Learning Domain-Specialised Representations for Cross-Lingual Biomedical Entity Linking

Fangyu Liu, Ivan Vulić, Anna Korhonen, Nigel Collier
Language Technology Lab, TAL, University of Cambridge
{fl399, iv250, alk23, nhc30}cam.ac.uk

Abstract

Injecting external domain-specific knowledge (e.g., UMLS) into pretrained language models (LMs) advances their capability to handle specialised in-domain tasks such as biomedical entity linking (BEL). However, such abundant expert knowledge is available only for a handful of languages (e.g., English). In this work, by proposing a novel cross-lingual biomedical entity linking task (XL-BEL) and establishing a new XL-BEL benchmark spanning 10 typologically diverse languages, we first investigate the ability of standard knowledge-agnostic as well as knowledge-enhanced monolingual and multilingual LMs beyond the standard monolingual English BEL task. The scores indicate large gaps to English performance. We then address the challenge of transferring domain-specific knowledge from resource-rich languages to resource-poor ones. To this end, we propose and evaluate a series of cross-lingual transfer methods for the XL-BEL task, and demonstrate that general-domain bitext helps propagate the available English knowledge to languages with little to no in-domain data. Remarkably, we show that our proposed domain-specific transfer methods yield consistent gains across all target languages, sometimes up to 20 Precision@1 points, without any in-domain knowledge in the target language, and without any in-domain parallel data.

1 Introduction

Recent work has demonstrated that it is possible to combine the strength of 1) Transformer-based encoders such as BERT (Devlin et al., 2019; Liu et al., 2019), pretrained on large general-domain data with 2) external linguistic and world knowledge (Zhang et al., 2019; Levine et al., 2020; Lauscher et al., 2020). Such expert human-curated knowledge is crucial for NLP applications in specialised domains such as biomedicine. There, Liu et al. (2021) recently proposed *self-alignment pretraining* (SAP), a technique to fine-tune BERT on phrase-level synonyms extracted from the Unified Medical Language System (UMLS; Bodenreider 2004).[1] Their SAPBERT model currently holds state-of-the-art (SotA) across all major English biomedical entity linking (BEL) datasets. However, this approach is not widely applicable to other languages: abundant external resources are available only for a few languages, hindering the development of domain-specific NLP models in all other languages.

Simultaneously, exciting breakthroughs in cross-lingual transfer for language understanding tasks have been achieved (Artetxe and Schwenk, 2019; Hu et al., 2020). However, it remains unclear whether such transfer techniques can be used to improve domain-specific NLP applications and mitigate the gap between knowledge-enhanced models in resource-rich versus resource-poor languages. In this paper, we thus investigate the current performance gaps in the BEL task beyond English, and propose several cross-lingual transfer techniques to improve domain-specialised representations and BEL in resource-lean languages.

In particular, we first present a novel cross-lingual BEL (XL-BEL) task and its corresponding evaluation benchmark in 10 typologically diverse languages, which aims to map biomedical names/mentions in any language to the controlled UMLS vocabulary. After empirically highlighting the deficiencies of multilingual encoders (e.g, MBERT and XLMR; Conneau et al. 2020) on XL-BEL, we propose and evaluate a multilingual extension of the SAP technique. Our main results suggest that expert knowledge can be transferred from English to resource-leaner languages, yielding huge gains over vanilla MBERT and XLMR, and English-only SAPBERT. We also show that

[1] UMLS is a large-scale biomedical knowledge graph containing more than 14M biomedical entity names.

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Short Papers), pages 565–574
August 1–6, 2021. ©2021 Association for Computational Linguistics

leveraging general-domain word and phrase translations offers substantial gains in the XL-BEL task.

Contributions. 1) We highlight the challenge of learning (biomedical) domain-specialised cross-lingual representations. **2)** We propose a novel multilingual XL-BEL task with a comprehensive evaluation benchmark in 10 languages. **3)** We offer systematic evaluations of existing knowledge-agnostic and knowledge-enhanced monolingual and multilingual LMs in the XL-BEL task. **4)** We present a new SotA multilingual encoder in the biomedical domain, which yields large gains in XL-BEL especially on resource-poor languages, and provides strong benchmarking results to guide future work. The code, data, and pretrained models are available online at: `github.com/cambridgeltl/sapbert`.

2 Methodology

Background and Related Work. Learning biomedical entity representations is at the core of BioNLP, benefiting, e.g., relational knowledge discovery (Wang et al., 2018) and literature search (Lee et al., 2016). In the current era of contextualised representations based on Transformer architectures (Vaswani et al., 2017), biomedical text encoders are pretrained via Masked Language Modelling (MLM) on diverse biomedical texts such as PubMed articles (Lee et al., 2020; Gu et al., 2020), clinical notes (Peng et al., 2019; Alsentzer et al., 2019), and even online health forum posts (Basaldella et al., 2020). However, it has been empirically verified that naively applying MLM-pretrained models as entity encoders does not perform well in tasks such as biomedical entity linking (Basaldella et al., 2020; Sung et al., 2020). Recently, Liu et al. (2021) proposed SAP (**S**elf-**A**lignment **P**retraning), a fine-tuning method that leverages synonymy sets extracted from UMLS to improve BERT's ability to act as a biomedical entity encoder. Their SAPBERT model currently achieves SotA scores on all major English BEL benchmarks.

