Validating Ontology-based Annotations of Biomedical Resources using Zero-shot Learning

Dimitrios A. Koutsomitropoulos
Computer Engineering and Informatics Dpt.
University of Patras
26500 Patras, Greece
koutsomi@ceid.upatras.gr

ABSTRACT

Authoritative thesauri in the form of web ontologies offer a sound representation of domain knowledge and can act as a reference point for automated semantic tagging. On the other hand, current language models achieve to capture contextualized semantics of text corpora and can be leveraged towards this goal. We present an approach for injecting subject annotations using query term expansion against such ontologies in the biomedical domain. For the user to have an indication of the usefulness of these suggestions we further propose an online method for validating the quality of annotations using NLI models such as BART and XLM-R. To circumvent training barriers posed by very large label sets and scarcity of data we rely on zero-shot classification and show that semantic matching can contribute above-average thematic annotations. Also, a web-based validation service can be attractive for human curators vs. the overhead of pretraining large, domain-tailored classification models.

KEYWORDS

Thesaurus; semantic matching; biomedical indexing; classification; MeSH; language models; machine learning.

1 Introduction

To promote the discovery and exploitation of open learning and research materials, users often rely on expert tagging and classification. Open Educational Resources (OERs) are a form of learning objects that are openly available and can be effectively reused and consolidated to support learning and research needs [7]. However, assigning subject annotations by experts is a costly task that is both time-consuming and error prone. The manual effort often required in this process makes automated thematic indexing and annotation methods appealing. In addition, resources' metadata can be sought for at and harvested from various content providers that frequently follow their own proprietary annotation schemes or disparate metadata standards.

To aid the discovery and synthesis of such resources the notion of a Learning Object Ontology Repository (LOOR) has been earlier proposed by the authors [15], which enables federated search over repositories and combines the semantics of results' metadata into a common learning object (LO) ontology. Next, we have shown that one can reuse the initial federated search keywords to construct and provide appropriate subject annotations for selected OERs in an automated manner, thus supporting the effort of content curators and instructors for careful and elaborate content selection [13]. These keywords are reused to discover matching terms and then expanded within authoritative, domain knowledge thesauri, expressed in the form of OWL ontologies and implemented in SKOS [19]. As a result, semantic interoperability is accommodated, and content retrieval is improved by boosting recall. Still, other than human intervention, there is no means to assess the appropriateness of discovered subject annotations and their quality is associated with the search indexing capabilities of the source providers.

Building on these premises, in this paper we propose a method for validating subject terms by tapping into deep language models based on word embeddings, including BART [17] and XLM-RoBERTa (XLM-R) [4]. Since multiple terms could be proposed for each item, the task addressed is one of multi-label classification. While item classification requires provision of the entire label set and is generally a hard task even for large models with millions of parameters[2], [16], in our work we are interested in acquiring validation scores for terms already suggested by the semantic matching process, rather than classifying from scratch. We show that it is possible to achieve this even with very low resources and in the absence of available training data by employing the idea of zero-shot classification [28]. We conduct experiments with a representative result set of biomedical OER metadata and Medical Subject Headings (MeSH) suggestions [24] and assess an average score for semantic matching subject annotations of over 70%. Moreover, the proposed validation process can be streamlined by implementing a low latency, RESTful communication service with the models' inference results through openly available APIs. As such, it can appeal to instructors by equipping them with a degree of confidence about the proposed terms amidst their task of content selection. Our prototype implementation and sources are available on GitHub: https://github.com/swigroup/federated-semantic-search.

To our knowledge, this is the first time a concrete, web-based procedure about zero-shot validation of subject annotations is proposed, with such annotations coming from the reuse of search query terms combined with SKOS-based matching and expansion.

In the following, we examine some related work involving approaches for semantic annotation and classification of learning material in repositories as well as Natural Language Inference

(NLI) machine-learning models and their use for zero-shot, multi-label classification. In Section 3 we outline the workings of the LOOR aggregation mechanism, briefly describing the main characteristics of the LO ontology and the thesauri term matching and expansion process. Section 4 details the method for zero-shot validation of thematic suggestions and presents the algorithm for assigning subject terms and computing their scores. Next, in Section 5, we carry out the validation experiments and compare scoring performance against pre-labeled datasets (ground truth). Finally, the last section summarizes our conclusions and gives pointers for future work.

