Further with Knowledge Graphs. M. Alam et al. (Eds.)
AKA Verlag and IOS Press, 2021

doi:10.3233/SSW210037

LLOD-Driven Bilingual Word Embeddings Rivaling Cross-Lingual Transformers in Quality of Life Concept Detection from French Online Health Communities

Katharina ALLGAIER [a], Susana VERÍSSIMO [a], Sherry TAN [a 1],
Matthias ORLIKOWSKI [a] and Matthias HARTUNG [a]
[a] *Semalytix GmbH, Bielefeld, Germany*
e-mail: {first.last}@semalytix.com

Abstract. We describe the use of Linguistic Linked Open Data (LLOD) to support a cross-lingual transfer framework for concept detection in online health communities. Our goal is to develop multilingual text analytics as an enabler for analyzing health-related quality of life (HRQoL) from self-reported patient narratives. The framework capitalizes on supervised cross-lingual projection methods, so that labeled training data for a source language are sufficient and are not needed for target languages. Cross-lingual supervision is provided by LLOD lexical resources to learn bilingual word embeddings that are simultaneously tuned to represent an inventory of HRQoL concepts based on the World Health Organization's quality of life surveys (WHOQOL). We demonstrate that lexicon induction from LLOD resources is a powerful method that yields rich and informative lexical resources for the cross-lingual concept detection task which can outperform existing domain-specific lexica. Furthermore, in a comparative evaluation we find that our models based on bilingual word embeddings exhibit a high degree of complementarity with an approach that integrates machine translation and rule-based extraction algorithms. In a combined configuration, our models rival the performance of state-of-the-art cross-lingual transformers, despite being of considerably lower model complexity.

Keywords. Multilingual Text Analytics, Linguistic Linked Open Data, Bilingual Word Embeddings, Cross-lingual Transformers, Health-related Quality of Life

1. Introduction

Increasingly, multilingual language resources are available as Linguistic Linked Open Data (LLOD) [1] which model relations between resources and include rich metadata with standardized, non-proprietary technologies – a trend which promises to lead to improved multilingual NLP systems. However, how to effectively utilize these resources is

[1] This author contributed to the results presented in this paper during an internship at Semalytix.

not self-evident, in particular for specialized domains. One example of such a domain are posts from online health communities, i.e., web fora and similar systems focused on health topics used by patients, caregivers and/or professionals in a wide range of languages. Online health communities are a relevant data source for a range of emerging application areas, such as public health monitoring or evidence generation for regulatory drug approval [2], which entail analysing patients' experiences beyond clinical trials. A central aspect of these so-called patient-reported outcomes is health-related quality of life (HRQoL) [3].

In this paper, we focus on classifying posts into categories derived from facets of HRQoL as described in the World Health Organization's quality of life surveys (WHO-QOL) [4], e.g., pain and discomfort, work capacity, financial resources. We approach the problem of predicting HRQoL facets across languages via a multitude of individual binary classifiers trained using a cross-lingual transfer learning framework based on bilingual lexica available as multilingual LLOD. The combination of LLOD and transfer learning is motivated by the flexibility required to predict a large number of HRQoL facets (we consider a total of 19 facets) in a multilingual setting: Transfer learning allows us to train classifiers for different languages based on training data in a single source language, without the need of additional annotated data for each target language. LLOD enables us to leverage a breadth of existing multilingual resources and infer lexica for additional language pairs using implicit links between resources. We demonstrate in the reported experiments that this not only a benefit in terms of flexibility, but also leads to improved performance for our cross-lingual transfer learning approach in comparison to a medical lexicon directly applicable to the evaluated language pair.

In more technical detail, our approach is based on word embeddings and cross-lingual supervision via token-level lexica (supervised bilingual word embeddings). Thus, the training procedure and resulting models are considerably less complex than state-of-the-art cross-lingual zero-shot models, which are based on contextualized representations learnt via pre-training transformer-based language models on massive multilingual corpora. Consequently, we present evaluation results comparing our approach to a language-model-based classifier for the case of transfer from English to French for detection of HRQoL facets in posts from online health communities. We find that our models, when combined with a baseline approach that integrates machine translation and rule-based extraction algorithms, are strong contestants to cross-lingual transformers.

2. Related Work

Given our focus on exploring the factors of effectively applying LLOD resources to cross-lingual transfer learning for text classification, we build on supervised approaches for learning bilingual word representations which are able to incorporate existing seed lexica (cf. [5]), but do not require additional supervision or resources, e.g. parallel or aligned corpora as in early work on cross-lingual transfer [6]. In particular, we adopt workflows for using LLOD in cross-lingual transfer learning based on task-informed, bilingual word embeddings (adopted from bilingual sentiment embeddings [7]) presented in [8] and apply them to a different target language (Spanish vs. French), a much more varied task (HRQoL aspect detection vs. sentiment analysis) and different text genre (online health community posts vs. medical experts' interview transcripts).

