and the related work (§6).

Entity Linking. Given a knowledge base \mathcal{E} and a query mention q , the task of Entity Linking (EL) involves identifying the correct entity $e \in \mathcal{E}$ to which the mention is referring. In principle, the structure of the knowledge base \mathcal{E} can range from a flat catalog of unrelated entities to a complex and heterogeneous ontology. In this work we focus on taxonomies of a single type of entity (i.e. occupations).

Inspired by the multilingual formulation proposed by Botha et al. [3], we consider each entity e as a language-agnostic concept with associated language-specific textual information. For each language l in a set of supported languages \mathcal{L}^{tax} , any entity may have a set of names (synonymous between each other), a description, and example sentences where the concept is used. The query q is a text string in some language l^q , with no prior assumptions about the relationship between l^q and the set \mathcal{L}^{tax} of supported languages¹.

In principle, the system may receive a query mention q that refers to an entity that does not exist in the taxonomy, or it may not refer to any entity at all. This problem, known as out-of-KB or NIL prediction [23], falls outside the scope of this work. Additionally, it is typical in the EL community to allow the system to know the textual context in which the mention occurs, aiding in the resolution of ambiguity [24]. This aspect is also beyond the scope of our work, as the data we use to build our datasets only includes unnormalized occupation names as queries.

Entity linking can be framed as a ranking task [25]: given a query q , the system produces a score $s(q, e)$ for each $e \in \mathcal{E}$ and the predicted entity \hat{e} is computed as:

$$\hat{e}(q) = \operatorname{argmax}_{e \in \mathcal{E}} s(q, e)$$

and rank-based evaluation metrics can be used to study the performance. A typical approach to this task breaks it into two stages. The first is the *Candidate Generation Stage*, where an initial ranking is obtained using a low-latency method, trying to optimize for recall. In the

¹For example, setting $\mathcal{L}^{tax} = \{l^q\}$ would result in a monolingual task, and $\mathcal{L}^{tax} = \{l^x\}$ with $l^q \neq l^x$ involves a cross-lingual task. More generally, a set \mathcal{L}^{tax} with higher cardinality can define a multilingual task.

second stage, the *Re-ranking Stage*, a more costly but higher-precision method is applied to evaluate the top elements in the preliminary rank.

Obtaining annotated data for training such systems is costly, particularly for tasks involving custom taxonomies or low-resource languages [4]. To mitigate this problem, many techniques have been proposed for leveraging transfer learning to obtain good performance in zero-shot EL scenarios [1, 2]. State-of-the-art methods typically use a bi-encoder for the candidate generation stage, and a cross-encoder for the re-ranking stage.

Multilingual Taxonomies. For the purposes of this work, we define a taxonomy \mathcal{E} as a directed acyclic graph (DAG) where nodes are concepts and edges represent binary IS-A relationships [26] between concepts. The tail concept (child) is a hyponym of the head concept (parent) and therefore represents a narrower meaning. Conversely, the parent is a hypernym of the child and represents a broader meaning, i.e. a category to which the child belongs. Concepts are allowed to have many parents.

In a multilingual taxonomy, concepts are language-agnostic but they have language-specific properties, such as a set of names, a description, or usage examples. In other words, every concept has one set of names for each language supported in the taxonomy. The set of names for a concept for a language are considered synonyms between each other. If a lexical entry is attached to more than one concept, this implies polysemy.

Occupation Taxonomies. Several public occupation taxonomies were developed to classify, standardize, and organize information related to job titles and roles found in the workforce.

One popular and influential occupations taxonomy is the European Skills, Competences, Qualifications, and Occupations (ESCO) ontology, a collection of multilingual and interrelated taxonomies created and maintained by the European Union [27, 28]. It includes 3,039 occupation concepts in its latest version, each with names and definitions (descriptions) in 28 languages. Every concept has one or more names in every supported language. The names are compliant with the terminological guidelines defined by ESCO [29]. All the names of a particular concept in a particular language are considered synonyms with each other. Also, for a particular concept, the language-specific name sets can be considered parallel data from a translation point of view.

Another important example is the O*NET-SOC taxonomy. The Occupational Information Network (O*NET) is developed and maintained by the United States government [30, 31] to standardize information relevant to the labor market, based on the 2018 Standard Occupational Classification (SOC) system². It contains information

²<https://www.bls.gov/soc/>

in English about 1,016 occupations, each with a set of names and a description.

Additionally, many other countries have developed their own national taxonomies or terminologies for occupations. For example, the Federal Employment Agency in Germany developed the *Klassifikation der Berufe 2010* (KldB 2010) which is a terminology used to standardize the information in the German language about occupations [32].

To achieve interoperability between some of these taxonomies, mappings—also called **crosswalks**—were developed and made public. These mappings establish an alignment between two given taxonomies. In particular, the European Union published many crosswalks [33] that map concepts from national taxonomies, which are typically monolingual, into ESCO. The process described in Section 4 uses this information as a gold standard to create the datasets for the MELO Benchmark.

3. Task

As mentioned already, the task consists of multilingual Entity Linking of occupations into the ESCO taxonomy, which we denote by \mathcal{E} . Given a query mention q , which is a text string expressing the non-normalized name of an occupation without surrounding context, we need to find the best semantic match in ESCO, namely the correct entity $e \in \mathcal{E}$. Every occupation in the taxonomy has textual information in all languages $l \in \mathcal{L}^{tax}$. The query is expressed in language l^q , which we make no prior assumptions about.

For evaluation, we operationalize the task as a ranking problem with binary-relevance annotations, where a query q is used to rank all the strings c_i in a corpus \mathcal{C} . The corpus is a collection of lexical terms denoting occupation names, and it is derived from the taxonomy \mathcal{E} .

To build the corpus \mathcal{C} , we first define the set of target languages for the corpus, as a subset $\mathcal{L}^c \subset \mathcal{L}^{tax}$. Then, we collect every surface form (name) for every occupation corresponding to those languages. That is, starting from an empty set, we traverse \mathcal{E} and, for each occupation e , we add every name available in any language in \mathcal{L}^c . As a result, \mathcal{C} is the collection of every name of every occupation in every target language.

The annotations consist of the set of relevant corpus elements for each query. Given the correct entity e for a query q , then those corpus elements c_i that were obtained from the surface forms of e are considered to be relevant, while any other element in the corpus is considered irrelevant.

Because the goal is to find the relevant concept e in the taxonomy for the given query (i.e. to solve the entity linking formulation of the task), obtaining at least one surface form c_i associated with the relevant concept at

the top of the ranking is sufficient for correctly performing the task. In other words, when ranking the corpus elements for a query, the position in the ranking of the highest-ranked relevant surface form is the measure we aim to evaluate. For this reason, we evaluate the baseline models with the following metrics: mean reciprocal rank (MRR) and top- k accuracy ($A@k$).

4. Datasets

The MELO Benchmark consists of 48 datasets, where each is an instance of the ranking task as described in Section 3. While the set of queries differs among the datasets, the target taxonomy is always ESCO Occupations. Although the underlying concepts in the corpus are the same, the surface forms—specifically, the occupation names—vary across datasets, since they are presented in different subsets of ESCO languages.

We leverage existing crosswalks³, which are high-quality mappings between ESCO Occupations and other taxonomies [34, 33], to build the datasets. Two datasets are derived from the mapping between ESCO and the O*NET-SOC Taxonomy, while the remaining ones are derived from the mapping between ESCO and the official occupation terminologies from several European countries. While ESCO is a multilingual taxonomy, the national terminologies are monolingual. Elements between the taxonomies are assigned SKOS relationships [35] such as *exact match*, *narrow match*, *broad match*, or *close match*.

For each crosswalk, we build two evaluation datasets: a monolingual dataset and a cross-lingual dataset. In both cases, the set of queries are those elements in the national terminologies (or O*NET) that either have only one *exact match* in ESCO or have zero *exact matches* and only one *narrow match*. Therefore, we are filtering out semantically ambiguous queries, e.g. if they have more than one *exact matches*, or that can't be assigned to a specific concept in ESCO because they are not specific enough, for example if they only have *broad* or *close matches*.

The language of the set of queries, l^q , depends on the national terminology. Regarding the languages used for the corpus, we select a different subset of the languages in ESCO for each modality. For the monolingual task we set $\mathcal{L}^c = \{l^q\}$, and for the cross-lingual we set $\mathcal{L}^c = \{\text{English}\}$. Exceptionally, since for O*NET the query language is already English, in this case instead of a cross-lingual task we define a multilingual task, where the corpus languages are English, German, Spanish, French, Italian, Dutch, Portuguese, and Polish (We intentionally include English, the query language.) As mentioned in the previous Section, the annotations

³<https://esco.ec.europa.eu/en/use-esco/eures-countries-mapping-tables>

Table 1

Datasets in the MELO Benchmark. † USA-en-xx is the only multilingual dataset. \mathcal{L}^{xx} denotes the set of languages of the elements in the corpus: English, German, Spanish, French, Italian, Dutch, Portuguese, and Polish.