2 Related Work

Efficient and effective classification of OERs is still an open problem that is often acknowledged in the literature. Other than the difficulty of producing accurate results in an automated manner, there are discrepancies in manual annotations that can be free-text or otherwise form unstructured pieces of metadata. Such discrepancies could be addressed by exploiting existing thematic vocabularies. For example, in [9] the authors try to further extend OERs discoverability and sharing by linking OERs descriptions with well-defined terms inside various established thesauri. As a result, data from several OER repositories can be integrated, exposed and finally enriched. Another approach for assigning and linking thematic entities and other types of resources to OERs is presented in [21]. This work describes how to transform IEEE LOM metadata into XML representation and RDF triples and finally link them to other datasets e.g., DBpedia.

OER retrieval can benefit from term and query expansion, more so when it leverages expert knowledge in the form of controlled vocabularies and thesauri. SKOS-based query expansion can achieve improved results in web search [11] or recommendation [26]. Expanded queries are also shown to allow retrieval of additional OERs that are relevant but would not be fetched otherwise [22]. The authors employ a non-SKOS ontology and address a specific repository, thus distancing from federated search.

Regarding classification in the biomedical domain using machine learning, recent advances include FullMeSH [6], which considers the full text of a publication and achieves and F-score of 0.67 and BertMeSH [29] that employs a deep language model (BERT [8]) and yields an F-score of 0.69. Both systems require extensive pretraining with existing labeled data that lasts for a few days (4-5). Another approach, closer to zero-shot learning, is investigated in [14], where out-of-the-box, already pretrained models are used to estimate semantic similarity between abstract and subject terms.

Zero-shot text classification aims to assign an appropriate label to a piece of text, irrespective of the text domain and the aspect described by the label [3]. Effectively, it attempts to classify text even in the absence of labeled data. NLI is a form of textual entailment that, given two sequences, determines if the one entails the other [27]. The idea of using general-language models finetuned on natural language inference tasks to achieve zero-shot classification is proposed in [28] and is further discussed in Section 4.

2 Federated Search and Term Expansion

3.1 Metadata Harvesting and Alignment

The overall system architecture of the federated search, semantic matching and term validation process is shown in Fig. 1. First, a federated query is initiated towards the various repositories. Data sources include *MERLOT II*, a large archive of OERs [18], *Europe PubMed Central*, a major repository of biomedical literature [10], *ARIADNE finder*, a European infrastructure for accessing and sharing learning resources [23] and *openarchives.gr*, the entry point for Greek scholarly content.

Next, OER metadata returned as responses to the query are harvested and aligned to a unified LO Ontology Schema. Despite the potentially disparate schemata that could hamper integration, it is possible to identify a least common set of elements and use well known educational metadata standards such as LOM, as mediators. This set can form the basis for a common schema to be used for direct ingestion of learning objects into the LOOR. This *LO Ontology Schema* contains entities representing elements of the IEEE LOM schema and combines terminology with the Dublin Core metadata terms specification. A detailed account of the design principles, development and description of this ontology can be found in [15]. The *lom:keyword* property is used to express the subject of an OER. In our LO ontology profile, it is implemented as an object- rather than a datatype- property, which would allow for literal-only values. This would allow to associate the subject of an OER to SKOS ontology concepts, thus increasing the value of

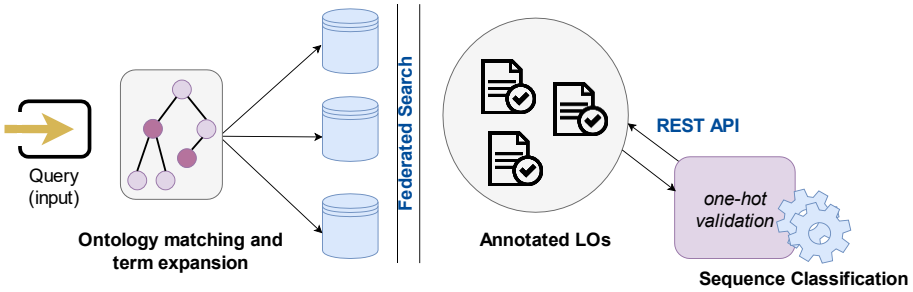

Figure 1: System architecture

our LO ontology when used in the context of knowledge discovery applications.