As our research questions imply the availability of applicable lexica, unsupervised or weakly supervised approaches for inducing bilingual word embeddings [9,10,11] are only indirectly relevant to our work. However, we plan to compare against them in future work, especially given that claims of comparable or even superior performance of unsupervised methods (e.g., [12]) have been called into question [13,14], in particular when evaluated on actual downstream tasks instead of bilingual lexicon induction [15].

Since the introduction of the Transformer neural architecture and pre-training via language modelling objectives on massive corpora, cross-lingual representations derived from these models, e.g. multilingual versions of BERT [16] or XLM-R [17], became state-of-the-art on a large number of multilingual problems. This comes, however, with a noticeable added cost in comparison to bilingual word embeddings in terms of model complexity and computational resources, especially during training (cf. [18]). We explore this performance-complexity trade-off by comparing our models based on bilingual word embeddings against a zero-shot cross-lingual classifier based on XLM-R.

Using indirect connections between translation lexica to automatically construct a bilingual lexicon via a pivot language goes at least back to [19]. Lexicon induction techniques using LLOD, and Apertium RDF in particular, were explored in [20,21].

3. Language- and Task-informed Cross-lingual Transfer Learning

Our approach to language- and task-informed transfer learning (LTTL) relies on the framework described in our previous work [8]. Using this architecture based on bilingual word embeddings [7], task-informed bilingual embedding spaces can be learned for any task which can be framed as text classification. Following this idea, we apply LTTL to HRQoL concept detection in this paper.

For training a task-specific model LTTL requires 1) monolingual word embeddings in both the source and target language, 2) ground-truth annotations in the source language, and 3) a bilingual dictionary that maps tokens from the source language to their translations in the target language (see Section 4). Annotations in the target language are required for evaluation only.

During training (Figure 1), word embeddings are looked up for the tokens in each document in the source-language annotated corpus and averaged in order to yield document representations a_S. A projection matrix M_S is trained to map a_S to a task-specific vector z_s, which is then passed to a softmax layer to derive the predicted label. By minimizing the cross-entropy loss between the predicted and the annotated labels, M_S and the parameters of the softmax layer are learnt to produce better task-specific predictions. Simultaneously, for every pair in the bilingual dictionary, we look up their word embeddings in the respective monolingual embedding space and project them using M_S and M_T (a corresponding matrix in the target language), respectively. Both matrices are jointly optimized to minimize the Euclidean distance between the projected embeddings in a shared bilingual embedding space, so that the projections from the target language are as close as possible to the projections from the source language for which monolingual task-specific supervision is available.

When using a trained LTTL model to classify a target-language document (Figure 2), we apply the same steps as during training based on target-language embeddings (embedding lookup, averaging, projection, prediction using the softmax layer). The projection

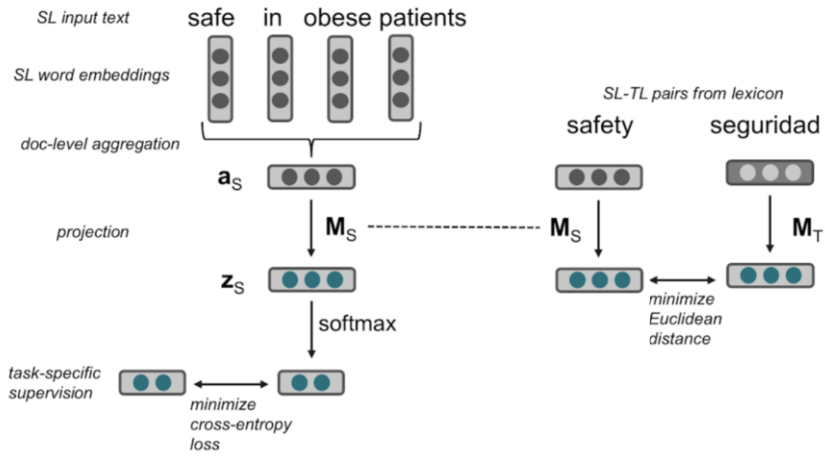

Figure 1. Training LTTL on a source-language (SL) annotated corpus and a source-language to target-language (TL) bilingual lexicon using TL and SL word embeddings to represent individual tokens

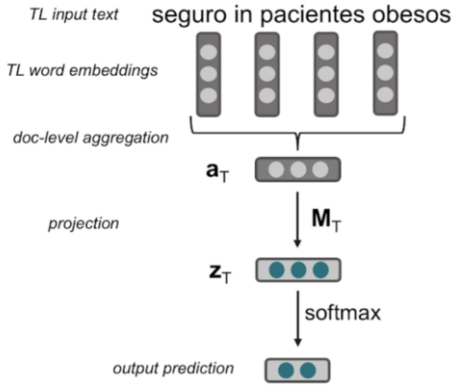

Figure 2. Predictions with LTTL on target-language (TL) text using TL word embeddings to represent individual tokens

step, however, is calculated using the matrix M_T which was optimized to project target-language embeddings close to the projections from the task-informed, source-language projection matrix M_S.