Task Name	Source Taxonomy	Queries		Target Taxonomy	Corpus Elements	
		Language	#		Language	#
USA-en-en	O*NET	en	633	ESCO v1.1.0	en	33,813
USA-en-xx†	O*NET	en	633	ESCO v1.1.0	\mathcal{L}^{xx}	150,140
AUT-de-de	Austria	de	1,120	ESCO v1.1.0	de	19,782
AUT-de-en	Austria	de	1,120	ESCO v1.1.0	en	33,813
BEL-fr-fr	Belgium	fr	328	ESCO v1.0.3	fr	15,227
BEL-fr-en	Belgium	fr	328	ESCO v1.0.3	en	33,609
BEL-nl-nl	Belgium	nl	328	ESCO v1.0.3	nl	24,070
BEL-nl-en	Belgium	nl	328	ESCO v1.0.3	en	33,609
BGR-bg-bg	Bulgaria	bg	4,438	ESCO v1.0.3	bg	21,082
BGR-bg-en	Bulgaria	bg	4,438	ESCO v1.0.3	en	33,609
CZE-cs-cs	Czechia	cs	988	ESCO v1.0.9	cs	13,333
CZE-cs-en	Czechia	cs	988	ESCO v1.0.9	en	33,583
DEU-de-de	Germany	de	1,779	ESCO v1.0.3	de	19,135
DEU-de-en	Germany	de	1,779	ESCO v1.0.3	en	33,609
DNK-da-da	Denmark	da	734	ESCO v1.0.8	da	10,410
DNK-da-en	Denmark	da	734	ESCO v1.0.8	en	33,583
ESP-es-es	Spain	es	1,580	ESCO v1.0.8	es	16,502
ESP-es-en	Spain	es	1,580	ESCO v1.0.8	en	33,583
EST-et-et	Estonia	et	1,068	ESCO v1.0.8	et	4,956
EST-et-en	Estonia	et	1,068	ESCO v1.0.8	en	33,583
FRA-fr-fr	France	fr	1,435	ESCO v1.0.9	fr	15,217
FRA-fr-en	France	fr	1,435	ESCO v1.0.9	en	33,583
HRV-hr-hr	Croatia	hr	2,347	ESCO v1.0.3	hr	17,390
HRV-hr-en	Croatia	hr	2,347	ESCO v1.0.3	en	33,609
HUN-hu-hu	Hungary	hu	362	ESCO v1.0.8	hu	16,923
HUN-hu-en	Hungary	hu	362	ESCO v1.0.8	en	33,583
ITA-it-it	Italy	it	362	ESCO v1.0.8	it	16,199
ITA-it-en	Italy	it	362	ESCO v1.0.8	en	33,583
LTU-lt-lt	Lithuania	lt	3,849	ESCO v1.0.8	lt	17,824
LTU-lt-en	Lithuania	lt	3,849	ESCO v1.0.8	en	33,583
LVA-lv-lv	Latvia	lv	3,251	ESCO v1.0.8	lv	9,733
LVA-lv-en	Latvia	lv	3,251	ESCO v1.0.8	en	33,583
NLD-nl-nl	Netherlands	nl	2,605	ESCO v1.0.3	nl	24,070
NLD-nl-en	Netherlands	nl	2,605	ESCO v1.0.3	en	33,609
NOR-no-no	Norway	no	96	ESCO v1.0.8	no	7,821
NOR-no-en	Norway	no	96	ESCO v1.0.8	en	33,583
POL-pl-pl	Poland	pl	1,937	ESCO v1.0.3	pl	8,879
POL-pl-en	Poland	pl	1,937	ESCO v1.0.3	en	33,609
PRT-pt-pt	Portugal	pt	379	ESCO v1.0.3	pt	11,671
PRT-pt-en	Portugal	pt	379	ESCO v1.0.3	en	33,609
ROU-ro-ro	Romania	ro	3,273	ESCO v1.0.8	ro	14,833
ROU-ro-en	Romania	ro	3,273	ESCO v1.0.8	en	33,583
SVK-sk-sk	Slovakia	sk	2,040	ESCO v1.0.8	sk	12,899
SVK-sk-en	Slovakia	sk	2,040	ESCO v1.0.8	en	33,583
SVN-sl-sl	Slovenia	sl	3,222	ESCO v1.0.8	sl	15,487
SVN-sl-en	Slovenia	sl	3,222	ESCO v1.0.8	en	33,583
SWE-sv-sv	Sweden	sv	2,883	ESCO v1.1.1	sv	7,506
SWE-sv-en	Sweden	sv	2,883	ESCO v1.1.1	en	33,802

consist of relevancy pairs, where the set of corpus elements that correspond to the correct occupation entity for a particular query are marked as relevant, while all other corpus elements are irrelevant.

To illustrate this with an example, given the national terminology of France, we use the corresponding cross-walk to build two datasets: the monolingual dataset,

where both the queries and the corpus elements are in French, and a cross-lingual dataset, where the queries are in French but the corpus elements are in English. We name these datasets FRA-fr-fr and FRA-fr-en, respectively. In Table 1 we list all the datasets in the benchmark, with information about the languages and number of elements in their query and corpus element sets. For

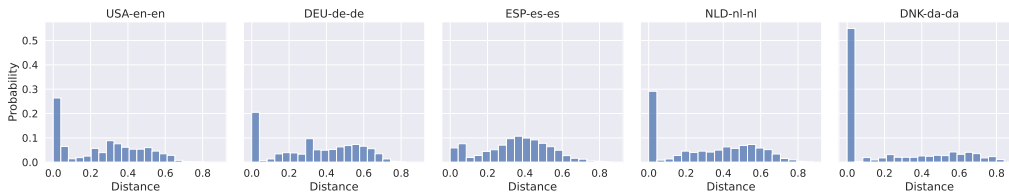


Figure 1: Histogram of minimum (normalized) edit distances between each query and the closest relevant corpus element for a selection of monolingual tasks in MELO.

further detail on the construction and composition of these datasets, as well as example queries and relevant corpus elements, please refer to Appendix A.

The benchmark is intended to represent realistic use cases, such as linking mentions into a taxonomy, enriching a custom taxonomy with new synonyms for the existing concepts, or aligning two taxonomies. It is also intended to study the cross-lingual and multilingual capabilities of proposed systems. Using extra information for solving this task, such as context for the mentions or descriptions and examples for the taxonomy concepts, is out of the scope of this work but represents an interesting line of future research that can take advantage of the MELO Benchmark.

To assess the lexical overlap between the surface forms in any national terminology and ESCO, we use the monolingual tasks, and measure the normalized edit distance between each query and the closest relevant corpus element. In Figure 1 we show a histogram with the distribution of such distances in a selection of tasks.

The lexical overlap is considerable in some cases, like with the Danish terminology. In the histogram, a big concentration of examples in the left-most bin implies that many queries are lexically very close to their relevant corpus elements. This, in principle, would make these tasks easier to solve using simple lexical scoring functions. In Appendix A we explain the procedure used to compute the lexical distances and we also present the same analysis for every task in the benchmark.

5. Experiments

To demonstrate the MELO Benchmark in use, we study the performance of several models when evaluated on the tasks we defined above. We explore both simple lexical baselines and advanced deep learning models using a bi-encoder, zero-shot setting.

Lexical Baselines. We evaluate the following baselines: edit-distance, word-level TF-IDF, word-level TF-IDF on lemmas, char-level TF-IDF, char-level TF-IDF on lemmas, BM25, and BM25 on lemmas. These models rely on surface-level text features.

Semantic Baselines. Additionally, we provide results for zero-shot evaluations using state-of-the-art deep learning models employed as symmetric bi-encoders. Under this setup, we use a sentence encoder to obtain a fixed-size representation for each surface form, and the score for a query and each corpus element is computed as the cosine similarity of their corresponding representations. This allows the system to capture deeper semantic relationships.

We experiment with the following pre-trained models in a zero-shot setup, without fine-tuning or in-context examples: ESCOXLN-R [10], mUSE-CNN [36], a multilingual variant of MPNet [37], BGE-M3 [38], GIST-Embedding [39], Multilingual E5 [40], E5 [41, 42], and the model text-embedding-3-large from OpenAI⁴. This selection of models represents a spectrum of trade-offs between performance and model complexity. We refer the reader to Appendix B and Table 3 for further details on the models and the inference procedure.

As described in Section 3, the goal of each task is to find the relevant concept in the taxonomy for the given query. Therefore, obtaining at least one surface form associated with the relevant concept at the top of the ranking is sufficient to achieve this goal. With that in mind, we use mean reciprocal rank (MRR) and top- k accuracy ($A@k$) as evaluation metrics.

Due to space constraints, in Table 2 we present results in terms of mean reciprocal rank (MRR) for a selected subset of tasks, while the complete set of results is provided in Table 5 and Table 6 in Appendix C.

In most monolingual datasets, the top-performing lexical baselines achieved MRR values ranging from 30% to 55%. Notably, in the French⁵ and Danish datasets, these baselines performed extraordinarily well in large part due to substantial lexical overlap, as indicated by the left-skewed distributions in Figure 4. In contrast, the Lithuanian, Norwegian, and Romanian datasets exhibited lower performance. Char-based TF-IDF variants deliver the highest performance among this group of baselines.

In a zero-shot setup, ESCOXLN-R performs poorly,

⁴<https://openai.com/index/new-embedding-models-and-api-updates>

⁵Results for every dataset are presented in Appendix C.

Table 2

Mean reciprocal rank (MRR) for each model, evaluated in the monolingual and the cross-lingual versions of a selection of tasks in MELO. † USA-en-xx is a multilingual dataset, with corpus elements that also cover the language of the query.

Model	USA		DEU		ESP		NLD		DNK	
	en-en	en-xx †	de-de	de-en	es-es	es-en	nl-nl	nl-en	da-da	da-en
Edit Distance	0.4858	0.4889	0.4392	0.0832	0.3297	0.0545	0.4275	0.0952	0.5650	0.1596
Word TF-IDF	0.3250	0.3207	0.4763	0.0388	0.2411	0.0127	0.4714	0.0460	0.5187	0.0398
Word TF-IDF (lemmas)	0.6056	0.5999	0.4666	0.0391	0.4318	0.0307	0.4674	0.0435	0.5179	0.0404
Char TF-IDF	0.5800	0.5764	0.5442	0.1301	0.4376	0.1238	0.4862	0.1281	0.5809	0.1576
Char TF-IDF (lemmas)	0.5957	0.5913	0.5474	0.1278	0.4697	0.1347	0.4811	0.1321	0.5801	0.1551
BM25	0.2936	0.2814	0.3377	0.0050	0.1916	0.0073	0.4433	0.0338	0.4987	0.0296
BM25 (lemmas)	0.6004	0.5978	0.4473	0.0198	0.4367	0.0275	0.4320	0.0393	0.5125	0.0334
ESCOXLM-R	0.3450	0.3426	0.4087	0.1002	0.2476	0.0854	0.3184	0.0829	0.3631	0.1095
mUSE-CNN	0.5532	0.5317	0.5606	0.3138	0.4176	0.3217	0.4255	0.2666	0.5026	0.1680
Paraph-mMPNet	0.5876	0.5822	0.4691	0.0916	0.3417	0.0899	0.3831	0.0955	0.4602	0.1148
BGE-M3	0.6226	0.6301	0.6083	0.3344	0.4927	0.3084	0.5045	0.3033	0.5839	0.3037
GIST-Embedding	0.6431	0.6464	0.5363	0.1325	0.3574	0.1534	0.4487	0.1316	0.5608	0.1348
mE5	0.6563	0.6588	0.6122	0.3858	0.5021	0.3480	0.5059	0.3246	0.5983	0.3325
E5	0.6735	0.6777	0.6639	0.5073	0.5557	0.4628	0.5650	0.4133	0.6178	0.4053
OpenAI	0.6842	0.6872	0.6778	0.5518	0.5371	0.4859	0.5723	0.4509	0.6173	0.4506

even falling behind simple lexical baselines across both monolingual and cross-lingual datasets. This result is consistent with previous research that has shown that encoders trained with masked language modeling (MLM) objectives often struggle to produce effective sentence representations when directly evaluated as sentence encoders [43]. In contrast, the other bi-encoders evaluated in this study were specifically optimized for generating useful sentence embeddings, which explains their superior performance in these tasks.