In addition, we replicate a least common subset of metadata elements and map them to the ontology. This includes *lom:title* (a literal), *lom:identifier* (anyURI) and *lom:description* (a literal) corresponding to the title, URL and description (e.g. the abstract) of the source item. To accommodate for validity scores of subject terms, another addition is the *sm:score* annotation property, with *sm* prefix denoting our custom namespace. The domain of this property is the set of lom:keyword assertions and its range is xsd:double. Consequently, it serves to reify the binary lom:keyword relation.

A path to the rest of the original OER's metadata is also maintained. These are easy to retrieve directly from their sources, using their unique URL or to harvest them through an OAI service provider. Further, a curator or instructor can manually review automatically assigned values, edit the rest of the fields of the unified schema and opt for the addition of the item into the LOOR.

3.2 Term Matching and Expansion

In earlier work we have investigated and documented the positive effects of query expansion when harvesting OERs and for subject classification for example, boosting recall of retrieval by a factor of 4-8 [13]. In essence, keywords that initiate harvesting are matched against expert terminological knowledge expressed in the form of SKOS thesauri. Each keyword can then be expanded into several narrower keywords which refine the initial query. This expansion is achieved by performing reasoning about the thesaurus hierarchy, taking advantage of the semantic relationships between matching terms. For example, to discover the refinements of a concept, we can expand on the transitive closure of the SKOS properties *skos:broader / skos:narrower*. A single thesaurus concept may also contain multiple lexical representations, including alternative labels and translations in different languages, as represented by the properties *skos:prefLabel* and *skos:altLabel*.

Reference thesauri used in the LOOR and implemented in SKOS include: the *Thesaurus of Greek Terms* (TGT) a bilingual (Greek, English) controlled vocabulary published by the National Documentation Center in Greece [20]; two thematic micro-thesauri extracted out of TGT for the fields of mathematics (*Maths Thesaurus*) and medicine (*Medicine Thesaurus*); and *Medical Subject Headings* (MeSH), a controlled vocabulary of biomedical terms in a hierarchical structure that is produced and maintained by the United States National Library of Medicine (NLM). MeSH consists of 29,917 concepts (called *Descriptors*) as of 2021 [24].

The information produced by the exploration of the thesauri term hierarchy is maintained and reused to provide thematic annotations for an item when it is considered for addition into the LOOR. As a result, original search keywords are also being used as seeds to generate appropriate subject annotations for selected OERs. Merely suppling such arbitrary keywords as subjects would prohibit further reuse, discovery and interoperability. Therefore, these keywords are first matched and refined against formal thematic thesauri and the matches are injected as semantic subject annotations into the selected OERs, using the lom:keyword property of the LO Ontology Schema. The uppermost parent concept that led to the currently matched term is also used to insert another subject annotation. This is reasonable since a broader concept is also a valid thematic subject for the resource in a classification hierarchy.

As an example, consider the seed keyword medicine. This is matched in the MeSH thesaurus by the concept with ID D008511 and refined, for example, by the concept with ID D009462 (neurology). Items, whose search keywords fetching them match one or more of the various labels of D009462 will be automatically indexed with the concept D009462. In addition, they will also get the concept D008511 as a subject annotation since this is the topmost parent matching term for the initial keyword.

4 Validation of Thematic Annotations

4.1 Computing scores and model communication

After the thematic subject assertions have been inserted into the LOM Ontology, as a result of keyword matching and expansion (*semantic matching*), it is time to validate these assertions and compute their scores. This is achieved by leveraging the Huggingface inference API [12]. This API offers an HTTP endpoint and allows to invoke any model available on the model hub. The free tier of the API allows to run inference tasks on CPU only.