In what follows, we first outline the SAP procedure, and then discuss the extension of the method to include multilingual UMLS synonyms (§2.1), and then introduce another SAP extension which combines domain-specific synonyms with general-domain translation data (§2.2).

2.1 Language-Agnostic SAP

Let $(x, y) \in \mathcal{X} \times \mathcal{Y}$ denote the tuple of a name and its categorical label. When learning from UMLS synonyms, $\mathcal{X} \times \mathcal{Y}$ is the set of all *(name, CUI*[2]*)* pairs, e.g., (*vaccination*, `C0042196`). While Liu et al. (2021) use only English names, we here consider names in other UMLS languages. During training, the model is steered to create similar representations for synonyms regardless of their language.[3] The learning scheme includes 1) an online sampling procedure to select training examples and 2) a metric learning loss that encourages strings sharing the same CUI to obtain similar representations.

Training Examples. Given a mini-batch of N examples $\mathcal{B} = \mathcal{X_B} \times \mathcal{Y_B} = \{(x_i, y_i)\}_{i=1}^{N}$, we start from constructing all possible triplets for all names $x_i \in \mathcal{X_B}$. Each triplet is in the form of (x_a, x_p, x_n) where x_a is the *anchor*, an arbitrary name from $\mathcal{X_B}$; x_p is a positive match of x_a (i.e., $y_a = y_p$) and x_n is a negative match of x_a (i.e., $y_a \neq y_n$). Let $f(\cdot)$ denote the encoder (i.e., MBERT or XLMR in this paper). Among the constructed triplets, we select all triplets that satisfy the following constraint:

$$\|f(x_a) - f(x_p)\|_2 + \lambda \geq \|f(x_a) - f(x_n)\|_2,$$

where λ is a predefined margin. In other words, we only consider triplets with the positive sample further to the negative sample by a margin of λ. These 'hard' triplets are more informative for representation learning (Liu et al., 2021). Every selected triplet then contributes one positive pair (x_a, x_p) and one negative pair (x_a, x_n). We collect all such positives and negatives, and denote them as $\mathcal{P}, \mathcal{N}$.

Multi-Similarity Loss. We compute the pairwise cosine similarity of all the name representations and obtain a similarity matrix $\mathbf{S} \in \mathbb{R}^{|\mathcal{X_B}| \times |\mathcal{X_B}|}$ where each entry $\mathbf{S}_{ij}$ is the cosine similarity between the i-th and j-th names in the mini-batch $\mathcal{B}$. The Multi-Similarity loss (MS, Wang et al. 2019), is then used for learning from the triplets:

$$\mathcal{L} = \frac{1}{|\mathcal{X_B}|} \sum_{i=1}^{|\mathcal{X_B}|} \left(\frac{1}{\alpha} \log \left(1 + \sum_{n \in \mathcal{N}_i} e^{\alpha(\mathbf{S}_{in} - \epsilon)}\right) + \frac{1}{\beta} \log \left(1 + \sum_{p \in \mathcal{P}_i} e^{-\beta(\mathbf{S}_{ip} - \epsilon)}\right)\right). \tag{1}$$

α, β are temperature scales; ϵ is an offset applied on the similarity matrix; $\mathcal{P}_i, \mathcal{N}_i$ are indices of positive and negative samples of the i-th *anchor*.

[2] In UMLS, "CUI" means **C**oncept **U**nique **I**dentifier.

[3] For instance, *vaccination* (EN), *active immunization* (EN), *vacunación* (ES) and 予防接種 (JA) all share the same Concept Unique Identifier (CUI; `C0042196`); thus, they should all have similar representations.

#↓, language→	EN	ES	DE	FI	RU	TR	KO	ZH	JA	TH
sentences	-	223,506	350,193	77,736	206,060	29,473	47,702	136,054	157,670	19,066
unique titles (Wiki page)	60,598	37,935	24,059	15,182	21,044	5,251	10,618	17,972	11,002	4,541
mentions	1,067,083	204,253	431,781	105,182	221,383	29,958	60,979	197,317	220,452	31,177
unique mentions	121,669	25,169	44,390	26,184	28,302	4,110	9,032	24,825	21,949	5,064
unique mentions$_{\text{mention!=title}}$	69,199	22,162	43,753	19,409	23,935	2,833	3,740	12,046	12,571	2,480

Table 1: Construction of the XL-BEL benchmark; key statistics. See the App. §A.1 for further details.

2.2 SAP with General-Domain Bitext

We also convert word and phrase translations into the same format (§2.1), where each 'class' now contains only two examples. For a translation pair (x_p, x_q), we create a unique pseudo-label y_{x_p,x_q} and produce two new name-label instances (x_p, y_{x_p,x_q}) and (x_q, y_{x_p,x_q}),[4] and proceed as in §2.1. This allows us to easily combine domain-specific knowledge with general translation knowledge within the same SAP framework.

3 The XL-BEL Task and Evaluation Data

A general cross-lingual entity linking (EL) task (McNamee et al., 2011; Tsai and Roth, 2016) aims to map a mention of an entity in free text of *any language* to a controlled English vocabulary, typically obtained from a knowledge graph (KG). In this work, we propose XL-BEL, a cross-lingual *biomedical* EL task. Instead of grounding entity mentions to English-specific ontologies, we use UMLS as a language-agnostic KG: the XL-BEL task requires a model to associate a mention in any language to a (language-agnostic) CUI in UMLS. XL-BEL thus serves as an ideal evaluation benchmark for biomedical entity representations: it challenges the capability of both 1) representing domain entities and also 2) associating entity names in different languages.