4. Language Resources

In this section we describe the relevant lexical resources used in LTTL. A detailed description of the LLOD pipeline used to generate these resources and the individual processing steps involved is presented in [8]. While our focus is on Apertium RDF as a

bilingual lexicon in this paper, these workflows have the potential of growing the LLOD cloud over time both in terms of data volume and richness of available resources.

4.1. Bilingual Translation Dictionaries

The bilingual lexica used in our experiments contain word-level translation pairs from a source to a target language. Lexica vary in terms of vocabulary size, the type of knowledge provided, origin, and purpose. In our experiments, we selected lexica according to the criteria of domain- and task specificity. Accordingly, broad-coverage, open-domain and medical lexica were used as described below. We deduplicated entries in all lexica during pre-processing.

Apertium lexica[2] are very comprehensive open-domain, broad-coverage lexica. Originally, this resource was generated for an open-source machine translation platform [22]. Apertium lexica used in our work were converted into RDF using the FINTAN platform [23] and published as linked data. These lexica contain entries annotated as nouns, proper nouns, verbs, adjectives and adverbs.

MeSpEn Glossaries[3] are lexica specific to the biomedical domain. A total of 46 bilingual medical glossaries for various language pairs are available. The lexica were generated based on hand-crafted glossaries made by professional translators [24].

4.2. Cross-lingual Lexicon Induction

In some cases, bilingual lexica of interest for a given task or domain may not be available for a language pair of interest. In this case, translation pairs can be inferred via triangulation [19]. This approach consists of leveraging available lexical resources in the source language and a pivot language, i.e., a language which has correspondences to the source and target languages, as a means to create a mapping between both. More specifically, we generated a bilingual dictionary for the language pair English-French based on Apertium RDF using Spanish as pivot language as follows: For each entry that links a source language term t_S to its translation t_P in the pivot language, if there is an entry linking t_P to a target language term t_T, a translation from t_S to t_T can be inferred and stored in a newly created source-target lexicon. Subsequently, (i) all duplicate entries and (ii) entries with divergent part-of-speech categories[4] in t_S and t_T are removed from the resulting lexicon. This induction procedure yields an induced open-domain EN–FR lexicon comprising 15,703 entries; for comparison, the existing MeSpEn Glossary resource comprises 6,571 domain-specific EN–FR entries.

4.3. Monolingual Word Embeddings

In addition to the bilingual lexical resources described above, LTTL also requires monolingual word embeddings. In our experiments, we use publicly available word embed-

[2] https://github.com/acoli-repo/acoli-dicts/tree/master/stable/apertium/apertium-rdf-2020-03-18

[3] https://doi.org/10.5281/zenodo.2205690

[4] This procedure relies on the PoS information that is integrated into Apertium 2.0 via mapping lexical entries to the LexInfo ontology [8].

Embeddings	Language	Type	Vocabulary Size	Vector Dimensions
google-news	English	open-domain	55,627	300
fr-wiki	French	open-domain	2,500,733	300

Table 1. Overview of monolingual embeddings used in our current experiments with the LTTL framework.

dings[5] pre-trained on different corpora. Table 1 describes the language, domain, vocabulary size and dimensionality of the used embeddings.

5. Data Sets

We use an English–French comparable corpus made up of anonymized posts from several openly accessible medical and health-related online fora to generate and/or annotate training and evaluation data sets for different HRQoL facets. The corpus contains extremely varied, uncontrolled language as the texts are mostly authored by patients and their relatives. This can be observed below for three representative examples from diverse QoL facets, with (a) denoting the original French texts and (b) their English machine-generated translations.

(1) **Sleep and rest** (SR)

(a) *Cetrizirine c'est quoi les filles ? Depuis un certain temps je suis insomniaque ...tisane...chronodorm... mélisse...rien n'y fait*

(b) What's Cetrizirine girls? I've been having insomnia for some time... herbal tea... chronodorm... lemon balm... nothing helps...

