Leveraging Linguistic Linked Data for Cross-lingual Model Transfer in the Pharmaceutical Domain

Jorge Gracia[1], Christian Fäth[2], Matthias Hartung[3], Max Ionov[2], Julia Bosque-Gil[1], Susana Veríssimo[3], Christian Chiarcos[2], and Matthias Orlikowski[3]

[1] Aragon Institute of Engineering Research, University of Zaragoza, Spain
{jogracia,jbosque}@unizar.es
[2] Goethe University Frankfurt, Germany
faeth@em.uni-frankfurt.de, ionov@cs.uni-frankfurt.de,
chiarcos@informatik.uni-frankfurt.de
[3] Semalytix GmbH, Bielefeld, Germany
{hartung,susana.verissimo,matthias.orlikowski}@semalytix.com

Abstract. We describe the use of linguistic linked data to support a cross-lingual transfer framework for sentiment analysis in the pharmaceutical domain. The proposed system dynamically gathers translations from the Linked Open Data (LOD) cloud, particularly from Apertium RDF, in order to project a deep learning-based sentiment classifier from one language to another, thus enabling scalability and avoiding the need of model re-training when transferred across languages. We describe the whole pipeline traversed by the multilingual data, from their conversion into RDF based on a new dynamic and flexible transformation framework, through their linking and publication as linked data, and finally their exploitation in the particular use case. Based on experiments on projecting a sentiment classifier from English to Spanish, we demonstrate how linked data techniques are able to enhance the multilingual capabilities of a deep learning-based approach in a dynamic and scalable way, in a real application scenario from the pharmaceutical domain.

Keywords: Apertium RDF · cross-lingual model transfer · Fintan

1 Introduction

One of the biggest challenges faced by international companies in Europe and worldwide is that markets are spread across countries and languages. Thus, their ability to adapt to new markets is of vital importance. To that end, language technologies (LT) and linked data (LD) have been recognised as core technologies to reduce language barriers between different national markets [16].

A major challenge faced by suppliers of LT services and products in global markets arises from the complexity of business use cases, technical components needed to address them, and input data that comes from multiple languages.

Approaching this challenge by attempting to build dedicated Natural Language Processing (NLP) stacks for each new language from scratch is not scalable, due to generally high on-boarding costs for initial model development and refinement.

As an alternative, *cross-lingual model transfer* methods are based on the idea that NLP models readily existing for a source language can be transferred to a new target language of interest without language-specific supervision in terms of manually created training data being required in this target language [17]. As a primary source of cross-lingual information, many transfer approaches rely on bilingual lexical resources in order to bridge the language gap.

A growing number of lexical resources is made publicly available as part of the Linguistic Linked Open Data (LLOD) cloud[4] [4]. In this paper, we demonstrate the strong potential of LLOD resources to be used as catalysers of cross-lingual transfer of NLP models in deep learning frameworks. This is illustrated by way of the multilingual lexical resource Apertium RDF v2.0 that has been created and published as LLOD in order to meet the requirements of an LT-based real-world evidence platform for the pharmaceutical industry in a software-as-a-service setting. The specific use case aims at rapidly and cost-effectively increasing the multilingual capabilities of the NLP components underlying the platform, which we demonstrate here for the case of a pharma-specific sentiment detection model that is transferred from English to Spanish. In order to allow for a flexible and automated way of transforming the Apertium data into the LLOD formats, we rely on Fintan [8], a newly developed RDF transformation platform.

The remainder of this paper is organised as follows. In Section 2 we describe the background and technological context of this research. In Section 3 we give an overview of the overall architecture. Then, Section 4 describes the transformation and linking steps carried out to convert the Apertium original data into RDF. In Section 5 the role of the Apertium RDF data to improve bilingual sentiment embeddings is explained, and Section 6 reports on some experimental validations. Finally, Section 7 contains conclusions and future work.

2 Background and Related Work

In this section we describe some core technologies needed to better understand our approach, namely the Ontolex-lemon model and the Apertium initiative. We also report on the recent advancements in cross-lingual transfer learning.

2.1 OntoLex-lemon

In the context of LLOD, OntoLex-lemon[5] is the primary community standard for representing lexical data in RDF [13]. This was originally developed with the aim to provide a rich linguistic grounding for ontologies, meaning that the natural language expressions used in labels, definitions or comments of ontology elements are equipped with an extensive linguistic description.