Evaluation Data Creation. For English, we take the available BEL dataset WikiMed (Vashishth et al., 2020), which links Wikipedia mentions to UMLS CUIs. We then follow similar procedures as WikiMed and create an XL-BEL benchmark covering 10 languages (see Table 2). For each language, we extract all sentences from its Wikipedia dump, find all hyperlinked concepts (i.e., words and phrases), lookup their Wikipedia pages, and retain only concepts that are linked to UMLS.[5] For each UMLS-linked mention, we add a triplet *(sentence, mention, CUI)* to our dataset.[6] Only one example per surface form is retained to ensure diversity. We then filter out examples with mentions that have the same surface form as their Wikipedia article page.[7] Finally, 1k examples are randomly selected for each language: they serve as the final test sets in our XL-BEL benchmark. The statistics of the benchmark are available in Table 1.

4 Experiments and Results

UMLS Data. We rely on the UMLS (2020AA) as our SAP fine-tuning data, leveraging synonyms in all available languages. The full multilingual fine-tuning data comprises ≈15M biomedical entity names associated with ≈4.2M individual CUIs. As expected, English is dominant (69.6% of all 15M names), followed by Spanish (10.7%) and French (2.2%). The full stats are in App. §A.3.

Translation Data. We use (a) "muse" word translations (Lample et al., 2018), and (b) the parallel Wikipedia article titles (phrase-level translations; referred to as "wt"). We also list results when using "muse" and "wt" combined ("wt+ muse").

Training and Evaluation Details. Our SAP fine-tuning largely follows Liu et al. (2021); we refer to the original work and the Appendix for further tech-

[4] These pseudo-labels are not related to UMLS, but are used to format our parallel translation data into the input convenient for the SAP procedure. In practice, for these data we generate pseudo-labels ourselves as 'LANGUAGE_CODE+index'. For instance, `ENDE2344` indicates that this word pair is our 2,344th English-German word translation. Note that the actual coding scheme does not matter as it is only used for our algorithm to determine what terms belong to the same (in this case - translation) category.

[5] For instance, given a sentence from German Wikipedia *Die [Inkubationszeit] von COVID-19 beträgt durchschnittlich fünf bis sechs Tage.*, we extract the hyperlinked word *Inkubationszeit* as an UMLS-linked entity mention. Since Wikipedia is inherently multilingual, if *Inkubationszeit* is linked to UMLS, its cross-lingual counterparts, e.g., *Incubation period* (EN), are all transitively linked to UMLS.

[6] Note that though each mention is accompanied with its context, we regard it as out-of-context mention following the tradition in prior work (Sung et al., 2020; Liu et al., 2021; Tutubalina et al., 2020). According to Basaldella et al. (2020), biomedical entity representations can be easily polluted by its context. We leave contextual modelling for future work.

[7] Otherwise, the problem is easily solved by comparing surface forms of the mention and the article title.

language→	EN		ES		DE		FI		RU		TR		KO		ZH		JA		TH		**avg**	
model↓	@1	@5	@1	@5	@1	@5	@1	@5	@1	@5	@1	@5	@1	@5	@1	@5	@1	@5	@1	@5	@1	@5
monolingual models																						
{$LANG}BERT	-	-	41.3	42.5	16.8	18.4	4.9	5.2	1.1	1.6	19.5	21.8	1.1	1.6	2.1	3.2	2.7	2.8	0.4	0.4	10.0	10.8
+ SAP$_{\text{all\_syn}}$	-	-	60.9	66.8	**35.5**	**40.0**	**18.8**	**23.9**	**36.4**	**42.4**	**44.9**	**49.7**	13.5	16.0	18.5	**23.8**	21.2	25.9	0.6	0.6	27.8	32.1
SAPBERT	**78.7**	**81.6**	47.3	51.4	22.7	24.7	8.2	10.2	5.8	6.0	26.4	29.7	2.0	2.4	1.9	2.2	3.0	3.2	3.1	3.4	19.9	21.6
SAPBERT$_{\text{all\_syn}}$	78.3	80.7	55.6	61.3	30.0	34.2	11.8	14.8	9.3	11.3	35.5	39.5	2.0	2.4	6.4	8.2	6.9	8.3	3.0	3.3	23.9	26.4
multilingual models																						
MBERT	0.8	1.7	0.5	0.7	0.3	0.4	0.4	0.8	0.0	0.0	0.7	1.2	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.3	0.5
+ SAP$_{\text{en\_syn}}$	75.5	79.9	50.6	55.8	26.0	29.6	8.7	10.7	10.1	12.6	31.0	34.4	2.7	3.2	4.1	5.7	4.7	5.9	3.1	3.5	21.7	24.1
+ SAP$_{\text{all\_syn}}$	75.0	79.7	**61.4**	**67.0**	33.4	37.8	18.4	21.9	35.1	40.3	44.5	47.7	15.1	17.6	**19.5**	22.7	19.9	25.0	2.8	3.4	32.5	36.3
XLMR	1.0	2.0	0.3	0.7	0.0	0.1	0.1	0.2	0.1	0.2	0.4	0.5	0.0	0.3	0.1	0.2	0.2	0.4	0.0	0.1	0.2	0.5
+ SAP$_{\text{en\_syn}}$	78.1	80.9	47.9	53.5	27.6	32.0	12.2	14.7	21.8	25.9	29.3	35.9	4.5	6.7	7.9	11.3	8.3	11.3	11.5	16.2	24.9	28.8
+ SAP$_{\text{all\_syn}}$	78.2	81.0	56.4	62.7	31.8	37.3	18.6	22.2	35.4	41.2	42.8	48.9	**16.7**	**21.4**	18.8	23.0	**24.0**	**28.1**	**20.6**	**27.5**	**34.3**	**39.3**