The mUSE-CNN model demonstrates fair performance on most monolingual tasks for languages included in its pre-training, especially when considering its relatively small model size and architecture type (see Table 3). However, as anticipated, its performance drops significantly for languages that were not included during its pre-training. Furthermore, its performance falls below the lexical baselines in almost all datasets. This can be observed in Figure 2b.

MPNet exhibits poor performance across all monolingual datasets, a surprising result given its larger model size, architecture type, and the fact that it was pre-trained in all the languages used in this experiment. Despite these advantages, it is generally outperformed by the smaller mUSE-CNN model, with the notable exception of the English datasets.

BGE-M3 and Multilingual E5 have similar characteristics, as described in Table 3, and both deliver strong performance across most monolingual tasks. In these cases, they generally outperform all lexical baselines and smaller bi-encoders. However, in the English datasets, Multilingual E5 outperforms BGE-M3.

GIST-Embedding demonstrates strong performance in English, outperforming many larger models. It also

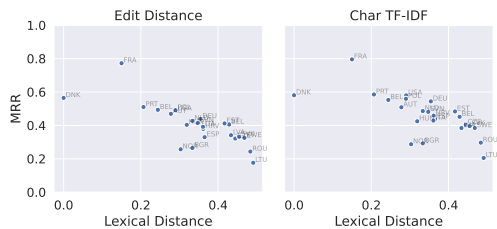
achieves reasonable results in most other languages, which is surprising considering its primary training was focused on English.

E5, a significantly larger Decoder-only model, outperforms the previously mentioned models across most tasks. This is also surprising since E5 was mainly trained in English. Finally, although limited details are available publicly about OpenAI’s text-embedding-3-large model, its performance is generally on par with or even surpasses that of E5. OpenAI’s model delivers the highest overall performance among all the models evaluated in our experiments.

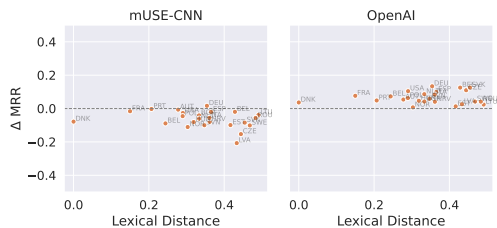
The performance of the models in each monolingual dataset is correlated with the lexical overlap in the dataset, as measured by the median of the distributions presented in Figure 4. As expected, lexical baselines exhibit a particularly strong correlation, with Spearman’s coefficients of -0.74 for Char TF-IDF and -0.80 for Edit Distance. Interestingly, bi-encoders also demonstrate a moderate correlation, such as mE5 (-0.65) and OpenAI (-0.62). In Figure 2 we visualize this correlation, as well as the correlation between the lexical overlap and the difference in performance between some bi-encoders and a lexical baseline⁶. We observe that, the less lexical overlap in the dataset, the more the OpenAI model outperforms the lexical baseline.

Comparing the results of datasets USA-en-en and USA-en-xx, which share the same queries, we observe that most methods significantly enhance their performance when the corpus elements visible to the system are expanded to include multiple languages, surpassing their performance in the monolingual task. An implication for this is that, when linking mentions into a multilingual taxonomy, the surface forms in other languages are valu-

⁶Same figure is displayed in full size in Appendix C



(a) Absolute performance (in MRR).



(b) Performance relative to the lexical baseline Char TF-IDF

Figure 2: Correlation between model performance and the median of the minimum edit distance between queries and relevant corpus elements in monolingual datasets.

able even if the taxonomy includes entity names in the language of the query.

As expected, the performance drop when moving from monolingual to cross-lingual datasets (excluding O*NET) is significantly more pronounced for the lexical baselines compared to the bi-encoders. The capacity for (zero-shot) cross-lingual EL of occupations varies for different models: ESCOXML-R, MPNet, and GIST-Embedding exhibit very low cross-lingual performance; mUSE-CNN, BGE-M3, and Multilingual E5 demonstrate fair cross-lingual performance; while E5 and OpenAI achieve the highest cross-lingual performance.

Since the techniques we experiment with—lexical scorers and bi-encoders—are commonly used for candidate generation in the first stage of EL [1, 2], it is interesting to measure the top- k accuracy ($A@k$) for different values of k to assess how well such techniques recover the first relevant item. Figure 3 presents these results for the same subset of tasks for the following systems: Edit Distance, Char-level TF-IDF, mUSE-CNN, and OpenAI. The complete set of $A@k$ is available in Appendix C, in Figure 6 and Figure 7. The results observed for top- k accuracy are consistent with those for mean reciprocal rank (MRR), particularly in terms of the relative ranking and comparative performance of the models.

6. Related Work

There has been significant research interest in systems that normalize HR information into ESCO and other taxonomies.

Decorte et al. [14] explore the extraction of ESCO skills from segmented job descriptions. They approach this problem as a massive multi-label classification task, and present a human-annotated evaluation set for this task. More recently, Decorte et al. [17] approach the same problem from an EL perspective. They use a large language model (LLM) to produce synthetic annotations and train a bi-encoder to extract ESCO skills from job description segments. Finally, Zhang et al. [11] apply and compare two supervised EL methods for solving the same task: BLINK [2] and GENRE [44]. In contrast to these other studies, our work focuses on occupations instead of skills, explores cross-lingual and multilingual scenarios, and the task as we formulate it does not use context for linking the query mentions.

There has also been a substantial amount of research focused on occupations. Decorte et al. [20] developed an unsupervised approach to fine-tune BERT [45] to encode the semantics of occupation names. Furthermore, they create a dataset for the normalization of free-form English occupation names into ESCO and they use it to evaluate their model. It has been reported that this dataset contains ambiguous input queries [20] as well as some mislabeled elements [46]. Closely related works by Zbib et al. [47] and by Bocharova et al. [48] propose alternative unsupervised representation learning schemes. They both release evaluation datasets, the former for occupation name ranking, and the latter for EL of unnormalized occupation names into ESCO.

Lake [16] studies the application of bi-encoders and cross-encoders to EL of occupations to a custom taxonomy. Yamashita et al. [21] work on a normalization task for occupations, which closely resembles our formulation of EL. They create a non-public dataset by collecting a large number of unnormalized occupation names and then automatically mapping them to ESCO occupations via exact match after removing proper nouns. Vrolijk et al. [22] build a synthetic dataset for zero-shot evaluation and fine-tuning of several language models using information from ESCO that includes the synonyms for each entity name, the relationship between entities, and their definitions. In particular, they use the set of name synonyms for each ESCO occupation to pose a binary relevance classification problem, where positive pairs involve two names belonging to the same synonym set.

Two important use cases of the EL task under study are enriching and aligning taxonomies. In order to maintain up-to-date but well-curated taxonomies, it is common to automatically identify new candidate concepts

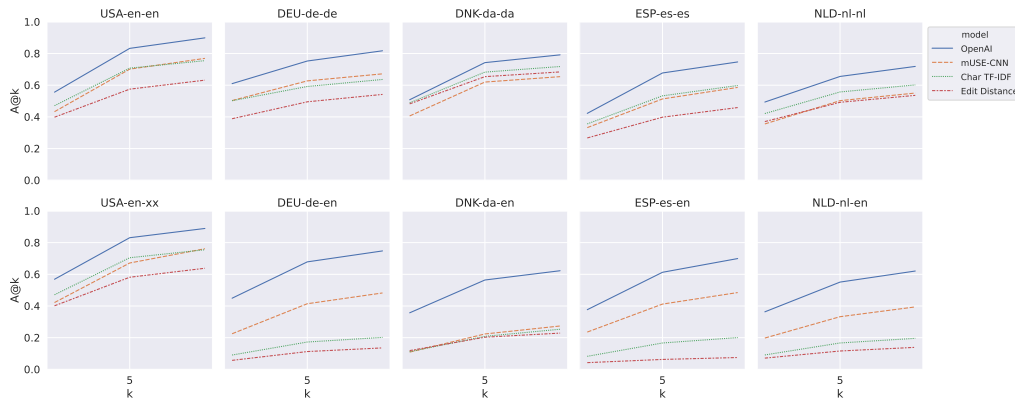


Figure 3: Top- k accuracy ($A@k$) for a selection of models in the MELO Benchmark tasks corresponding to O*NET, Germany, Spain, the Netherlands, and Denmark.

to be included, and to use human annotators to validate their inclusion. Similarly, when aligning two taxonomies—i.e. building a crosswalk—it is common to use automatic systems to propose and explore candidate matches between the concepts in each taxonomy.

Giabelli [19] and colleagues have worked on several approaches for enriching [49] and aligning [18, 50] taxonomies using word embeddings to model concepts via their names, together with structural information about the taxonomy. All these methods automatically score candidates for inclusion or mapping, and can be used within a human-in-the-loop framework for further validation.

During the creation of the crosswalk between O*NET and ESCO, the teams responsible for maintaining both taxonomies worked together to ensure a high-quality mapping [33]. Interestingly, they report employing a human-in-the-loop methodology where a fine-tuned BERT model [45] is used as a bi-encoder to rank the ESCO occupations for each O*NET occupation. They explore different methods for encoding each, leveraging occupation names (and synonyms) as well.