An HTTP request body is constructed in JSON format, by feeding the item title and abstract as well as the lexical representations of the injected subject terms. An important parameter is *multi_label* which we set to *true*, because an item can get multiple thematic annotations: each term can have an independent score in 0, 1, rather than all summing up to one. Next, this request is sent using HTTP POST method through the API, towards the selected model for inference. Transparent to the user, the response, also in JSON format, is parsed and the scores are assigned to the item's subject terms (lom:keyword assertions) through the sm:score annotation property. Figure 2 shows the expected outcome of the semantic matching and term validation process.

A SKOS concept acting as a filler to the lom:keyword property can have multiple lexical representations (multiple languages, alternative labels and so on). So, the question may arise as to what label is to be fed to the model to produce the scoring inference. We

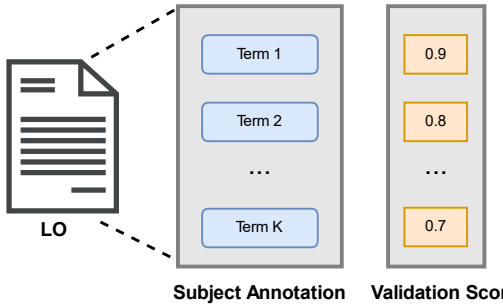

Figure 2. Assigning validation scores to term assertions

have chosen to take advantage of the skos:prefLabel property which denotes the preferred lexical representation for a vocabulary term. Still, by definition, a term can have many skos:prefLabel assertions, each for a different language tag [19]. Therefore, we find the preferred label which maximizes the score and use this value to assign an overall score for the term assertion (sm:score property). For cross-lingual models, this might not even be necessary.

Note also that it is desirable but not necessary for a keyword to match a thesaurus concept. In case there is no match, the keyword is still retained into the LOs metadata by keeping it as an additional lom:keyword. This is helpful for queries that may rely solely on text search or that are SKOS oblivious and is a useful fallback measure.

The following algorithm in Table 1 summarizes the scoring approach. Given a BFS traversing of a thesauri tree, it is easily seen that the complexity of the algorithm is polynomial to the number of axioms in the thesaurus ontology.

Table 1: Term validation and score assignment algorithm.

k, ℓ: keywords, I_k: result set for keyword k, $A(i)$: set of subject assertions for item $i \in I_k$ K: the *lom:keyword* property, $\mathcal{C}_k$, $\mathcal{D}_k$: sets of concepts
$\mathcal{C}_k \leftarrow$ **find_matching_concepts** (k) $\forall\, c \in \mathcal{C}_k$: $\mathcal{D}_k \leftarrow$ **find_narrower_concepts** (c) // used for term expansion (Section 3) $\forall\, i \in I_k$: $A(i) \leftarrow K\,(i, k)$ $\forall\, c \in \mathcal{C}_k$: //assign subject terms for i $A(i) \leftarrow A(i) \cup \{K\,(i, c)\}$ $A(i) \leftarrow A(i) \setminus \{K(i, k)\}$: $<c, k> \in$ skos:prefLabel //already there **If** $c \in \mathcal{D}_\ell$: // $\ell \neq k$ another keyword visited previously and expanded $A(i) \leftarrow A(i) \cup \{K\,(i, d)\}$: $d \equiv \mathcal{C}_\ell$ 1 $\forall\, a \in A(i)$: $s = \max_\ell$ (**compute_score** (ℓ)): $<i, a> \in K, <a, \ell> \in$ skos:prefLabel **assert** $< K\,(i, a), s> \in$ sm:score

4.2 Tasks and models for zero-shot validation

The main premise behind zero-shot classification is the repurposing of the NLI task. Natural Language Inference is the task of determining whether a "hypothesis" is true (entailment), false (contradiction), or undetermined (neutral) given a "premise" [28]. Not all models available on the Hugginface model hub support this task. For a model to support this task it has to be specifically trained or finetuned on an NLI dataset, such as the Stanford Natural Language Inference corpus (SNLI) [1] or the Multi-Genre Natural Language Inference corpus (MNLI) [27]. These corpora are in English only. It is shown that cross-lingual capabilities can be achieved when employing multilingual datasets, such as the Cross-lingual Natural Language Inference corpus (XNLI), which is an extension of the SNLI/MNLI corpus in 15 languages, including English and Greek [5].