Table 2: Various base models combined with SAP, using either all synonyms (*all_syn*) or only English synonyms (*en_syn*) in UMLS. {$LANG} denotes the language of the corresponding column (also in Table 4). See Table 6 (App. §A.3) for the language codes. **avg** refers to the average performance across all target languages. Grey and light blue rows are off-the-shelf base models and models fine-tuned with the UMLS knowledge, respectively.

language→	ES		DE		FI		RU		TR		KO		ZH		JA		TH		**avg**	
model↓	@1	@5	@1	@5	@1	@5	@1	@5	@1	@5	@1	@5	@1	@5	@1	@5	@1	@5	@1	@5
XLMR + SAP$_{\text{en\_syn}}$	47.9	53.5	27.6	32.0	12.2	14.7	21.8	25.9	29.3	35.9	4.5	6.7	7.9	11.3	8.3	11.3	11.5	16.2	19.0	23.1
+ en-{$LANG} wt	55.0	62.2	34.6	41.4	18.6	24.4	35.0	41.5	43.3	50.6	15.9	22.3	15.9	23.0	18.7	24.4	25.1	32.4	29.1	35.8
+ en-{$LANG} muse	54.4	61.0	28.7	34.4	16.7	20.6	33.6	39.0	41.9	48.8	11.9	16.3	12.3	16.7	15.7	19.9	18.6	25.1	26.0	31.3
+ en-{$LANG} wt+muse	49.4	59.6	30.3	36.9	20.4	28.9	33.2	41.9	42.7	51.7	16.1	22.3	16.0	22.9	17.8	24.3	26.2	34.0	28.0	35.8
XLMR + SAP$_{\text{all\_syn}}$	56.4	62.7	31.8	37.3	18.6	22.2	35.4	41.2	42.8	48.9	16.7	21.4	18.8	23.0	24.0	28.1	20.6	27.5	29.5	34.7
+ en-{$LANG} wt	57.2	63.7	35.1	42.3	20.3	27.6	35.8	43.8	48.8	55.0	22.1	27.9	20.6	27.3	24.8	**31.3**	30.0	37.6	32.7	**39.6**
+ en-{$LANG} muse	57.9	63.9	33.0	38.4	23.0	27.3	39.8	45.9	47.2	54.5	22.1	25.7	19.2	25.6	**25.2**	30.2	25.9	32.8	32.6	38.3
+ en-{$LANG} wt+ muse	51.4	61.2	31.3	38.9	22.8	28.4	36.4	45.2	42.2	51.6	**24.4**	**29.2**	21.1	28.2	23.2	30.4	**30.9**	**37.9**	31.5	39.0
MBERT + SAP$_{\text{all\_syn}}$	**61.4**	67.0	33.4	37.8	18.4	21.9	35.1	40.3	44.5	47.7	15.1	17.6	19.5	22.7	19.9	25.0	2.8	3.4	27.8	31.5
+ en-{$LANG} wt	59.2	66.9	**37.5**	**43.9**	25.6	33.0	39.6	47.2	52.7	59.7	19.8	24.3	24.1	31.9	23.5	28.7	4.8	5.9	31.9	37.9
+ en-{$LANG} muse	59.9	66.2	34.3	38.8	21.6	27.5	36.5	41.7	51.0	56.7	18.1	21.2	22.2	26.4	22.0	25.5	3.4	3.8	29.2	34.2
+ en-{$LANG} wt+ muse	59.2	**67.5**	35.3	42.4	**30.5**	**37.3**	**41.6**	**49.2**	**57.2**	**64.7**	19.8	25.0	**24.6**	**32.1**	24.3	28.0	5.2	6.3	**33.1**	39.2

Table 3: Results when applying SAP with 1) UMLS knowledge + 2) word and/or phrase translations .

nical details. The evaluation measure is standard Precision@1 and Precision@5. In all experiments, SAP always denotes fine-tuning of a base LM with UMLS data. `[CLS]` of the last layer's output is used as the final representation (Liu et al., 2021). Without explicit mentioning, we use the BASE variants of all monolingual and multilingual LMs. At inference, given a query representation, a nearest neighbour search is used to rank all candidates' representations. We restrict the target ontology to only include CUIs that appear in WikiMed (62,531 CUIs, 399,931 entity names).

4.1 Main Results and Discussion

Multilingual UMLS Knowledge Always Helps (Table 2). Table 2 summarises the results of applying multilingual SAP fine-tuning based on UMLS knowledge on a wide variety of monolingual, multilingual, and in-domain pretrained encoders. Injecting UMLS knowledge is consistently beneficial to the models' performance on XL-BEL across all languages and across all base encoders. Using multilingual UMLS synonyms to SAP-fine-tune the biomedical PUBMEDBERT (SAPBERT$_{\text{all\_syn}}$) instead of English-only synonyms (SAPBERT) improves its performance across the board. SAP-ing monolingual BERTs for each language also yields substantial gains across all languages; the only exception is Thai (TH), which is not represented in UMLS. Fine-tuning multilingual models MBERT and XLMR leads to even larger relative gains.