More recently, the ESCO team presented an analysis [46] on a task that is very similar to the one we present here. They fine-tune a XLM-RoBERTa model [51] on HR-related data, including the textual information from ESCO, but with no supervision signal for any specific EL task. They then use this model as a bi-encoder to suggest ESCO occupations for elements taken from the national terminologies of Latvia, Spain, Sweden, and Italy, as well as from O*NET. Using the respective crosswalks, they evaluate this as an EL task. They explore monolingual and cross-lingual (to English) modalities. A key difference between this work and ours is that they consider any SKOS relationship as a legitimate annotation, while we only use exact and narrow matches. We also filter

out semantically ambiguous queries for which experts determined that they should be related as an exact match to more than one ESCO concept. For those reasons, their results are not comparable to those we present in this work.

7. Conclusion

We have introduced the MELO Benchmark, a suite of 48 datasets for multilingual entity linking of occupations in 21 languages. We experimented with several out-of-the-box lexical and semantic baselines, demonstrating that there is still significant room for improvement. Our aim is that MELO will serve as a valuable resource for the research community, providing a standardized benchmark for assessing progress in multilingual EL within the HR domain, and fostering innovation and the development of new methodologies in this important area of research.

In future work, several research directions could be explored. First, the current evaluation scheme can be extended to incorporate NIL prediction or prediction using entity descriptions rather than relying solely on entity names, with the presented source code being easily adaptable for such modifications. Second, domain-adapting or fine-tuning encoders specifically for this task, in a manner similar to ESCOXLM-R but optimized for semantic text similarity, presents another possible direction. Third, exploring advanced deep learning techniques beyond bi-encoders, such as cross-encoders combined with re-ranking stages, could enhance model performance. Finally, investigating the meta-learning paradigm by dividing MELO tasks into meta-training and meta-testing tasks, and applying meta-learning context to solve the meta-testing tasks, exploiting multi-lingual transfer capabilities of modern deep-learning models, offers another

interesting direction for future work.

Acknowledgment

This publication uses the ESCO classification of the European Commission. We gratefully acknowledge the work done by the team involved in curating the ESCO Occupations taxonomy, as well as the teams responsible for the O*NET-SOC 2019 taxonomy and the other national taxonomies used in this work. Furthermore, we would also like to thank the teams responsible for creating the crosswalks between ESCO and these taxonomies.

References

- [1] L. Logeswaran, M.-W. Chang, K. Lee, K. Toutanova, J. Devlin, H. Lee, Zero-Shot Entity Linking by Reading Entity Descriptions, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 3449–3460. URL: <https://aclanthology.org/P19-1335>. doi:10.18653/v1/P19-1335.
- [2] L. Wu, F. Petroni, M. Josifoski, S. Riedel, L. Zettlemoyer, Scalable Zero-shot Entity Linking with Dense Entity Retrieval, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 6397–6407. URL: <https://aclanthology.org/2020.emnlp-main.519>. doi:10.18653/v1/2020.emnlp-main.519.
- [3] J. A. Botha, Z. Shan, D. Gillick, Entity Linking in 100 Languages, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 7833–7845. URL: <https://aclanthology.org/2020.emnlp-main.630>. doi:10.18653/v1/2020.emnlp-main.630.
- [4] X. Fu, W. Shi, X. Yu, Z. Zhao, D. Roth, Design Challenges in Low-resource Cross-lingual Entity Linking, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 6418–6432. URL: <https://aclanthology.org/2020.emnlp-main.521>. doi:10.18653/v1/2020.emnlp-main.521.
- [5] M. de Groot, J. Schutte, D. Graus, Job Posting-Enriched Knowledge Graph for Skills-based Matching, The 1st Workshop on Recommender Systems for Human Resources (RecSys in HR’21), in conjunction with the 15th ACM Conference on Recommender Systems (2021). URL: https://ceur-ws.org/Vol-2967/paper_3.pdf.
- [6] S. Tu, O. Cannon, Beyond Human-in-the-loop: Scaling Occupation Taxonomy at Indeed, The 2nd Workshop on Recommender Systems for Human Resources (RecSys in HR’22), in conjunction with the 16th ACM Conference on Recommender Systems (2022). URL: https://recsyshr.aau.dk/wp-content/uploads/2022/09/RecSysHR2022-paper_2.pdf.
- [7] S. Avlonitis, D. Lavi, M. Mansoury, D. Graus, Career Path Recommendations for Long-term Income Maximization: A Reinforcement Learning Approach, The 3rd Workshop on Recommender Systems for Human Resources (RecSys in HR’23), in conjunction with the 17th ACM Conference on Recommender Systems (2023). URL: https://recsyshr.aau.dk/wp-content/uploads/2023/09/RecSysHR2023-paper_2.pdf.
- [8] J.-J. Decorte, J. V. Haute, J. Deleu, C. Devellder, T. Demeester, Career Path Prediction using Resume Representation Learning and Skill-based Matching, The 3rd Workshop on Recommender Systems for Human Resources (RecSys in HR’23), in conjunction with the 17th ACM Conference on Recommender Systems (2023). URL: https://recsyshr.aau.dk/wp-content/uploads/2023/09/RecSysHR2023-paper_1.pdf.
- [9] M. Zhang, K. Jensen, S. Sonniks, B. Plank, SkillSpan: Hard and Soft Skill Extraction from English Job Postings, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Seattle, United States, 2022, pp. 4962–4984. URL: <https://aclanthology.org/2022.naacl-main.366>. doi:10.18653/v1/2022.naacl-main.366.
- [10] M. Zhang, R. van der Goot, B. Plank, ESCOXMLR: Multilingual Taxonomy-driven Pre-training for the Job Market Domain, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 11871–11890. URL: <https://aclanthology.org/2023.acl-long.662>. doi:10.18653/v1/2023.acl-long.662.
- [11] M. Zhang, R. van der Goot, B. Plank, Entity Linking in the Job Market Domain, in: Findings of the Association for Computational Linguistics: EACL 2024, Association for Computational Linguistics, St. Julian’s, Malta, 2024, pp. 410–419. URL: <https://aclanthology.org/2024.findings-eacl.28>.
- [12] E. Senger, M. Zhang, R. van der Goot, B. Plank, Deep Learning-based Computational Job Market Analysis: A Survey on Skill Extraction and Classi-

- fication from Job Postings, in: Proceedings of the First Workshop on Natural Language Processing for Human Resources (NLP4HR 2024), Association for Computational Linguistics, St. Julian's, Malta, 2024, pp. 1–15. URL: <https://aclanthology.org/2024.nlp4hr-1.1>.
- [13] M. Zhang, K. N. Jensen, B. Plank, Kompetencer: Fine-grained Skill Classification in Danish Job Postings via Distant Supervision and Transfer Learning, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 436–447. URL: <https://aclanthology.org/2022.lrec-1.46>.
- [14] J.-J. Decorte, J. V. Hautte, J. Deleu, C. Develder, T. Demeester, Design of Negative Sampling Strategies for Distantly Supervised Skill Extraction, The 2nd Workshop on Recommender Systems for Human Resources (RecSys in HR'22), in conjunction with the 16th ACM Conference on Recommender Systems (2022). URL: https://recsysshr.aau.dk/wp-content/uploads/2022/09/RecSysHR2022-paper_4.pdf.
- [15] M. Zhang, K. N. Jensen, R. van der Goot, B. Plank, Skill Extraction from Job Postings using Weak Supervision, The 2nd Workshop on Recommender Systems for Human Resources (RecSys in HR'22), in conjunction with the 16th ACM Conference on Recommender Systems (2022). URL: https://recsysshr.aau.dk/wp-content/uploads/2022/09/RecSysHR2022-paper_10.pdf.
- [16] T. Lake, Flexible Job Classification with Zero-Shot Learning, The 2nd Workshop on Recommender Systems for Human Resources (RecSys in HR'22), in conjunction with the 16th ACM Conference on Recommender Systems (2022). URL: https://recsysshr.aau.dk/wp-content/uploads/2022/09/RecSysHR2022-paper_8.pdf.
- [17] J.-J. Decorte, S. Verlinden, J. V. Hautte, J. Deleu, C. Develder, T. Demeester, Extreme Multi-Label Skill Extraction Training using Large Language Models, 2023. URL: <https://arxiv.org/abs/2307.10778>. arXiv:2307.10778.
- [18] A. Giabelli, L. Malandri, F. Mercorio, M. Mezzanzanica, WETA: Automatic taxonomy alignment via word embeddings, Computers in Industry 138 (2022) 103626. URL: <https://www.sciencedirect.com/science/article/pii/S0166361522000215>. doi:<https://doi.org/10.1016/j.compind.2022.103626>.
- [19] A. Giabelli, Integrating Word Embeddings and Taxonomy Learning for Enhanced Lexical Domain Modelling, Phd thesis, Università degli Studi di Milano-Bicocca, 2024.
- [20] J.-J. Decorte, J. V. Hautte, T. Demeester, C. Develder, JobBERT: Understanding Job Titles through Skills, FEAST, ECML-PKDD 2021 Workshop (2021). URL: https://feast-ecmlpkdd.github.io/archive/2021/papers/FEAST2021_paper_6.pdf.
- [21] M. Yamashita, J. T. Shen, T. Tran, H. Ekhtiari, D. Lee, JAMES: Normalizing Job Titles with Multi-Aspect Graph Embeddings and Reasoning, in: 2023 IEEE 10th International Conference on Data Science and Advanced Analytics (DSAA), 2023, pp. 1–10. URL: <https://arxiv.org/abs/2202.10739>. doi:10.1109/DSAA60987.2023.10302559.
- [22] J. Vrolijk, D. Graus, Enhancing PLM Performance on Labour Market Tasks via Instruction-based Finetuning and Prompt-tuning with Rules, The 3rd Workshop on Recommender Systems for Human Resources (RecSys in HR'23), in conjunction with the 17th ACM Conference on Recommender Systems (2023). URL: https://recsysshr.aau.dk/wp-content/uploads/2023/09/RecSysHR2023-paper_4.pdf.
- [23] F. Zhu, J. Yu, H. Jin, L. Hou, J. Li, Z. Sui, Learn to Not Link: Exploring NIL Prediction in Entity Linking, in: Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 10846–10860. URL: <https://aclanthology.org/2023.findings-acl.690>. doi:10.18653/v1/2023.findings-acl.690.
- [24] N. Gupta, S. Singh, D. Roth, Entity Linking via Joint Encoding of Types, Descriptions, and Context, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 2681–2690. URL: <https://aclanthology.org/D17-1284>. doi:10.18653/v1/D17-1284.
- [25] Z. Zheng, F. Li, M. Huang, X. Zhu, Learning to Link Entities with Knowledge Base, in: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Los Angeles, California, 2010, pp. 483–491. URL: <https://aclanthology.org/N10-1072>.
- [26] R. J. Brachman, What IS-A Is and Isn't: An Analysis of Taxonomic Links in Semantic Networks, Computer 16 (1983) 30–36. doi:10.1109/MC.1983.1654194.
- [27] M. le Vrang, A. Papantoniou, E. Pauwels, P. Fannes, D. Vandestein, J. De Smedt, ESCO: Boosting Job Matching in Europe with Semantic Interoperability, Computer 47 (2014) 57–64. doi:10.1109/MC.2014.283.
- [28] European Commission, ESCO Handbook: European Skills, Competences, Qualifications and