To use NLI for classification, we need to map the hypothesis/premise pairs in such a way that would accommodate the classification task. Both hypothesis and premise can be arbitrary texts. Therefore, the premise can be mapped to an item's title and abstract; in turn, the hypothesis can be mapped to each classification label, one at a time, prepended by an appropriate template e.g. "this text is about ?" or "this example is ?" where ? denotes the label. Thus, classification can be reduced to as many NLI tasks as there are labels. In case of large label sets, such as MeSH, this would mean to supply all terms and initiate this many NLI tasks, which would be impractical from a performance point of view. For only a few labels however, such as the ones proposed by semantic matching, zero-shot classification performs well (see next section) and is the basis of our term validation approach. To compute the score for a label, the inference API considers the score of the 'entailment' and 'contradiction' outcomes so that they add up to 1 (softmax) and outputs the value for 'entailment'.

Out of the models available on the hub that implement zero-shot inference we have chosen *bart-large-mnli*, which is a version of the pretrained BART autoencoder [17] that has been specifically finetuned on MNLI and thus supports NLI tasks; and *xlm-roberta-large-xnli*, which takes the pretrained XLM-RoBERTa (XLM-R) model [4] and finetunes it on XNLI, thus achieving, among others, multilingual capabilities.

It is worth noting that neither of these models have been specifically trained or finetuned on biomedical or other domain data and are rather general-domain language models. They can also be used out-of-the-box, without the need of additional training or finetuning, which makes them appealing for online inference or low-resource requirements. On the other hand, some domains, such as biomedicine, are characterized by abundant amounts of unlabeled text [16] and using them for training from scratch can offer improvements over finetuning general models. Even so, large language models, such as GPT-3 can perform competitively on downstream tasks with far less domain-specific data than would be required by smaller models [2], a fact that further supports the intuition behind our zero-shot validation approach.

5 Experiments and Results

5.1 Dataset for subject term validation

For evaluating our term validation approach, we focus on the biomedical domain. We consider a series of result sets fetched by our federated search and term expansion mechanism (Section 3.1) for 40 different search keywords. These keywords are matched against the MeSH thesaurus and the matches provide subject annotations for items, corresponding to MeSH terms (Section 3.2). These are to be validated by the zero-shot approach and get their scores. The keywords have been chosen to be evenly distributed among high-level thesauri terms (lots of children in the hierarchy), mid-level terms (fewer children in the hierarchy), bottom-level terms (no children) and keywords having no matches whatsoever in the thesaurus used.

Since federated search spans over various repositories, search results may include items in different languages namely, English and Greek. Even though MeSH is not multilingual, some items with Greek titles and abstracts can still get English-only MeSH annotations, because search looks within their full metadata which may include translations. A sample record with its subject suggestions is shown in Figure 3 in OWL format. The dataset contains a total of 2,123 such items and 3,541 MeSH term lom:keyword property assertions (

Table **2**). On average, an item is classified under 1.7 MeSH headings over a total of 56 unique terms.

To evaluate the significance of the scores we compare them with the scores produced for an already labeled dataset annotated by experts (ground truth). This dataset comprises of biomedical citation records and their respective MeSH headings from the PubMed database [25]. It has been constructed to have similar properties with the first i.e., the number of items and unique MeSH term occurrences is virtually the same (

Table **2**).

Table 2. Test datasets for term validation and their properties.

Dataset	Total items	Total MeSH assertions	Unique terms	Average terms per item
Semantic matching dataset	2,123	3,541	56	1.7
PubMed dataset	2,000	2,473	50	1.2

5.2 Validation scores and performance

In the following, we present the results of term validation experiments for expert vs. automatic annotations. These experiments would help answer mainly two questions: *a)* how well semantic matching performs vs. expert suggestions and *b)* what is the score distribution of terms assigned by experts vs. terms assigned by semantic matching.

First, we show results with the bart-large-mnli model and then for xlm-roberta-large-xnli which is cross-lingual. Table 3 summarizes the mean scores, their standard deviation, and average response time for each dataset and each model, respectively.