Performance across Languages (Table 2). UMLS data is heavily biased towards Romance and Germanic languages. As a result, for languages more similar to these families, monolingual LMs (upper half, Table 2) are on par or outperform multilingual LMs (lower half, Table 2). However, for other (distant) languages (e.g., KO, ZH, JA, TH), the opposite holds. For instance, on TH, XLMR+SAP$_{\text{all\_syn}}$ outperforms THBERT+SAP$_{\text{all\_syn}}$ by 20% Precision@1.

General Translation Knowledge is Useful (Table 3). Table 3 summarises the results where we

language→	ES		DE		RU		KO		**avg**	
model↓	@1	@5	@1	@5	@1	@5	@1	@5	@1	@5
mBERT										
+ SAP$_{\text{en\_syn}}$	50.6	55.8	26.0	29.6	10.1	12.6	2.7	3.2	22.4	25.3
+ SAP$_{\{\$\text{LANG}\}\text{\_syn}}$	57.1	62.8	28.9	33.6	25.8	31.7	2.1	2.6	28.5	32.7
+ SAP$_{\text{en+}\{\$\text{LANG}\}\text{\_syn}}$	61.1	**68.5**	**35.2**	**39.8**	**35.6**	**40.9**	14.4	16.3	**36.6**	**41.4**
+ SAP$_{\text{all\_syn}}$	**61.4**	67.0	33.4	37.8	35.1	40.3	**15.1**	**17.6**	36.6	40.7
XLMR										
+ SAP$_{\text{en\_syn}}$	47.9	53.5	27.6	32.0	21.8	25.9	4.5	6.7	25.5	29.5
+ SAP$_{\{\$\text{LANG}\}\text{\_syn}}$	52.9	55.8	25.9	30.4	28.7	34.2	2.4	2.9	24.5	30.8
+ SAP$_{\text{en+}\{\$\text{LANG}\}\text{\_syn}}$	55.8	62.5	27.7	32.3	**36.4**	**42.2**	15.8	19.8	33.9	39.2
+ SAP$_{\text{all\_syn}}$	**56.4**	**62.7**	**31.8**	**37.3**	35.4	41.2	**16.7**	**21.4**	**35.1**	**40.7**

Table 4: Varying UMLS synonymy sets.

data split→	EN		**avg**	
model↓	@1	@5	@1	@5
XLMR	1.0	2.0	0.2	0.5
+ SAP$_{\text{all\_syn}}$	78.2	81.0	34.3	39.3
XLMR$_{\text{LARGE}}$	73.0	75.0	12.3	13.3
+ SAP$_{\text{all\_syn}}$	**78.3**	**81.3**	**39.0**	**44.2**

Table 5: Comparing BASE and LARGE models on XL-BEL. Both EN results and **avg** across all languages are reported. Full table available in Appendix Table 9.

continue training on general translation data (§2.2) after the previous UMLS-based SAP. With this variant, base multilingual LMs become powerful multilingual biomedical experts. We observe additional strong gains (cf., Table 2) with out-of-domain translation data: e.g., for MBERT the gains range from 2.4% to 12.7% on all languages except ES. For XLMR, we report Precision@1 boosts of >10% on RU, TR, KO, TH with XLMR+SAP$_{\text{en\_syn}}$, and similar but smaller gains also with XLMR+SAP$_{\text{all\_syn}}$.

We stress the case of TH, not covered in UMLS. Precision@1 rises from 11.5% (XLMR+SAP$_{\text{en\_syn}}$) to 30.9%$^{\uparrow 19.4\%}$ (XLMR+SAP$_{\text{all\_syn}}$(+en-th wt+ muse)), achieved through the synergistic effect of both knowledge types: **1) UMLS synonyms in other languages** push the scores to 20.6%$^{\uparrow 9.1\%}$; **2) translation knowledge** increases it further to 30.9%$^{\uparrow 10.3\%}$. In general, these results suggest that both external in-domain knowledge and general-domain translations boost the performance in resource-poor languages.

The More the Better (Table 4)? According to Table 4 (lower half), it holds almost universally that all_syn > en+{$LANG}_syn > en_syn/{$LANG}_syn on XLMR, that is, it seems that more in-domain knowledge (even in non-related languages) benefit cross-lingual transfer. However, for MBERT (Table 4, upper half), the trend is less clear, with en+{$LANG}_syn sometimes outperforming the all_syn variant. Despite modest performance differences, this suggests that the choice of source languages for knowledge transfer also plays a role; this warrants further investigations in future work.