(2) **Activities of daily living** (DL)

(a) *Actuellement en arrêt maladie du a mon cancer j'ai de la chimiothérapie a l'hôpital. Les écoles de ma commune ferme je suis incapable de m'occuper de mes enfants.*

(b) Currently on hold disease from my cancer I have chemotherapy in the hospital. The schools in my commune are closed I am unable to take care of my children.

(3) **Body image and appearance** (BA)

(a) *je ne veux pas forcément que ça se sache que je suis malade et avec perruque et maquillage je veux passer incognito lol ... car je suis qlq un qui manque bcp de confiance en soi.*

(b) But I don't want him to shout it from the roof-tops at school or anything else because I don't necessarily want it to be known that I'm sick and with wigs and make-up I want to go incognito lol ... because I'm one who lacks a lot of self-confidence.

[5] Available from https://drive.google.com/open?id=1GpyF2h0j8K5TKT7y7Aj0OyPgpFc8pMNS and https://wikipedia2vec.github.io/wikipedia2vec/pretrained/, respectively.

Quality of life dimensions	Facets within dimensions	Size of training set	
		positive	negative
Physical health	Energy and Fatigue (EF)	3000	3000
	Pain and discomfort (PD)	3000	3000
	Sleep and rest (SR)	707	689
Psychological health	Body image and appearance (BA)	1164	1139
	Negative feelings (NF)	1464	1428
	Positive feelings (PF)	3000	3000
	Thinking, learning, memory and concentration (TM)	380	379
Level of independence	Mobility (MB)	2112	2006
	Activities of daily living (DA)	842	830
	Work capacity (WO)	606	590
Social relations	Personal relationships (PR)	3000	3000
	Sexual activity (SA)	44	44
	Social support (SO)	834	813
Environment	Financial resources (FR	1488	1446
	Health and social care (HC)	1625	1577
	Home environment (HE)	751	745
	Participation in and opportunities for recreation and leisure (RL)	3000	3000
	Physical environment (PE)	3000	3000
	Transport (TR)	1567	1518

Table 2. Overview of QoL dimensions and contained facets with their training data size in terms of number of posts (facets in boldface are part of the manually annotated gold standard)

5.1. English training data

We generate annotation labels for the English data using a rule-based pattern matching engine from the in-house Semalytix technology stack. These rules, in addition to plain text matching, include regular expressions to capture morphological variation, part-of-speech tagging, dependency syntax or knowledge graph type constraints. This rule-based system allows for rapid generation of labeled training data for 19 HRQoL concepts that are in scope in this paper. Data set sizes vary per HRQoL concept depending on the available number of matches produced by the monolingual rules and are capped at 3,000 positive and negative examples per concept (6,000 in total). The resulting English data sets are randomly split into a training and development set (80/20). Table 2 provides an overview of all concepts and their respective data volume.

Figure 3. Label propagation from labeled source language (EN) to unlabeled target language (FR) documents.

5.2. French evaluation data (Silver Standard)

In light of the multitude of HRQoL concepts under investigation in this study, we rely on a heuristic label propagation procedure in order to create a large-scale evaluation corpus for validation purposes in the target language. To make use of the described monolingual rule system for texts that are not written in English, the target language texts are algorithmically translated into English via DeepL[6]. Thus, the rule engine can be run on the translated texts in the same way as on originally English ones. The resulting concept labels are then propagated back to the target-language documents. An illustration of this process is depicted in Figure 3.

However, it needs to be emphasized that the resulting target language labels were not manually checked for correctness. Hence, even though the underlying rule-based classifiers available for English are optimized for precision, the test collection resulting from this procedure must be considered a silver standard. Again, data set sizes vary depending on the available number of matches. They are capped at 100 positive and 100 negative examples per facet. The French target data sets are used for evaluation exclusively and thus are divided into a development and test set (50/50).

5.3. French evaluation data (Gold Standard)

In the interest of a thorough evaluation of concept detection performance in the target language for at least a subset of concepts, we selected one concept from each QoL domain (highlighted in bold face in Table 2) to create a hand-curated gold standard. These gold standard data sets consist of 100 positive and 100 negative texts samples per HRQoL facet which each were verified to be correct by manual annotation. As the French silver standard data, these data sets are also exclusively used for evaluation and evenly divided into a development and test set (50/50).

[6] https://www.deepl.com/pro?cta=header-pro/

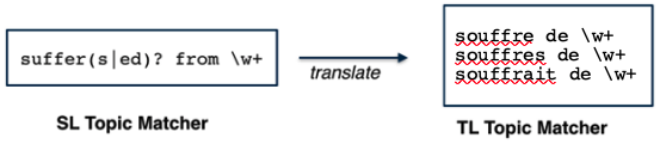

Figure 4. Illustration of translation procedure for Baseline 1.