Table 3. Performance of term validation for test datasets and zero-shot models.

Dataset	model	mean	std	response time (ms)	response time *use_cache* (ms)
Semantic Matching PubMed dataset	bart-large-mnli	0.57	0.32	1764	166
		0.62	0.34	1885	160
Semantic Matching PubMed dataset	xlm-roberta-large-xnli	0.71	0.27	1321	162
		0.66	0.31	1640	160

To produce scores, the models require a forward pass for the text (title and abstract) as well as additional passes, one for each label. Observed latency remains below *2 sec* for each item, which indicates that term suggestions for several items can be validated within user-acceptable timeframes. This can get a considerable speed-up once inferences are allowed to run on GPU. In addition, when the *use_cache* option of the API is enabled (default), responses are quite rapid. This can be of benefit for multiple users exploring adjacent knowledge domains and is possible due to the deterministic operation of the models.

bart-large-mnli

Fig. 4 shows the distribution of scores for the two datasets using BART. We notice that semantic matching is capable of suggesting terms with an average score of 0.57. This can be considered quite high, once compared with the 0.62 of the experts' terms.

```
<owl:NamedIndividual rdf:ID="N66069">
<lom:keyword xml:lang="en">Gene Proteins</lom:keyword>
<lom:keyword rdf:resource="http://www.nlm.nih.gov/mesh/2006#D000602"/>
<lom:keyword rdf:resource="http://www.nlm.nih.gov/mesh/2006#D011506"/>
<lom:identifier rdf:datatype="xsd:anyURI">http://europepmc.org/...</lom:identifier>
<lom:title>Trypsin inhibitors: promising candidate satietogenic...</lom:title>
<lom:description>The increase in non-communicable chronic disea...</lom:description>
</owl:NamedIndividual>
```

Figure 3: Sample record with MeSH terms assigned by semantic matching

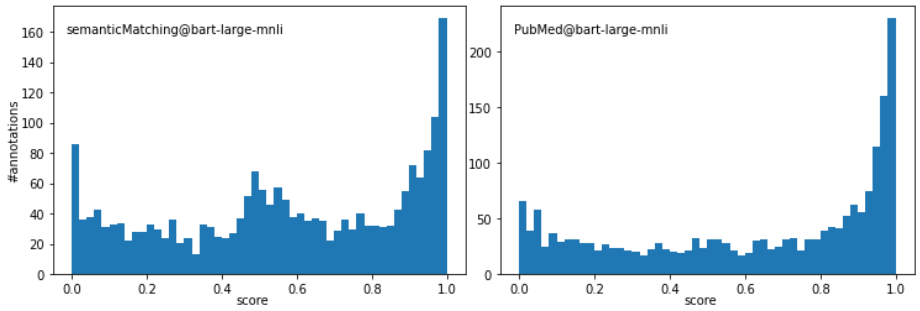

Fig. 4. Term validation scores for the two datasets with bart-large-mnli

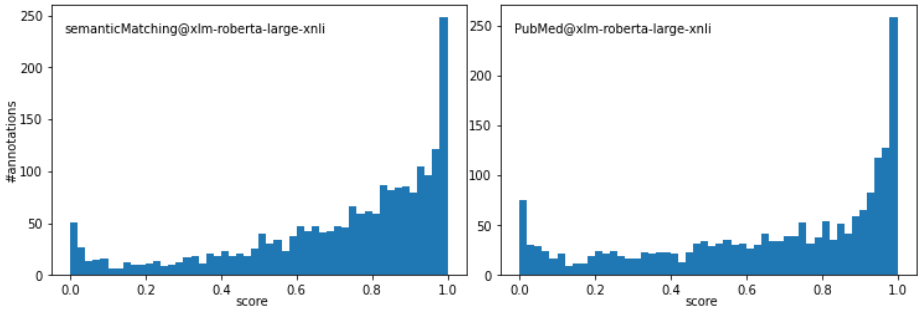

Fig. 5. Term validation scores for the two datasets with xlm-roberta-large-xnli

In addition, almost half of the term assertions (49.7%) appear to pass the 0.62 threshold. The two distributions appear similar, with a growing number of assertions having larger scores (up to 1). This is intuitively satisfying, since scores do not appear to be uniformly or normally distributed and thus results from zero-shot models are far from random. Instead, both distributions exhibit a "long-tail" property. However, semantic matching suggestions have some concentration around the mean which may be due to some Greek texts slipping into results, but still getting MeSH terms. These are treated neutrally by the mono-lingual, zero-shot model (neither contradict, nor entail). Also note that in both cases, there are some terms the model cannot properly infer (scores close to zero) possibly due to their specificity and lack of training.