- Occupations, Technical Report, European Union, 2019. URL: <https://esco.ec.europa.eu/system/files/2021-07/Handbook.pdf>.
- [29] European Commission, ESCO Terminological Guidelines, Technical Report, European Union, 2021. URL: <https://esco.ec.europa.eu/en/about-esco/publications/publication/esco-terminological-guidelines>.
- [30] E. C. Dierdorff, D. W. Drewes, J. J. Norton, O*NET Tools and Technology: A Synopsis of Data Development Procedures, Technical Report, North Carolina State University, 2006. URL: https://www.onetcenter.org/dl_files/T2Development.pdf.
- [31] M. J. Handel, The O*NET Content Model: Strengths and Limitations, *Journal for Labour Market Research* 49 (2016) 157–176. doi:10.1007/s12651-016-0199-8.
- [32] W. Paulus, B. Matthes, Klassifikation der Berufe: Struktur, Codierung und Umsteigeschlüssel, Technical Report, Bundesagentur für Arbeit, 2013. URL: https://doku.iab.de/fdz/reporte/2013/MR_08-13.pdf.
- [33] European Commission, The Crosswalk Between ESCO and O*NET, Technical Report, European Union, 2022. URL: <https://esco.ec.europa.eu/system/files/2022-12/ONET%20ESCO%20Technical%20Report.pdf>.
- [34] European Commission, ESCO implementation manual, Technical Report, European Union, 2018. URL: https://esco.ec.europa.eu/system/files/2021-07/425b7a5f-3048-4377-a816-5402c00e9a9505_A_Annex_Draft_ESCO_Implementation_manual.pdf.
- [35] A. Miles, S. Bechhofer, SKOS Simple Knowledge Organization System Reference, W3C Recommendation, World Wide Web Consortium, 2009. URL: <https://www.w3.org/TR/skos-reference/>, w3C Recommendation.
- [36] Y. Yang, D. Cer, A. Ahmad, M. Guo, J. Law, N. Constant, G. Hernandez Abrego, S. Yuan, C. Tar, Y.-h. Sung, B. Strope, R. Kurzweil, Multilingual Universal Sentence Encoder for Semantic Retrieval, in: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Association for Computational Linguistics, Online, 2020, pp. 87–94. URL: <https://aclanthology.org/2020.acl-demos.12>. doi:10.18653/v1/2020.acl-demos.12.
- [37] K. Song, X. Tan, T. Qin, J. Lu, T.-Y. Liu, MP-Net: Masked and Permuted Pre-training for Language Understanding, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), *Advances in Neural Information Processing Systems*, volume 33, Curran Associates, Inc., 2020, pp. 16857–16867. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/c3a690be93aa602ee2dc0ccab5b7b67e-Paper.pdf.
- [38] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, Z. Liu, BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation, 2024. URL: <https://arxiv.org/abs/2402.03216>. arXiv:2402.03216.
- [39] A. V. Solatorio, GISTEmbed: Guided In-sample Selection of Training Negatives for Text Embedding Fine-tuning, 2024. URL: <https://arxiv.org/abs/2402.16829>. arXiv:2402.16829.
- [40] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, F. Wei, Multilingual E5 Text Embeddings: A Technical Report, 2024. URL: <https://arxiv.org/abs/2402.05672>. arXiv:2402.05672.
- [41] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, F. Wei, Improving Text Embeddings with Large Language Models, 2023. URL: <https://arxiv.org/abs/2401.00368>. arXiv:2401.00368.
- [42] L. Wang, N. Yang, X. Huang, B. Jiao, L. Yang, D. Jiang, R. Majumder, F. Wei, Text Embeddings by Weakly-Supervised Contrastive Pre-training, 2022. URL: <https://arxiv.org/abs/2212.03533>. arXiv:2212.03533.
- [43] B. Li, H. Zhou, J. He, M. Wang, Y. Yang, L. Li, On the Sentence Embeddings from Pre-trained Language Models, in: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, Online, 2020, pp. 9119–9130. URL: <https://aclanthology.org/2020.emnlp-main.733>. doi:10.18653/v1/2020.emnlp-main.733.
- [44] N. D. Cao, G. Izacard, S. Riedel, F. Petroni, Autoregressive Entity Retrieval, *International Conference on Learning Representations (2021)*. URL: <https://openreview.net/forum?id=5k8F6UU39V>.
- [45] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: <https://aclanthology.org/N19-1423>. doi:10.18653/v1/N19-1423.
- [46] European Commission, Machine Learning Assisted Mapping of Multilingual Occupational Data to ESCO, Technical Report, European Union, 2022. URL: <https://shorturl.at/REcDd>.
- [47] R. Zbib, L. A. Lacasa, F. Retyk, R. Poves, J. Aizpuru, H. Fabregat, V. Šimkus, E. García-Casademont, Learning Job Titles Similarity from Noisy Skill Labels, FEAST, ECML-PKDD 2021 Workshop (2022). URL: <https://feast-ecmlpkdd.github.io/>

- archive/2022/papers/FEAST2022_paper_4972.pdf.
- [48] M. Bocharova, E. Malakhov, V. Mezhuyev, VacancySBERT: the approach for representation of titles and skills for semantic similarity search in the recruitment domain, *Applied Aspects of Information Technology* 6 (2023) 52–59. URL: <https://aait.od.ua/index.php/journal/article/view/161/212>. doi:10.15276/aait.06.2023.4.
- [49] A. Giabelli, L. Malandri, F. Mercorio, M. Mezzanzanica, A. Seveso, NEO: A Tool for Taxonomy Enrichment with New Emerging Occupations, in: *The Semantic Web – ISWC 2020*, Springer International Publishing, Cham, 2020, pp. 568–584.
- [50] A. Giabelli, L. Malandri, F. Mercorio, M. Mezzanzanica, JoTA: Aligning Multilingual Job Taxonomies through Word Embeddings (Student Abstract), *Proceedings of the AAAI Conference on Artificial Intelligence* 36 (2022) 12955–12956. URL: <https://ojs.aaai.org/index.php/AAAI/article/view/21614>. doi:10.1609/aaai.v36i11.21614.
- [51] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised Cross-lingual Representation Learning at Scale, in: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Online, 2020, pp. 8440–8451. URL: <https://aclanthology.org/2020.acl-main.747>. doi:10.18653/v1/2020.acl-main.747.
- [52] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. St. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, B. Strope, R. Kurzweil, Universal Sentence Encoder for English, in: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 169–174. URL: <https://aclanthology.org/D18-2029>. doi:10.18653/v1/D18-2029.
- [53] N. Reimers, I. Gurevych, Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, in: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3982–3992. URL: <https://aclanthology.org/D19-1410>. doi:10.18653/v1/D19-1410.

A. Details on the Datasets

We release the source code used to build the datasets⁷, providing researchers with a tool to easily generate new datasets by combining different sets of languages for query and corpus elements. Using this code, new instantiations of the task can be derived from the input data by defining custom language combinations. For example, it is possible to use the Italian national terminology to set up an Italian-to-Greek cross-lingual task, or even combine the query sets of several national classifications and leverage all languages in ESCO to create a more complex multilingual task.

The input data consists of files with the multilingual ESCO Occupations taxonomy (one for each relevant version) and files containing the queries in each national terminology, which are mapped to the ESCO concept ID of the relevant occupation. To create a dataset, the user can select a national terminology and a set of languages for the corpus (any subset of the languages supported by ESCO).

In Table 4 we present example queries and their relevant corpus elements, sampled from the NLD-nl-nl, PRT-pt-pt, and PRT-pt-en datasets.

Finally, we analyze the lexical overlap between the national classifications and ESCO. In Figure 4, we present a histogram showing the normalized edit distance between queries and their closest relevant corpus element, for all the tasks in MELO.

To compute the distances, we first lowercase the surface forms of both the query and the corpus element, and we use the method `ratio` from the Python package `RAPIDFUZZ`⁸. This is a measure of the normalized edit distance between the two strings. In the histograms, for each query, we compute the distance for all its relevant corpus elements and report the minimum distance.

In the histograms, the left-most bin represents the fraction of queries for which the closest relevant element is either identical or very similar. The Danish national terminology has the highest concentration of such cases. To a lesser extent, this is also true for Hungarian, Estonian, and Polish.

Excluding those lexically trivial cases, the more the distribution is skewed to the left, the easier the task. For example, comparing the Belgian (in the French language) and the French tasks, the queries from the French terminology show greater lexical overlap with their relevant corpus elements.

In Appendix C, we use this analysis to compare the performance of lexical baselines across different monolingual tasks.

⁷<https://github.com/Avature/melo-benchmark>

⁸<https://rapidfuzz.github.io/RapidFuzz/Usage/fuzz.html>



Figure 4: Histogram of minimum (normalized) edit distances between each query and the closest relevant corpus element for each monolingual task in MELO.