Are Large Models (Cross-Lingual) Domain Experts (Table 5)? We also investigate the LARGE variant of XLMR, and compare it to its BASE variant. On English, XLMR$_{\text{LARGE}}$ gets 73.0% Precision@1, being in the same range as SAPBERT (78.7%), without SAP-tuning (Table 5). The scores without SAP fine-tuning on XLMR$_{\text{LARGE}}$, although much higher than of its BASE variant, decrease on other ('non-English') languages. At the same time, note that XLMR BASE achieves random-level performance without SAP-tuning. After SAP fine-tuning, on average, XLMR$_{\text{LARGE}}$+SAP still outperforms BASE models, but the gap is much smaller: e.g., we note that the performance of the two SAP-ed models is on par in English. This suggests that with sufficient knowledge injection, the underlying base model is less important (English); however, when the external data are scarce (other languages beyond English), a heavily parameterised large pretrained encoder can boost knowledge transfer to resource-poor languages.

5 Conclusion

We have introduced a novel cross-lingual biomedical entity task (XL-BEL), establishing a wide-coverage and reliable evaluation benchmark for cross-lingual entity representations in the biomedical domain in 10 languages, and have evaluated current SotA biomedical entity representations on XL-BEL. We have also presented an effective transfer learning scheme that leverages general-domain translations to improve the cross-lingual ability of domain-specialised representation models. We hope that our work will inspire more research on multilingual *and* domain-specialised representation learning in the future.

Acknowledgements

We thank the three reviewers and the AC for their insightful comments and suggestions. FL is supported by Grace & Thomas C.H. Chan Cambridge Scholarship. IV and AK are supported by the ERC Consolidator Grant LEXICAL (no. 648909) awarded to AK. NC kindly acknowledges grant-in-aid support from the UK ESRC for project EPI-AI (ES/T012277/1).

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Marco Basaldella, Fangyu Liu, Ehsan Shareghi, and Nigel Collier. 2020. COMETA: A corpus for medical entity linking in the social media. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3122–3137, Online. Association for Computational Linguistics.

Rachel Bawden, Kevin Bretonnel Cohen, Cristian Grozea, Antonio Jimeno Yepes, Madeleine Kittner, Martin Krallinger, Nancy Mah, Aurelie Neveol, Mariana Neves, Felipe Soares, Amy Siu, Karin Verspoor, and Maika Vicente Navarro. 2019. Findings of the WMT 2019 biomedical translation shared task: Evaluation for MEDLINE abstracts and biomedical terminologies. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 29–53, Florence, Italy. Association for Computational Linguistics.

Olivier Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32:D267–D270.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific language model pretraining for biomedical natural language processing. *arXiv:2007.15779*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Anne Lauscher, Ivan Vulić, Edoardo Maria Ponti, Anna Korhonen, and Goran Glavaš. 2020. Specializing unsupervised pretraining models for word-level semantic similarity. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1371–1383, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Sunwon Lee, Donghyeon Kim, Kyubum Lee, Jaehoon Choi, Seongsoon Kim, Minji Jeon, Sangrak Lim, Donghee Choi, Sunkyu Kim, Aik-Choon Tan, et al. 2016. BEST: next-generation biomedical entity search tool for knowledge discovery from biomedical literature. *PloS one*, 11(10):e0164680.

Yoav Levine, Barak Lenz, Or Dagan, Ori Ram, Dan Padnos, Or Sharir, Shai Shalev-Shwartz, Amnon Shashua, and Yoav Shoham. 2020. SenseBERT: Driving some sense into BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4656–4667, Online. Association for Computational Linguistics.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021. Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4228–4238, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Paul McNamee, James Mayfield, Dawn Lawrie, Douglas Oard, and David Doermann. 2011. Crosslanguage entity linking. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 255–263, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65, Florence, Italy. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Mujeen Sung, Hwisang Jeon, Jinhyuk Lee, and Jaewoo Kang. 2020. Biomedical entity representations with synonym marginalization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3641–3650, Online. Association for Computational Linguistics.

Chen-Tse Tsai and Dan Roth. 2016. Cross-lingual wikification using multilingual embeddings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 589–598, San Diego, California. Association for Computational Linguistics.

Elena Tutubalina, Artur Kadurin, and Zulfat Miftahutdinov. 2020. Fair evaluation in concept normalization: a large-scale comparative analysis for BERT-based models. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6710–6716, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Shikhar Vashishth, Rishabh Joshi, Denis Newman-Griffis, Ritam Dutt, and Carolyn Rose. 2020. MedType: Improving Medical Entity Linking with Semantic Type Prediction. *arXiv e-prints*, page arXiv:2005.00460.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R. Scott. 2019. Multi-similarity loss with general pair weighting for deep metric learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 5022–5030. Computer Vision Foundation / IEEE.

Yanshan Wang, Sijia Liu, Naveed Afzal, Majid Rastegar-Mojarad, Liwei Wang, Feichen Shen, Paul Kingsbury, and Hongfang Liu. 2018. A comparison of word embeddings for the biomedical natural language processing. *Journal of biomedical informatics*, 87:12–20.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, Florence, Italy. Association for Computational Linguistics.

A Appendix A

A.1 XL-BEL: Full Statistics

Table 1 in the main paper summarises the key statistics of the XL-BEL benchmark. It was extracted from the 20200601 version of Wikipedia dump. "sentences" refers to the number of sentences that contain biomedical mentions in the Wiki dump. "unique titles (Wiki page)" denotes the number of unique Wikipedia articles the biomedical mentions link to. "mentions" denotes the number of all biomedical mentions in the Wikipedia dump. "unique mentions" refers to the number of mentions after filtering out examples containing duplicated mention surface forms. "unique mentions$_{\text{mention!=title}}$" denotes the number of unique mentions that have surface forms different from the Wikipedia articles they link to. The 1k test sets for each language are then randomly selected from the examples in "unique mentions$_{\text{mention!=title}}$".