xlm-roberta-large-xnli

Fig. 5 shows the distribution of scores for the two datasets using XLM-R. There is now an increased average score for semantic matching of 0.71. The quality of suggestions appears to surpass that of expert annotations by a small mark (0.05). Distributions keep the same properties as before, but now there is no 'bump' towards the middle for semantic matching, possibly because now the model recognizes Greek texts and classifies them correctly. Indeed, XLM-R has been trained and finetuned on cross-lingual datasets, including English and Greek. Therefore, it is capable of capturing the semantics of texts in both and infer about them more accurately.

There is an improvement also on the PubMed dataset which, however, does not contain texts other than English. This is consistent with the performance of the models employed: In [17] it is shown that BERT outperforms BART on the MNLI task. Next in [4] it is shown that multilingual models can outperform (monolingual) BERT even, and surprisingly, for single-language tasks. This is due to the capacity of multilingual models to leverage training data coming from multiple languages for a particular task. Thus, it is expected for xlm-roberta-large-xnli to perform generally better than bart-large-mnli.

As a side note, as far as MeSH classification is concerned, validation scores of semantic matching are state-of-the-art (0.71 vs. 0.69 of BertMeSH). However, these are not directly comparable because BertMeSH considers the entire 29K label set for making assignments from scratch. In this paper, we deal with validation of existing assignments only.

6 Conclusions and Future Work

Zero-shot inference can be leveraged to validate an OER's thematic annotations, be they either human-supplied or produced by automated techniques. The low-cost and hassle-free invocation of the models makes the use of inference APIs desirable. Even for end-to-end classification involving all available labels, the zero-shot approach can be of value, but may require deployment on premises and accelerated hardware. Still, resource consumption remains low since training from scratch can be avoided.

The semantic matching approach can bootstrap the task of thematic content tagging by tapping into domain-knowledge thesauri and is shown to produce quality classification results.

Then, content curators, instructors and researchers can take benefit of our validation service to verify the appropriateness of proposed terms and decide their addition into the LOOR for further reuse.

As a next step, we intend to investigate further the effect that fine-tuning a zero-shot model may have on validation scores. By finetuning an NLI model to the problem domain for example, a set of biomedical corpora we can expect an even greater accuracy in the validation task. The very semantics of the ontology used, such as the relationships between terms and their hierarchy, could be involved in this process.

REFERENCES

[1] Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In Proc. of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015), pp. 632-642

[2] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165.

[3] Chang, M. W., Ratinov, L. A., Roth, D., & Srikumar, V. (2008). Importance of Semantic Representation: Dataless Classification. In Aaai (Vol. 2, pp. 830-835).

[4] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L. and Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. In Proc. of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 8440–8451.

[5] Conneau, A., Lample, G., Rinott, R., Williams, A., Bowman, S. R., Schwenk, H., & Stoyanov, V. (2018). XNLI: Evaluating cross-lingual sentence representations. In Proc. of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018), pp. 2475–2485.

[6] Dai S, You R, Lu Z, Huang X, Mamitsuka H, Zhu S (2020) FullMeSH: improving large-scale MeSH indexing with full text. Bioinformatics (Oxford, England), 36(5), 1533–1541. https://doi.org/10.1093/bioinformatics/btz756

[7] Davis, E., Cochran, D., Fagerheim, B., & Thoms, B. (2016) Enhancing Teaching and Learning: Libraries and Open Educational Resources in the Classroom. Public Services Quarterly, 12(1), 22-35.