Table 3

Characteristics of models used in bi-encoder experiments. Encoder-only and decoder-only architectures refer to Transformers. Model size is given in millions of parameters. Some specifications are unknown for the OpenAI model. The *Language Support* column indicates the extent to which the languages involved in the benchmark are supported by each model. † The mUSE-CNN model supports only English, German, French, Spanish, Dutch, Portuguese, Italian, and Polish. ‡ Although GIST-Embedding and E5 are reported to be trained primarily in English, the pre-training of these models did involve examples in other languages as well.

Model	Architecture	Model Size	Output Dims	Language Support
ESCOXLM-R	Encoder-only Transformer	561	1024	Complete
mUSE-CNN	CNN	69	512	Partial †
Paraph-mMPNet	Encoder-only Transformer	278	768	Complete
BGE-M3	Encoder-only Transformer	560	1024	Complete
GIST-Embedding	Encoder-only Transformer	109	768	Mainly English ‡
mE5	Encoder-only Transformer	560	1024	Complete
E5	Decoder-only Transformer	7111	4096	Mainly English ‡
OpenAI	Unknown	Unknown	3072	Unknown

B. Details on the Models

Here, we provide further details about the models explored in this work.

Regarding the lexical baselines, we always apply a simple preprocessing in which we lowercase the input strings and, for all languages except Bulgarian, also perform ASCII normalization. For the edit distance baseline, we use RAPIDFUZZ as described above. For the TF-IDF baselines, we use the SCIKIT-LEARN⁹ Python package, while for the BM5 variants, we use the Okapi BM25 implementation from RANK-BM25¹⁰.

For the baseline variants that involve lemmatization, we use SPACY¹¹ models whenever available. However, SPACY models were not available for the following languages: Bulgarian, Czech, Estonian, Hungarian, Latvian, and Slovak. Lemmatization is applied before ASCII normalization.

In the case of bi-encoders, we experiment with several deep learning sentence encoders that have demonstrated strong performance in other semantic text similarity tasks.

The first model is ESCOXLM-R, proposed by Zhang et al. [10], which is based on XLM-RoBERTa. We use the PyTorch implementation and the pre-trained weights that are available on HuggingFace with the model name `jjzha/esco-xlm-roberta-large`. The base model was pre-trained on data in 88 languages, including all those involved in our datasets, and the fine-tuning by Zhang and colleagues involved learning objectives that leverage information in ESCO. Although it is usual to experiment with the XLM-RoBERTa family of models only after fine-tuning, in our experiment we use it out-of-the-box in a zero-shot setup. During inference, the

input to the model is the surface form of the query or the corpus element, with no preprocessing.

We also present results for the Multilingual Universal Sentence Encoder (mUSE-CNN) model variant with a CNN architecture, proposed by Cer et al. [52, 36]. In our experiments, we use the TensorFlow implementation and the pre-trained weights available on TensorFlow Hub with the handle `google/universal-sentence-encoder-multilingual/3`. This model was pre-trained on data in Arabic, Chinese, English, French, German, Italian, Japanese, Korean, Dutch, Polish, Portuguese, Spanish, Thai, Turkish, and Russian. (Note that, during training, mUSE-CNN has not seen text for languages such as Bulgarian, Czech, or Danish.) During inference, the input to the model is the surface form of the query or the corpus element without any preprocessing or enclosing prompt template.

Other open-source models we experiment with are implemented in PyTorch within the HuggingFace package SENTENCE-TRANSFORMERS [53]. These models are the following: a multilingual model based on MPNet [37] that was pre-trained on 50 languages, including all of MELO languages¹²; the BGE-M3 model [38], which supports more than 100 languages, including also all MELO languages¹³; GIST Embedding [39], which is a model reported to be primarily trained in English¹⁴; Multilingual E5 [40], which was pre-trained on 94 languages, including all of MELO languages¹⁵; and E5 [41, 42] pre-trained on many languages but reported to perform best on English-language input¹⁶.

Finally, we also experiment with the text-embedding-

⁹https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

¹⁰<https://pypi.org/project/rank-bm25/>

¹¹<https://spacy.io/api/lemmatizer>

¹²<https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2>

¹³<https://huggingface.co/BAAI/bge-m3>

¹⁴<https://huggingface.co/avsolatorio/GIST-Embedding-v0>

¹⁵<https://huggingface.co/intfloat/multilingual-e5-large>

¹⁶<https://huggingface.co/intfloat/e5-mistral-7b-instruct>

3-large model from OpenAI¹⁷, which is reported to be state-of-the-art for many semantic text similarity tasks.

For HuggingFace and OpenAI models, during inference, we wrap the input text (the surface form of the query or corpus element) with the following prompt template:

The candidate’s job title is “{{surf_form}}”.
What skills are likely required for this job?

where {{surf_form}} is replaced with the surface form of the element that is being encoded.

This decision was informed by preliminary experiments in which we evaluated various models with different wrapping prompt templates, including no template (as with ESCOXML-R and mUSE-CNN). We speculate that such prompts are particularly beneficial for LLM-based encoders, as they may better capture the semantics of the occupation names we aim to rank.

Although we also experimented with prompts in the same language as each query, this did not improve performance. Consistently using a single prompt ensures a language-agnostic and symmetric bi-encoder approach.

C. Full Results

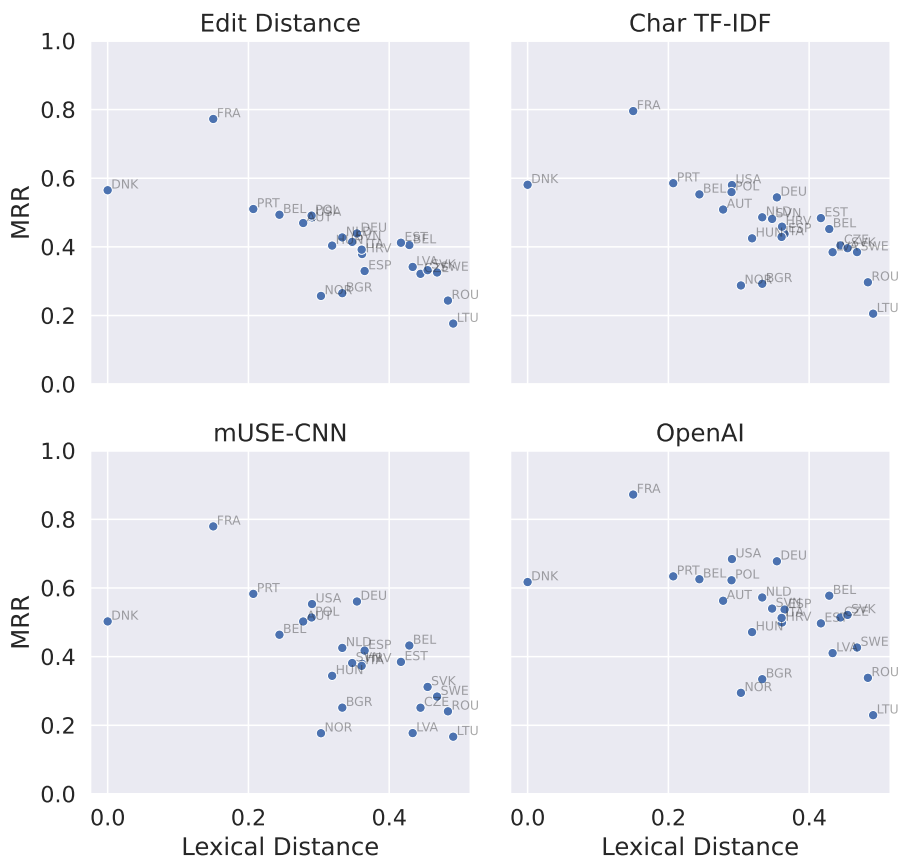
This Section presents the full set of experimental results. Table 5 and Table 6 include the mean reciprocal rank (MRR) for each model across all tasks in MELO.

Although not included with the main results, we also evaluated a random baseline for each dataset, where the score $s(q, c_i)$ for any query and any corpus element is drawn from a uniform distribution. The performance of this baseline varies depending on the number of corpus elements and the distribution of relevant elements per query, but in general, its MRR is close to 0.020.

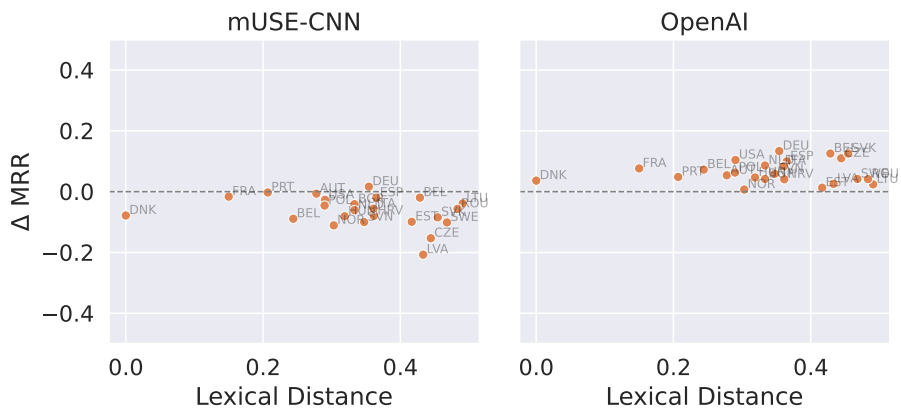
Additionally, Figure 5 shows scatterplots illustrating the correlation between model performance and the median of the lexical overlap index described in Appendix A: the minimum normalized edit distance per query.

Finally, in Figure 6 and in Figure 7 we show the top- k accuracy ($A@k$) for a selection of models in every task in MELO.

¹⁷<https://openai.com/index/new-embedding-models-and-api-updates>



(a) Absolute performance (in MRR).



(b) Performance relative to the lexical baseline Char TF-IDF

Figure 5: Correlation between model performance and the median of the minimum edit distance between queries and relevant corpus elements in monolingual datasets.

Table 4

Example queries and their relevant corpus elements for various datasets in MELO. Surface forms in Dutch are presented in red, while those in Portuguese are in blue.