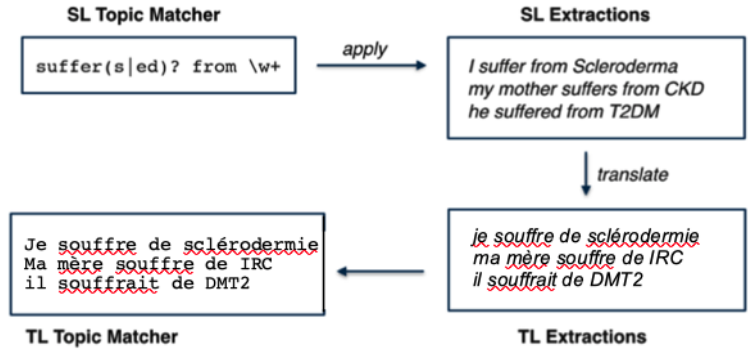

Figure 5. Illustration of translation procedure for Baseline 2.

6. Baseline Models based on Machine Translation and Rules

As comparison to our LTTL model, we generate two baseline models based on machine translation in combination with the previously described rule engine (cf. Section 5.1). As illustrated in Figure 4, our approach for Baseline 1 (BL1) is to first extract all rules used in the monolingual rule engine for English for each required concept in the source language (SL). These are then directly translated into the target language (TL) using the DeepL translation API[7]. The resulting TL rule sets can subsequently be used as rule-based extractors on the TL test set such that matching documents are classified as positive instances of the respective concept, others as negative ones.

Baseline 2 (BL2) is following a slightly different approach (cf. Figure 5). First the original monolingual rules for each concept are applied to the English training data. Then, all English phrases that match those patterns are extracted and translated into the target language. Subsequently, those extractions (which in comparison to Baseline 1 do not usually contain any regular expressions or other formal constraints) are then used as target language extraction rules and run on the TL test set, analogously to Baseline 1.

7. Evaluation

The experiments reported in this section address the problem of HRQoL concept detection from French online health communities. We simulate a real-world setting in which

[7]This includes a shallow post-processing step to remove broken rule syntax.

no labeled training examples are available in the target language. Therefore, we approach the task in a cross-lingual manner, transferring knowledge that is available in existing models or resources for English to French as target language. Our primary interest is in answering the following research questions:

1. How does cross-lingual concept detection performance via LTTL compare to state-of-the-art cross-lingual transformer architectures?
2. Focusing on the specific lexical resource needs of LTTL, what is the impact of a large, open-domain lexicon induced from Apertium RDF [25] via a pivot language vs. a smaller, biomedical, directly applicable lexicon [26]?
3. How does LTTL concept detection performance differ across HRQoL concepts, i.e., can our approach effectively be applied to a large number of different concepts?

7.1. Experiment 1: Gold Standard Evaluation

7.1.1. Settings.

In a first experiment, we evaluate LTTL against the gold standard described in Section 5.3. For comparison, we obtained results for both baselines BL1 and BL2 introduced in Section 6. Additionally, we explore the setting of combining the LTTL model with each baseline in a sequential way. This was done by first executing LTTL and subsequently feeding all data points (from both the positive and the negative samples) that had been classified as negative by the model into the respective baseline.

Furthermore, we challenge LTTL in another comparison against the state-of-the art cross-lingual XLM-R model [17]. It is a transformer-based multilingual masked neural language model that is pre-trained for cross-lingual NLP tasks. In our use case, the model is fine-tuned on the English task-specific data and then tested on French evaluation data where it performs zero-shot cross-lingual classification.

7.1.2. Results and Discussion.

Results for this experiment are shown in Table 3 in terms of precision, recall and F1 measure for the positive class. We observe that, for four among the five concepts under investigation, LTTL outperforms both baselines based on machine translation (BL1 and BL2). While both baselines show divergent patterns across concepts (favoring precision on some concepts, recall on others), they are largely complementary with LTTL: With *Positive Feelings* as an exception, the sequential combinations of LTTL with one of BL1 or BL2 yield a boost in classification performance over LTTL in isolation. Apparently, this blend of cross-lingual word embeddings with cross-lingual rule engineering constitutes an effective approach to the HRQoL concept detection problem. To some extent, this still holds in view of the performance of the neural state-of-the-art XLM-R model, which outperforms LTTL+BL1/2 in three out of five cases, but obtains lower results for *Recreation and Leisure* and *Positive Feelings*.