A.2 XL-BEL: Selection of Languages

Our goal is to select a diverse and representative sample of languages for the resource and evaluation from the full set of possibly supported languages. For this reason, we exclude some Romance and Germanic languages which were too similar to some languages already included in the resource (e.g., since we include Spanish as a representative of the Romance language, evaluating on related languages such as Portuguese or Italian would not yield additional and new insights, while it would just imply running additional experiments). The language list covers languages that are close to English (Spanish, German); languages that are very distant from English (Thai, Chinese, etc.); and also languages that are *in the middle* (e.g., Turkish, which is typologically different, but shares a similar writing script with English).

The availability of biomedical texts in Wikipedia also slightly impacted our choice of languages. The overlapping entities of Wikipedia and UMLS are not evenly distributed in the biomedical domain. For example, since animal species are comprehensively encoded in UMLS, they become rather dominant for certain low-resource languages. We manually inspected the distribution of the covered entities in each language to ensure that they are indeed representative biomedical concepts. Languages with heavily skewed entity distributions are filtered out. E.g., biomedical concepts in Basque Wikipedia are heavily skewed towards plant and animal species (which are valid UMLS concepts but not representative enough). As a result, we dropped Basque as our evaluation language. The current 10 languages all have a reasonably fair distribution over biomedical concepts categories.

A.3 UMLS Data Preparation

All our UMLS fine-tuning data for SAP is extracted from the MRCONSO.RRF file downloaded at https://www.nlm.nih.gov/research/umls/licensedcontent/umlsarchives04.html#2020AA. The extracted data includes 147,706,62 synonyms distributed in more than 20 languages. The detailed statistics are available in Table 6.

code	language	# synonyms	percentage
EN	English	10,277,246	69.6%
ES	Spanish	1,575,109	10.7%
JA	Japanese	329,333	2.2%
RU	Russian	291,554	2.0%
DE	German	231,098	1.6%
KO	Korean	145,865	1.0%
ZH	Chinese	80,602	0.5%
TR	Turkish	51,328	0.3%
FI	Finnish	24,767	0.2%
TH	Thai	0	0.0%
FR	French	428,406	2.9%
PT	Portuguese	309,448	2.1%
NL	Dutch	290,415	2.0%
IT	Italian	242,133	1.3%
CS	Czech	196,760	0.7%
NO	Norwegian	63,075	0.4%
PL	Polish	51,778	0.4%
ET	Estonian	31,107	0.2%
SV	Swedish	29,716	0.2%
HR	Croatian	10,035	0.1%
EL	Greek	2,281	<0.1%
LV	Latvian	1,405	<0.1%
	Total	147,706,62	100%

Table 6: The amount of UMLS synonyms per language. The first 10 languages are included in our XL-BEL test languages. However, note that Thai has no UMLS data.

A.4 Translation Data

The full statistics of the used word and phrase translation data are listed in Table 7. The "muse" word translations are downloaded from https://github.com/facebookresearch/MUSE while the Wikititle pairs ("wt") are extracted by us, and are made publicly available.

A.5 Pretrained Encoders

A complete listing of URLs for all used pretrained encoders hosted on huggingface.co is provided in Table 8. For monolingual models of each language,

#↓, language→	EN-ES	EN-DE	EN-FI	EN-RU	EN-TR	EN-KO	EN-ZH	EN-JA	EN-TH
muse	112,583	101,931	43,102	48,714	68,611	20,549	39,334	25,969	25,332
wt	1,079,547	1,241,104	338,284	886,760	260,392	319,492	638,900	547,923	107,398

Table 7: Statistics of muse word translations ("muse") and Wikipedia title pairs ("wt").

we made the best effort to select the most popular one (based on download counts).

A.6 Full Table for Comparing with LARGE Models

Table 9 list results across all languages for comparing BASE and LARGE models.

A.7 Future Work

Investigating Other Cross-Lingual Transfer Learning Schemes. We also explored adapting multilingual sentence representation transfer techniques like Reimers and Gurevych (2020) that leverage parallel data. However, we observed no improvement comparing to the main transfer scheme reported in the paper. We plan to investigate existing techniques more comprehensively, and benchmark more results on XL-BEL in the future.

Comparison with in-Domain Parallel Data. While we used general-domain bitexts to cover more resource-poor languages, we are aware that in-domain bitexts exist among several "mainstream" languages (EN, ZH, ES, PT, FR, DE, Bawden et al. 2019).[8] In the future, we plan to also compare with biomedical term/sentence translations on these languages to gain more insights on the impact of domain-shift.

A.8 Number of Model Parameters

All BASE models have ≈110M parameters while LARGE models have ≈340M parameters.

A.9 Hyperparameter Optimisation

Table 10 lists the hyperparameter search space. Note that the chosen hyperparameters yield the overall best performance, but might be suboptimal in any single setting. We used the same random seed across all experiments.

A.10 Software and Hardware Dependencies

All our experiments are implemented using PyTorch 1.7.0 with Automatic Mixed Precision (AMP)[9] turned on. The hardware we use is listed in Table 11. On this machine, the SAP fine-tuning procedure generally takes 5-10 hours with UMLS data. SAP fine-tuning with translation data takes 10 minutes to 5 hours, depending on the amount of the data. Inference generally takes <10 minutes.