Dataset Name	Query	Relevant Corpus Elements
NLD-nl-nl	Kredietbeoordelaar	kredietanalist kredietadviseur analist kredieten en risico's
	Woonbegeleider gezinsvervangend huis, wooncentrum	medewerker verzorgingshuis medewerker verzorgingstehuis medewerkster verzorgingscentrum medewerkster verzorgingstehuis medewerkster verzorgingshuis medewerker verzorgingscentrum
	PR-adviseur	lobbyist lobbyiste
PRT-pt-pt	Guia intérprete	Guias-intérpretes
	Fumigador e outros controladores, de pragas e ervas daninhas	Pulverizador de pesticidas/Pulverizadora de pesticidas Pulverizador de pesticidas Pulverizadora de pesticidas
	Empregado de serviços de apoio à produção	Coordenador de montagem de máquinas/Coordenadora de montagem de máquinas Coordenador de montagem de máquinas Coordenadora de montagem de máquinas
PRT-pt-en	Guia intérprete	Travel guides
	Fumigador e outros controladores, de pragas e ervas daninhas	pesticides sprayer lawn care chemical applicator spray technician pesticides applicator trees sprayer sprayer of pesticides
	Empregado de serviços de apoio à produção	machinery assembly coordinator production line coordinator manufacturing co-ordinator assembly line coordinator machinery manufacturing co-ordinator machinery production inspector machinery production co-ordinator assembly line co-ordinator machinery assembly co-ordinator machinery manufacturing manager production line co-ordinator

Table 5

Mean reciprocal rank (MRR) for every model, evaluated in the monolingual and the cross-lingual versions of the MELO tasks.

Model	USA		AUT		BEL		BEL		BGR	
	en-en	en-xx	de-de	de-en	fr-fr	fr-en	nl-nl	nl-en	bg-bg	bg-en
Edit Distance	0.4858	0.4889	0.4695	0.1337	0.4053	0.1072	0.4936	0.1456	0.2651	0.0007
Word TF-IDF	0.3250	0.3207	0.4104	0.0319	0.4589	0.0735	0.4914	0.0618	0.2740	0.0033
Word TF-IDF (lemmas)	0.6056	0.5999	0.4115	0.0288	0.4677	0.0947	0.4907	0.0593	-	-
Char TF-IDF	0.5800	0.5764	0.5088	0.1008	0.4520	0.1827	0.5529	0.1970	0.2925	0.0006
Char TF-IDF (lemmas)	0.5957	0.5913	0.5096	0.1269	0.4597	0.1781	0.5474	0.1750	-	-
BM25	0.2936	0.2814	0.0252	0.0041	0.4130	0.0583	0.4553	0.0398	0.2581	0.0033
BM25 (lemmas)	0.6004	0.5978	0.3808	0.0186	0.4651	0.0723	0.4598	0.0389	-	-
ESCOXLM-R	0.3450	0.3426	0.4150	0.0767	0.2575	0.0537	0.3720	0.1084	0.2215	0.0269
mUSE-CNN	0.5532	0.5317	0.5024	0.2656	0.4324	0.3213	0.4638	0.3148	0.2514	0.1044
Paraph-mMPNet	0.5876	0.5822	0.3726	0.0852	0.3824	0.1459	0.4283	0.1498	0.2146	0.0167
BGE-M3	0.6226	0.6301	0.5330	0.2819	0.5225	0.4005	0.5709	0.3529	0.3192	0.1825
GIST-Embedding	0.6431	0.6464	0.4819	0.0947	0.4803	0.1848	0.5113	0.1706	0.2700	0.0033
mE5	0.6563	0.6588	0.5334	0.3092	0.5407	0.4266	0.5683	0.3851	0.3106	0.1870
E5	0.6735	0.6777	0.5612	0.4143	0.5606	0.5380	0.6133	0.4991	0.3406	0.2371
OpenAI	0.6842	0.6872	0.5628	0.4304	0.5775	0.5736	0.6255	0.5698	0.3343	0.2367

(a) Results for tasks: O*NET, Austria, Belgium (French), Belgium (Dutch), and Bulgaria.

Model	CZE		DEU		DNK		ESP		EST	
	cs-cs	cs-en	de-de	de-en	da-da	da-en	es-es	es-en	et-et	et-en
Edit Distance	0.3215	0.0524	0.4392	0.0832	0.5650	0.1596	0.3297	0.0545	0.4121	0.1146
Word TF-IDF	0.2410	0.0023	0.4763	0.0388	0.5187	0.0398	0.2411	0.0127	0.3675	0.0097
Word TF-IDF (lemmas)	-	-	0.4666	0.0391	0.5179	0.0404	0.4318	0.0307	-	-
Char TF-IDF	0.4043	0.0843	0.5442	0.1301	0.5809	0.1576	0.4376	0.1238	0.4838	0.1095
Char TF-IDF (lemmas)	-	-	0.5474	0.1278	0.5801	0.1551	0.4697	0.1347	-	-
BM25	0.2189	0.0023	0.3377	0.0050	0.4987	0.0296	0.1916	0.0073	0.2982	0.0055
BM25 (lemmas)	-	-	0.4473	0.0198	0.5125	0.0334	0.4367	0.0275	-	-
ESCOXLM-R	0.1835	0.0195	0.4087	0.1002	0.3631	0.1095	0.2476	0.0854	0.2995	0.0374
mUSE-CNN	0.2512	0.0914	0.5606	0.3138	0.5026	0.1680	0.4176	0.3217	0.3847	0.0811
Paraph-mMPNet	0.2418	0.0464	0.4691	0.0916	0.4602	0.1148	0.3417	0.0899	0.3505	0.0635
BGE-M3	0.4285	0.3021	0.6083	0.3344	0.5839	0.3037	0.4927	0.3084	0.4726	0.2882
GIST-Embedding	0.3383	0.0854	0.5363	0.1325	0.5608	0.1348	0.3574	0.1534	0.3996	0.0597
mE5	0.4498	0.3406	0.6122	0.3858	0.5983	0.3325	0.5021	0.3480	0.4531	0.2757
E5	0.5145	0.4148	0.6639	0.5073	0.6178	0.4053	0.5557	0.4628	0.4913	0.2465
OpenAI	0.5141	0.4356	0.6778	0.5518	0.6173	0.4506	0.5371	0.4859	0.4969	0.3915

(b) Results for tasks: Czechia, Germany, Denmark, Spain, and Estonia.

Model	FRA		HRV		HUN		ITA		LTU	
	fr-fr	fr-en	hr-hr	hr-en	hu-hu	hu-en	it-it	it-en	lt-lt	lt-en
Edit Distance	0.7726	0.0964	0.3791	0.0325	0.4037	0.0362	0.3919	0.1069	0.1766	0.0530
Word TF-IDF	0.7743	0.0646	0.4565	0.0058	0.3604	0.0035	0.1886	0.0164	0.1890	0.0033
Word TF-IDF (lemmas)	0.7824	0.0810	0.4416	0.0073	-	-	0.4452	0.0142	0.1973	0.0036
Char TF-IDF	0.7956	0.1954	0.4588	0.0995	0.4249	0.0273	0.4290	0.1560	0.2054	0.0410
Char TF-IDF (lemmas)	0.7936	0.1890	0.4657	0.0936	-	-	0.4800	0.1760	0.2118	0.0361
BM25	0.7514	0.0484	0.4050	0.0021	0.3247	0.0030	0.1609	0.0036	0.1874	0.0033
BM25 (lemmas)	0.8042	0.0707	0.4445	0.0075	-	-	0.4425	0.0081	0.1984	0.0037
ESCOXLM-R	0.6603	0.1098	0.3128	0.0479	0.2952	0.0311	0.2686	0.1124	0.1381	0.0209
mUSE-CNN	0.7794	0.3681	0.3790	0.0769	0.3441	0.0324	0.3732	0.2861	0.1668	0.0521
Paraph-mMPNet	0.7660	0.1624	0.3678	0.0524	0.3198	0.0219	0.3580	0.0902	0.1747	0.0196
BGE-M3	0.8454	0.4171	0.4827	0.2473	0.4496	0.1878	0.4730	0.3271	0.2212	0.1310
GIST-Embedding	0.8047	0.2011	0.3968	0.0737	0.3901	0.0306	0.4242	0.1651	0.1876	0.0355
mE5	0.8427	0.4464	0.4734	0.2712	0.4327	0.2155	0.4825	0.3583	0.2258	0.1206
E5	0.8632	0.5760	0.5074	0.3516	0.4973	0.3372	0.5384	0.4459	0.2269	0.1121
OpenAI	0.8721	0.6160	0.4995	0.3795	0.4715	0.3455	0.5128	0.4573	0.2295	0.1754

(c) Results for tasks: France, Croatia, Hungary, Italy, and Lithuania.

Table 6

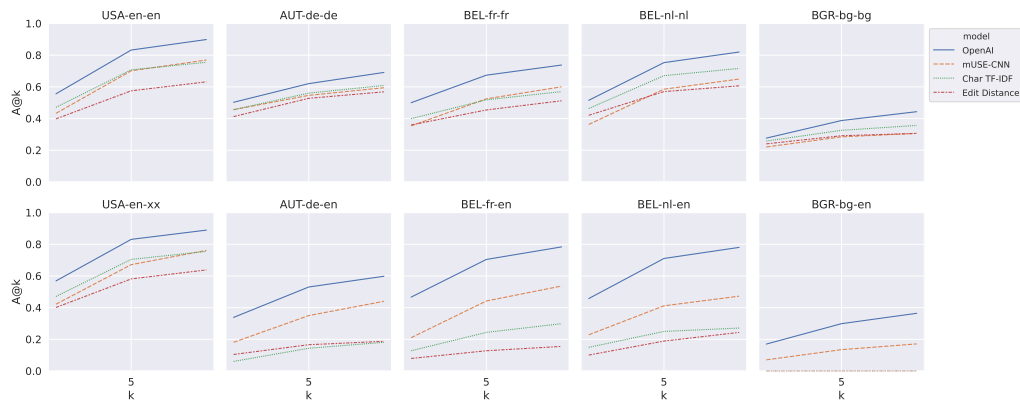
Mean reciprocal rank (MRR) for every model, evaluated in the monolingual and the cross-lingual versions of the MELO tasks.