The excellent generalization properties of XLM-R notwithstanding, these results suggest that cross-lingual HRQoL concept detection does not necessarily require the heavy machinery of cross-lingual transformer models in all facets of interest. We argue that, given the much higher model complexity of cross-lingual transformers, architectures based on bilingual word embeddings such as LTTL may pose a practical compro-

Model	Lexicon	WO			SA			SR			RL			PF		
		Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1
LTTL	ES PoS	0.60	0.74	0.66	0.53	0.92	0.67	0.63	0.92	0.75	0.71	0.78	0.74	0.65	0.74	**0.69**
	MedGl	0.59	0.78	0.67	0.49	0.81	0.61	0.70	0.84	0.76	0.73	0.44	0.55	0.75	0.12	0.21
LTTL +BL1	ES PoS	0.64	0.82	0.72	0.56	0.99	0.71	0.64	0.94	0.76	0.50	1.00	0.67	0.51	0.81	0.63
	MedGl	0.59	0.81	0.68	0.53	0.95	0.68	0.69	0.90	0.78	0.50	1.00	0.67	0.50	0.79	0.62
LTTL +BL2	ES PoS	0.65	0.84	0.73	0.56	0.99	0.71	0.64	0.95	0.76	0.71	0.87	0.78	0.68	0.23	0.34
	MedGl	0.60	0.83	0.70	0.53	0.93	0.67	0.69	0.91	0.78	0.73	0.69	0.71	0.69	0.20	0.31
Baseline 1		1.00	0.08	0.15	1.00	0.84	**0.92**	0.97	0.37	0.54	0.50	0.98	0.66	0.50	0.79	0.62
Baseline 2		1.00	0.15	0.26	1.00	0.82	0.90	0.96	0.63	0.76	0.89	0.41	0.56	0.69	0.20	0.31
XLM-R		0.97	0.68	**0.80**	0.76	0.92	0.83	0.98	0.80	**0.88**	0.80	0.74	**0.77**	0.74	0.58	0.65

Table 3. Results for EN-FR concept transfer via both baseline models and LTTL, separately and in sequential combination for all 5 gold standard data sets. Lexicons refer to the ES single pivot version including PoS information (ES PoS) and the MeSpEn medical glossary (MedGl). Results are reported in terms of precision, recall and F1 measure for the positive class. Best F1 performance per concept is highlighted in bold. Concept abbreviations are in line with Table 2.

mise in many application scenarios. Moreover, we noted in additional experiments not reported here that when using smaller data sets (comparable to the one available for *Sexual Activity*), the margin between LTTL and XLM-R results becomes much narrower, which might suggest that LLOD-based bilingual word embeddings can cope better with smaller sets of training data. This conjecture requires deeper investigation in future work.

7.2. Experiment 2: Evaluation against Large-scale Silver Standard

7.2.1. Settings.

In a second experiment, we investigate cross-lingual classification performance for all 19 HRQoL concepts summarized in Table 2. We run individual text classification models that are instantiated from LTTL on each of these concepts and evaluate them against the silver standard presented in Section 5.3. Besides enabling a large-scale comparison across this multitude of concepts, this experiment is mainly designed to explore the resource requirements of LLOD-based cross-lingual transfer learning: Given that Apertium RDF does not include English–French translations, we induced a bilingual lexicon via the pivot language Spanish. Here, we want to assess the prospects of LLOD-based lexicon induction relative to a readily existing English-French bilingual medical lexicon MeSpEn Glossaries.

7.2.2. Results and Discussion.

Table 4 shows the results of this experiment in terms of F1 measure for the positive class. While LTTL surpasses both baselines for a substantially large number of concepts, only in a minority of cases (4 out of 19) it is conversely outperformed by one of them, with BL1 and BL2 not showing a clear trend to one outperforming the other in the majority of cases. Being designed as precision-oriented extraction rules for English documents,

Facet	BL 1	BL 2	LTTL	LTTL +BL1	LTTL +BL2	LTTL	LTTL +BL1	LTTL +BL2
				ES PoS			MedGl	
DA	0.56	0.56	0.69	**0.75**	0.73	0.62	0.73	0.72
BA	0.67	0.58	0.68	0.68	0.68	**0.70**	**0.70**	**0.70**
EF	0.65	**0.83**	0.35	0.68	*0.83*	0.60	0.68	0.77
FR	0.56	0.67	0.67	0.67	0.66	0.73	**0.75**	0.67
HC	0.55	0.23	0.67	0.67	0.67	**0.71**	0.70	0.70
HE	0.52	0.34	0.71	**0.74**	0.71	0.20	0.55	0.43
MB	0.55	0.49	0.66	**0.74**	0.71	0.67	0.67	0.67
NF	0.20	0.29	0.52	0.56	0.60	0.67	0.66	**0.68**
PD	0.49	0.49	0.57	**0.74**	**0.74**	0.65	0.66	0.66
PR	0.24	**0.82**	0.67	0.67	0.67	0.63	0.67	0.77
PE	0.50	0.52	0.71	**0.73**	**0.73**	0.43	0.71	0.72
PF	0.65	0.43	0.49	**0.67**	**0.67**	0.00	0.65	0.43
RL	0.66	0.56	0.73	0.67	**0.74**	0.67	0.66	0.69
SA	**0.86**	0.83	0.45	0.83	0.81	0.62	0.77	0.76
SR	0.43	0.71	0.72	0.72	**0.79**	0.63	0.69	0.75
SO	0.04	0.67	**0.78**	0.77	0.67	0.58	0.52	0.67
TM	0.41	0.36	0.58	0.65	0.63	**0.69**	0.68	0.68
TR	0.52	0.48	**0.73**	0.71	0.71	0.36	0.57	0.54
WO	0.17	0.31	0.67	0.69	0.70	**0.77**	0.76	0.76