[8] Devlin, J, Chang, M. W., Lee, K., Toutanova, K. (2019) BERT: Pre-training of Deep Bidirectional Transformers for Language Un-derstanding, In Proc. of NAACL-HLT 2019, pp. 4171–4186.

[9] Dietze, S., Yu, H. Q., Giordano, D., Kaldoudi, E., Dovrolis, N. & Taibi, D. (2012). Linked education: interlinking educational resources and the web of data. In: The 27th ACM Symposium On Applied Computing (SAC-2012), Special Track on Semantic Web and Applications.

[10] Europe PMC Consortium. (2017) Europe PMC: A Full-Text Literature Database for the Life Sciences and Platform for Innovation. Nucleic Acids Research 43. Database issue (2015): D1042–D1048.

[11] Haslhofer, B., Martins, F., & Magalhães, J. (2013). Using SKOS vocabularies for improving web search. In Proceedings of the 22nd international conference on World Wide Web companion (pp. 1253-1258). International World Wide Web Conferences Steering Committee.

[12] Huggingface (2021). Accelerated Inference API (online). Available : https://api-inference.huggingface.co/docs/python/html/index.html

[13] Koutsomitropoulos, D. (2019) Semantic annotation and harvesting of federated scholarly data using ontologies. Digital Library Perspectives 35 (3–4), 157–171 (2019)

[14] Koutsomitropoulos, D. A., and Andriopoulos, A. (2021). Thesaurus-based Word Embeddings for Automated Biomedical Literature Classification. Neural Computing and Applications, in press.

[15] Koutsomitropoulos, D. A., and Solomou, G. D. (2018). A learning object ontology repository to support annotation and discovery of educational resources using semantic thesauri. IFLA journal 44 (1), 4-22.

[16] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4), 1234-1240.

[17] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V. and Zettlemoyer, L. (2019). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461.

[18] McMartin, F. (2006) MERLOT: a model for user involvement in digital library design and implementation. Journal of Digital Information, 5 (3).

[19] Miles, A., and Bechhofer, S., eds. (2009) SKOS Simple Knowledge Organization System Reference. W3C Recommendation. Available: http://www.w3.org/TR/skos-reference

[20] National Documentation Center. (2021) Thesaurus of Greek Terms. Available: http://general-terms.thesaurus.ekt.gr/vocab/index.php

[21] Rajabi, E., Alonso, S.S., & Sicilia, M. (2015). Interlinking educational resources to Web of Data through IEEE LOM. Computer Science and Information Systems, 12(1), 233–255.

[22] Segura, N. A., García-Barriocanal, E., & Prieto, M. (2011). An empirical analysis of ontology-based query expansion for learning resource searches using MERLOT and the Gene ontology. Knowledge-Based Systems, 24(1), 119-133.

[23] Ternier, S., Verbert, K., Parra, G., Vandeputte, B., Klerkx, J., Duval, E., et al. (2009). The ariadne infrastructure for managing and storing metadata. IEEE Internet Computing, 13(4).

[24] U.S. National Library of Medicine. Medical Subject Headings, 2021. Online. Available: https://www.nlm.nih.gov/mesh/meshhome.html

[25] U.S. National Library of Medicine. PubMed.gov Online. https://www.nlm.nih.gov/databases/download/pubmed_medline.html

[26] Wenige, L., Berger, G., & Ruhland, J. (2018). SKOS-based concept expansion for LOD-enabled recommender systems. In Proc. of the 12th International Conference on Metadata and Semantics Research (MTSR 2018), pp. 101-112, Springer.

[27] Williams, A., Nangia, N., & Bowman, S. R. (2017). A broad-coverage challenge corpus for sentence understanding through inference. In Proc. of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics (ACL): Human Language Technologies, Volume 1, pp. 1112-1122

[28] Yin, W., Hay, J., & Roth, D. (2019). Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In Proc. of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP 2019), pp. 3914-3923.

[29] You, R., Liu, Y., Mamitsuka, H. and Zhu, S.(2020) BERTMeSH: Deep Contextual Representation Learning for Large-scale High-performance MeSH Indexing with Full Text. Bioinformatics, 2020. https://doi.org/10.1093/bioinformatics/btaa837