[8] http://www.statmt.org/wmt19/biomedical-translation-task.html

[9] https://pytorch.org/docs/stable/amp.html

model	URL
monolingual models	
SapBERT	https://huggingface.co/cambridgeltl/SapBERT-from-PubMedBERT-fulltext
esBERT	https://huggingface.co/dccuchile/bert-base-spanish-wwm-uncased
deBERT	https://huggingface.co/dbmdz/bert-base-german-uncased
fiBERT	https://huggingface.co/TurkuNLP/bert-base-finnish-uncased-v1
ruBERT	https://huggingface.co/DeepPavlov/rubert-base-cased
trBERT	https://huggingface.co/loodos/bert-base-turkish-uncased
krBERT	https://huggingface.co/snunlp/KR-BERT-char16424
zhBERT	https://huggingface.co/bert-base-chinese
jaBERT	https://huggingface.co/cl-tohoku/bert-base-japanese
thBERT	https://huggingface.co/monsoon-nlp/bert-base-thai
cross-lingual models	
mBERT	https://huggingface.co/bert-base-multilingual-uncased
XLMR	https://huggingface.co/xlm-roberta-base
XLMR$_{\text{LARGE}}$	https://huggingface.co/xlm-roberta-large
XLMR$_{\text{LARGE-XNLI}}$	https://huggingface.co/joeddav/xlm-roberta-large-xnli
XLMR$_{\text{LARGE-SQUAD2}}$	https://huggingface.co/deepset/xlm-roberta-large-squad2

Table 8: A listing of HuggingFace URLs of all pretrained models used in this work.

language→	EN		ES		DE		FI		RU		TR		KO		ZH		JA		TH		**avg**	
model↓	@1	@5	@1	@5	@1	@5	@1	@5	@1	@5	@1	@5	@1	@5	@1	@5	@1	@5	@1	@5	@1	@5
SapBert	**78.7**	**81.6**	47.3	51.4	22.7	24.7	8.2	10.2	5.8	6.0	26.4	29.7	2.0	2.4	1.9	2.2	3.0	3.2	3.1	3.4	19.9	21.6
SapBert$_{\text{all\_syn}}$	78.3	80.7	55.6	61.3	30.0	34.2	11.8	14.8	9.3	11.3	35.5	39.5	2.0	2.4	6.4	8.2	6.9	8.3	3.0	3.3	23.9	26.4
XLMR	1.0	2.0	0.3	0.7	0.0	0.1	0.1	0.2	0.1	0.2	0.4	0.5	0.0	0.3	0.1	0.2	0.2	0.4	0.0	0.1	0.2	0.5
XLMR + Sap$_{\text{all\_syn}}$	78.2	81.0	56.4	62.7	31.8	37.3	18.6	22.2	35.4	41.2	42.8	48.9	16.7	21.4	18.8	23.0	24.0	28.1	20.6	27.5	34.3	39.3
XLMR$_{\text{LARGE}}$	73.0	75.0	20.7	24.6	7.8	9.1	1.9	2.7	3.0	3.3	11.8	13.5	1.2	1.2	0.7	0.9	1.6	1.8	0.9	1.2	12.3	13.3
XLMR$_{\text{LARGE-XNLI}}$	72.6	75.1	30.1	33.5	10.7	12.2	3.4	4.6	5.9	7.4	16.4	18.4	1.9	2.6	1.3	2.0	2.0	2.5	1.3	2.0	14.6	16.0
XLMR$_{\text{LARGE-SQUAD2}}$	74.6	76.2	31.4	35.3	11.9	13.2	3.5	4.4	5.2	6.5	16.9	19.2	1.4	1.5	0.6	0.9	1.8	2.1	2.0	2.3	14.9	16.2
XLMR$_{\text{LARGE}}$ + Sap$_{\text{all\_syn}}$	78.3	81.3	**61.0**	**66.8**	**35.0**	**40.0**	**25.2**	**29.2**	**41.9**	**47.3**	**46.1**	**52.4**	**22.2**	**26.7**	**23.5**	**29.0**	**28.5**	**33.6**	**28.7**	**35.5**	**39.0**	**44.2**

Table 9: A comparison of Base (upper half) and Large (lower half) multilingual encoders on XL-BEL.

hyperparameters	search space
pretraining learning rate	2e-5
pretraining batch size	512
pretraining training epochs	1
bitext fine-tuning learning rate	2e-5
bitext fine-tuning batch size	{64, 128, 256*}
bitext fine-tuning epochs	{1, 2, 3, 4, 5*, 10}
max_seq_length of tokeniser	25
λ in Online Mining	0.2
α in MS loss (Eq. (1))	2
β in MS loss (Eq. (1))	50
ϵ in MS loss (Eq. (1))	1

Table 10: Hyperparameters along with their search grid. $*$ marks the values used to obtain the reported results. The hparams without any defied search grid are adopted directly from Liu et al. (2021).

hardware	specification
RAM	192 GB
CPU	Intel Xeon W-2255 @3.70GHz, 10-core 20-threads
GPU	NVIDIA GeForce RTX 2080 Ti (11 GB) × 4

Table 11: Hardware specifications of the used machine. For Large model training, we use another server with two NVIDIA GeForce RTX 3090 (24 GB).