Table 4. Results for EN–FR concept transfer evaluated on 19 silver standard QoL facets for both baseline models and LTTL separately and in sequential combination, denoted by F1-measure for the positive class. Abbreviations stem from Table 2. Lexicons refer to the ES single pivot version including PoS information (ES PoS) and the MeSpEn medical glossary (MedGl). Bold results highlight best performance for a given concept.

most of the baselines still favour precision after being transferred to French: For roughly 90% of the concepts at least one of the baselines shows a noticeably better precision than LTTL. However, apart from a small number of cases, LTTL benefits from a much higher recall, which results in a better overall performance in terms of F1.

Therefore, the strong degree of complementarity between LTTL and BL1/2 observed in Experiment 1 is confirmed in this large-scale setting as well: For the majority of concepts the performance of the LTTL+Baselines combination exceeds LTTL and is among the best configurations. This is the case for about 70% of the tasks, while in 90-95% of them LTTL+BL1/2 performs better or at least equally well as LTTL. Winning results

are obtained by the combined models by a sometimes considerable margin compared to LTTL.

Regarding the different lexicon configurations employed, it becomes evident that the automatically created open-domain lexicon which was induced leveraging LLOD information from Apertium RDF (denoted as ES-PoS in the table) performs better than the directly available medical domain-specific lexicon (denoted as MedGl in the table) in 12 out of 19 concepts. This shows that LLOD-based lexicon induction is a useful flexible process that could potentially be refined further in future work in order to boost performance even more.

8. Summary and Conclusions

We presented LTTL as a language- and task-informed framework for cross-lingual transfer learning. LTTL can be flexibly used in order to induce bilingual task-specific word embeddings as lexical representations for NLP models that are needed for multilingual text classification tasks. Being embedded into an LLOD exploitation pipeline, LTTL is flexibly applicable to different languages and for various tasks, which we demonstrated in this paper for the task of detecting HRQoL concepts from French online health communities.

In the experiments reported, we showed that it can be employed even when a bilingual lexicon for a particular language pair is not readily available, thanks to LLOD-driven lexicon induction via one or more pivot languages. Furthermore, the LTTL model can effectively be combined sequentially with rule-based concepts detectors, resulting in a noticeable increase of classification performance, all while making use of openly available LLOD resources. Comparing LTTL against the state-of-the-art cross-lingual neural language model XLM-R, we find that the HRQoL concept detection task does not necessarily lean itself against the high complexity of transformer-based architectures. In fact, our results suggest that architectures based on bilingual word embeddings such as LTTL may pose a practical compromise for the task at hand in many real-world application contexts. In future work, we plan to exploit the richness of Apertium RDF and other existing LLOD lexical resources in large-scale experiments on lexicon induction. Our goal is to further enhance bilingual dictionaries via multiple pivot languages, with the potential goal of bringing the performance of LTTL even closer to cross-lingual transformers.

Acknowledgments

This work was funded by the Prêt-à-LLOD project within the European Union's Horizon 2020 research and innovation programme under grant agreement no. 825182.

References

[1] Declerck T, McCrae JP, Hartung M, Gracia J, Chiarcos C, Montiel-Ponsoda E, et al. Recent Developments for the Linguistic Linked Open Data Infrastructure. In: Proc. of LREC; 2020. p. 5660-7.

[2] McDonald L, Malcolm B, Ramagopalan S, Syrad H. Real-world Data and the Patient Perspective: the PROmise of Social Media? BMC Medicine. 2019;17.