Model	LVA		NLD		NOR		POL		PRT	
	lv-lv	lv-en	nl-nl	nl-en	no-no	no-en	pl-pl	pl-en	pt-pt	pt-en
Edit Distance	0.3416	0.0900	0.4275	0.0952	0.2571	0.0472	0.4911	0.0637	0.5103	0.1119
Word TF-IDF	0.3802	0.0066	0.4714	0.0460	0.0453	0.0008	0.5630	0.0143	0.6051	0.0272
Word TF-IDF (lemmas)	-	-	0.4674	0.0435	0.1292	0.0009	0.5588	0.0216	0.5947	0.0266
Char TF-IDF	0.3845	0.0774	0.4862	0.1281	0.2876	0.0582	0.5596	0.1109	0.5855	0.1896
Char TF-IDF (lemmas)	-	-	0.4811	0.1321	0.3272	0.0472	0.5528	0.1115	0.5860	0.1904
BM25	0.3664	0.0054	0.4433	0.0338	0.0316	0.0002	0.5535	0.0085	0.5736	0.0236
BM25 (lemmas)	-	-	0.4320	0.0393	0.1307	0.0004	0.5482	0.0181	0.5886	0.0258
ESCOXLM-R	0.1569	0.0276	0.3184	0.0829	0.1101	0.0267	0.4063	0.0882	0.4846	0.1312
mUSE-CNN	0.1773	0.0357	0.4255	0.2666	0.1769	0.1109	0.5141	0.3258	0.5829	0.3363
Paraph-mMPNet	0.2718	0.0343	0.3831	0.0955	0.1485	0.0492	0.4582	0.0585	0.5362	0.0996
BGE-M3	0.3842	0.1922	0.5045	0.3033	0.2662	0.1984	0.5916	0.3542	0.5878	0.4124
GIST-Embedding	0.3450	0.0603	0.4487	0.1316	0.2427	0.0716	0.4859	0.1107	0.5330	0.2017
mE5	0.3839	0.1899	0.5059	0.3246	0.2558	0.2229	0.5836	0.3793	0.5834	0.4372
E5	0.4008	0.1716	0.5650	0.4133	0.2899	0.3512	0.6220	0.4844	0.6416	0.5336
OpenAI	0.4103	0.2418	0.5723	0.4509	0.2946	0.4358	0.6225	0.5085	0.6339	0.5413

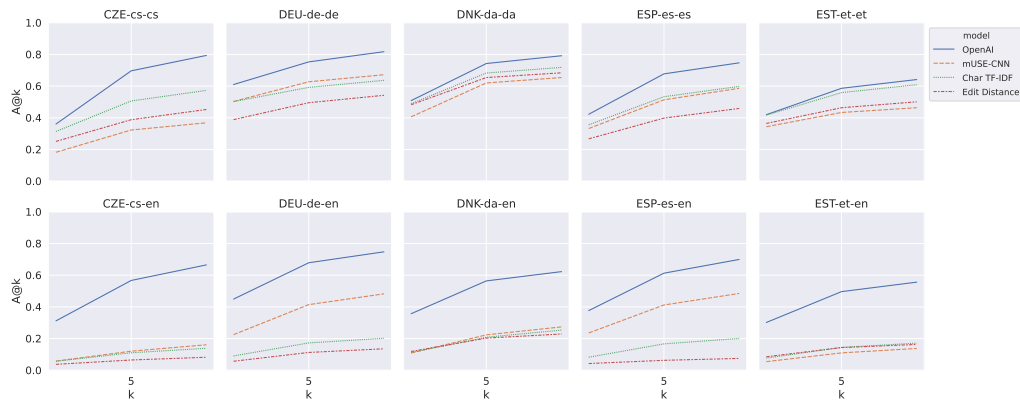
(a) Results for tasks: Latvia, the Netherlands, Norway, Poland, and Portugal.

Model	ROU		SVK		SVN		SWE	
	ro-ro	ro-ro	sk-sk	sk-en	sl-sl	sl-en	sv-it	sv-en
Edit Distance	0.2436	0.0521	0.3321	0.0725	0.4145	0.0665	0.3254	0.0845
Word TF-IDF	0.2849	0.0261	0.3695	0.0156	0.4808	0.0083	0.2997	0.0187
Word TF-IDF (lemmas)	0.2768	0.0332	-	-	0.4821	0.0133	0.3034	0.0191
Char TF-IDF	0.2969	0.1043	0.3961	0.1123	0.4814	0.0759	0.3848	0.0905
Char TF-IDF (lemmas)	0.3038	0.1054	-	-	0.4850	0.0757	0.3904	0.0937
BM25	0.2687	0.0224	0.3477	0.0120	0.4645	0.0051	0.2421	0.0125
BM25 (lemmas)	0.2621	0.0300	-	-	0.4862	0.0131	0.3002	0.0170
ESCOXLM-R	0.1458	0.0556	0.2295	0.0429	0.3179	0.0535	0.2111	0.0681
mUSE-CNN	0.2407	0.1171	0.3118	0.1040	0.3814	0.0836	0.2837	0.0948
Paraph-mMPNet	0.2649	0.1036	0.2799	0.0639	0.3593	0.0486	0.2662	0.0611
BGE-M3	0.3167	0.1946	0.4568	0.3127	0.4999	0.2848	0.3905	0.1905
GIST-Embedding	0.2852	0.1084	0.3491	0.1041	0.4223	0.0765	0.3265	0.0772
mE5	0.3176	0.2043	0.4632	0.3308	0.4897	0.2815	0.4001	0.2017
E5	0.3314	0.2308	0.5087	0.3604	0.5339	0.3912	0.4286	0.2605
OpenAI	0.3383	0.2572	0.5216	0.4122	0.5400	0.4092	0.4266	0.2909

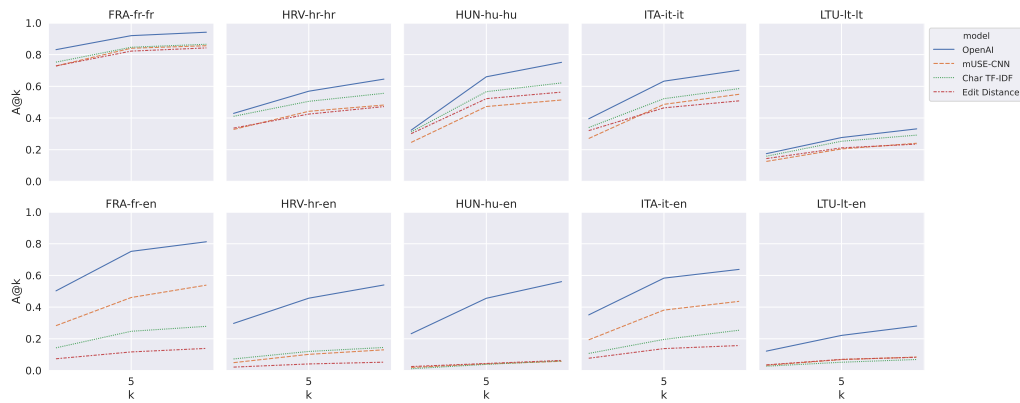
(b) Results for tasks: Romania, Slovakia, Slovenia, and Sweden.



(a) Results for tasks: O*NET, Austria, Belgium (fr), Belgium (nl), and Bulgaria.

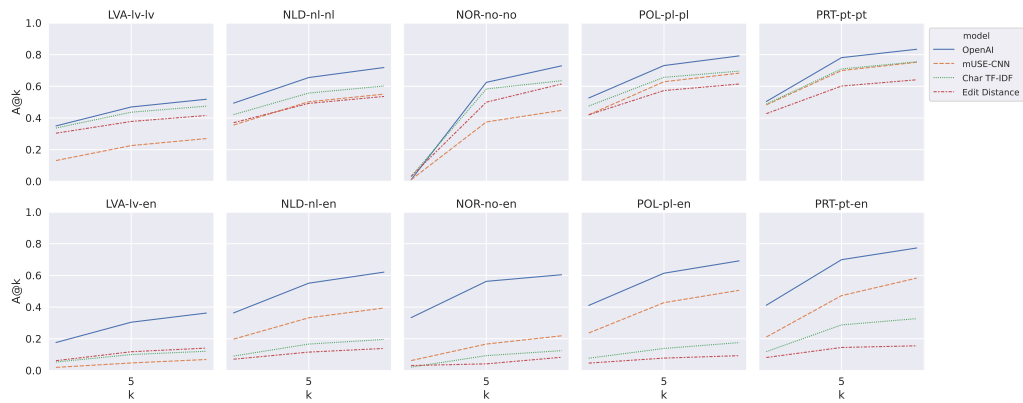


(b) Results for tasks: Czechia, Germany, Denmark, Spain, and Estonia.

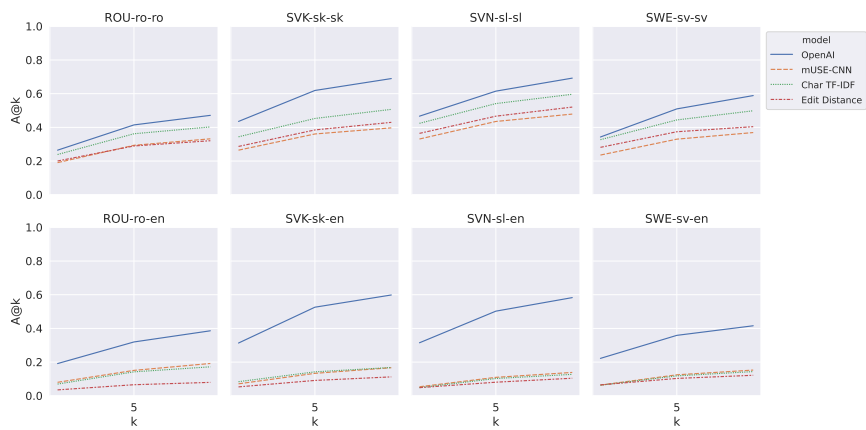


(c) Results for tasks: France, Croatia, Hungary, Italy, and Lithuania.

Figure 6: Top- k accuracy ($A@k$) for a selection of models in the MELO Benchmark tasks.



(a) Results for tasks: Latvia, the Netherlands, Norway, Poland, and Portugal.



(b) Results for tasks: Romania, Slovakia, Slovenia, and Sweden.

Figure 7: Top- k accuracy ($A@k$) for a selection of models in the MELO Benchmark tasks.