[3] Bullinger M, Quitmann J. Quality of life as patient-reported outcomes: principles of assessment. Dialogues in Clinical Neuroscience. 2014 Jun;16(2):137-45.
[4] World Health Organization. WHOQOL: Measuring Quality of Life; 1997. Available from: https://apps.who.int/iris/handle/10665/63482.
[5] Søgaard A, Vulić I, Ruder S, Faruqui M. Cross-Lingual Word Embeddings. Morgan & Claypool; 2019.
[6] Yarowsky D, Ngai G, Wicentowski R. Inducing Multilingual Text Analysis Tools via Robust Projection across Aligned Corpora. In: Proc. of HLT; 2001. .
[7] Barnes J, Klinger R, Schulte im Walde S. Bilingual Sentiment Embeddings: Joint Projection of Sentiment Across Languages. In: Proc. of ACL; 2018. p. 2483-93.
[8] Gracia J, Fäth C, Hartung M, Ionov M, Bosque-Gil J, Veríssimo S, et al. Leveraging Linguistic Linked Data for Cross-Lingual Model Transfer in the Pharmaceutical Domain. In: The Semantic Web – ISWC 2020. Cham: Springer; 2020. .
[9] Conneau A, Lample G, Ranzato M, Denoyer L, Jégou H. Word Translation Without Parallel Data. arXiv:171004087 [cs]. 2017. Available from: http://arxiv.org/abs/1710.04087.
[10] Chen X, Cardie C. Unsupervised Multilingual Word Embeddings. In: Proc. of EMNLP; 2018. p. 261-70.
[11] Ruder S, Søgaard A, Vulić I. Unsupervised Cross-Lingual Representation Learning. In: Proc. of ACL; 2019. p. 31-8.
[12] Feng Y, Wan X. Learning Bilingual Sentiment-Specific Word Embeddings without Cross-lingual Supervision. In: Proc. of NAACL:HLT; 2019. p. 420-9.
[13] Vulić I, Glavaš G, Reichart R, Korhonen A. Do We Really Need Fully Unsupervised Cross-Lingual Embeddings? In: Proc. of EMNLP-IJCNLP; 2019. p. 4407-18.
[14] Artetxe M, Ruder S, Yogatama D, Labaka G, Agirre E. A Call for More Rigor in Unsupervised Cross-lingual Learning. In: Proc. of ACL; 2020. p. 7375-88.
[15] Glavaš G, Litschko R, Ruder S, Vulić I. How to (Properly) Evaluate Cross-Lingual Word Embeddings: On Strong Baselines, Comparative Analyses, and Some Misconceptions. In: Proc. of ACL; 2019. p. 710-21.
[16] Karthikeyan K, Wang Z, Mayhew S, Roth D. Cross-Lingual Ability of Multilingual BERT: An Empirical Study; 2020. Available from: https://arxiv.org/abs/1912.07840.
[17] Conneau A, Khandelwal K, Goyal N, Chaudhary V, Wenzek G, Guzmán F, et al. Unsupervised Cross-lingual Representation Learning at Scale. In: Proc. of ACL; 2020. p. 8440-51.
[18] Strubell E, Ganesh A, McCallum A. Energy and Policy Considerations for Deep Learning in NLP. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics; 2019. p. 3645-50.
[19] Tanaka K, Umemura K. Construction of a bilingual dictionary intermediated by a third language. In: Proc. of COLING; 1994. p. 297-303.
[20] Villegas M, Melero M, Bel N, Gracia J. Leveraging RDF Graphs for Crossing Multiple Bilingual Dictionaries. In: Proc. of LREC; 2016. p. 868-76.
[21] Gracia J, Kabashi B, Kernerman I, editors. Proceedings of TIAD-2019 Shared Task – Translation Inference Across Dictionaries. vol. 2493. CEUR-WS.org; 2019.
[22] Forcada ML, Ginestí-Rosell M, Nordfalk J, O'Regan J, Ortiz-Rojas S, Pérez-Ortiz JA, et al. Apertium: a free/open-source platform for rule-based machine translation. Machine Translation. 2011;25(2):127-44.
[23] Fäth C, Chiarcos C, Ebbrecht B, Ionov M. Fintan - Flexible, Integrated Transformation and Annotation eNgineering. In: Proc. of LREC; 2020. p. 7212-21.
[24] Marimon M, Krallinger M. MeSpEn_Glossaries. Zenodo; 2018.
[25] Gracia J, Villegas M, Gómez-Pérez A, Bel N. The Apertium Bilingual Dictionaries on the Web of Data. Semantic Web. 2018;9(2):231-40.
[26] Villegas M, Intxaurrondo A, Gonzalez-Agirre A, Marimón M, Krallinger M. The MeSpEN Resource for English-Spanish Medical Machine Translation and Terminologies: Census of Parallel Corpora, Glossaries and Term Translations. In: Workshop on Multilingual Biomedical Text Processing; 2018. .