[4] `http://linguistic-lod.org/llod-cloud`
[5] `https://www.w3.org/2016/05/ontolex/`

The main class for linguistic description in OntoLex is `LexicalEntry`, which corresponds to a word, a multi-word expression, or an affix. Lexical entries have different lexical forms (through the `Form` class) with their corresponding written and/or phonetic representations. The connection of a lexical entry to an ontological entity is marked mainly by the `denotes` property or is mediated by the `LexicalSense` or the `LexicalConcept` classes.[6]

Other modules extend the core module such as the variation and translation (*vartrans*) module, which introduces the representation of translations as a subtype of `SenseRelation`, i.e., a relation established between lexical senses.[7]

2.2 Apertium

Apertium[8] is a free/open-source machine translation platform [6] that mostly relies on the use of symbolic methods and currently includes over 50 language pairs.[9] It provides NLP components for many languages, as well as transfer rules and bilingual dictionaries for their respective translation.

A subset of the family of bilingual dictionaries developed in Apertium was converted to the LMF [7] ISO standard as part of the METANET4U Project.[10] From that subset of Apertium dictionaries, only the entries in Apertium which were annotated as nouns, proper nouns, verbs, adjectives and adverbs were considered (from a long list of heterogeneous parts of speech present across datasets). This LMF subset constituted the basis for the first RDF representation of the Apertium dictionaries [10], which was released as LLOD[11] (we will refer to it as `Apertium RDF v1.0` in the rest of this paper). Such an RDF version of the Apertium dictionary data was based on the *lemon* model, the predecessor of Ontolex-lemon, and its translation module [9].

Given that Apertium RDF v1.0 only covered the language pairs for which an LMF version was available, and that the initiative to convert Apertium dictionaries into LMF was not continued, we decided to expand Apertium RDF by accessing the Apertium source data directly and converting them into the more recent OntoLex version of the *lemon* model.

2.3 Cross-lingual Transfer Learning

Cross-lingual induction of resources for multilingual text analytics, instead of creating them from scratch, has attracted much attention in the NLP literature over the last decades, dating back at least to [19]. These early works are

[6] See `https://www.w3.org/2016/05/ontolex/#core` for a diagram and complete description of the OntoLex-lemon core module

[7] See the whole diagram of the vartrans module at `https://www.w3.org/2016/05/ontolex/#variation-translation-vartrans`

[8] `https://www.apertium.org/`

[9] `http://wiki.apertium.org/wiki/Main_Page`

[10] `http://www.meta-net.eu/projects/METANET4U/`

[11] `http://linguistic.linkeddata.es/resource/id/apertium`

comparatively resource-intensive themselves, as they assume the availability of parallel or aligned corpora, which is a requirement that is hard to meet for many language pairs, and even more so in technical domains.

In more recent work, these requirements are substantially alleviated by representation learning approaches capitalizing on bilingual word embeddings which can be induced from parallel and non-parallel corpora (cf. [17] for an overview). Due to their generality, bilingual embedding approaches are sufficiently versatile in order to be applied to a variety of cross-lingual text classification problems [15]. Cross-lingual sentiment analysis, as a special case, is investigated in multiple studies from a representation learning perspective [21,20,1].

UBiSE [5] presents a projection approach based on bilingual sentiment-specific word embeddings without any cross-lingual supervision, thus reducing resource requirements to a minimum: Only relying on a labeled sentiment corpus in the source language, as well as monolingual embeddings for both languages, their method outperforms Bilingual Sentiment Embeddings (BLSE) [1] on online customer reviews. In light of our results presented in this paper, it remains to be evaluated as to whether UBiSE can be scaled to technical domains as well.

Our assumption is that the use of the LD version of the Apertium data (and in general of any linguistic data) for cross-lingual model transfer has a number of advantages: It does not rely on proprietary formats and APIs but on standard representation mechanisms and access means (RDF, ontologies, SPARQL, etc.), which also makes linkage and combination with other LD resources easier. Further, the continuous enrichment and growth of the LLOD cloud (e.g., more translations among new language pairs are available) can lead to the improvement of the NLP stack exploiting them with little or no extra effort.

3 Overall architecture

In this section we describe the whole pipeline that the multilingual data traverse: the Apertium source data is taken as input and converted into RDF based on the OntoLex-lemon model. Then, it is linked to the LexInfo[12] catalogue of linguistic categories and published as LD. In the next step, the RDF data is consumed by a sentiment analysis component in a user application, to support cross-lingual model transfer. Such a pipeline is illustrated in Figure 1. Two main components implement such a pipeline, namely Fintan and Pharos®:

Fintan, the Flexible, Integrated Transformation and Annotation eNgineering platform [8] has been developed in the context of the Prêt-à-LLOD project[13] and allows for creating complex transformation pipelines between widely used formats for representing linguistic resources. Fintan allows existing RDF converters to be integrated with stream-based graph processing steps which modify the resulting data to comply with standard data models such as OntoLex-lemon and interlink it with existing LD resources. Fintan thus poses an ideal framework

[12] `https://www.lexinfo.net/ontology/2.0/lexinfo`
[13] `https://www.pret-a-llod.eu/`

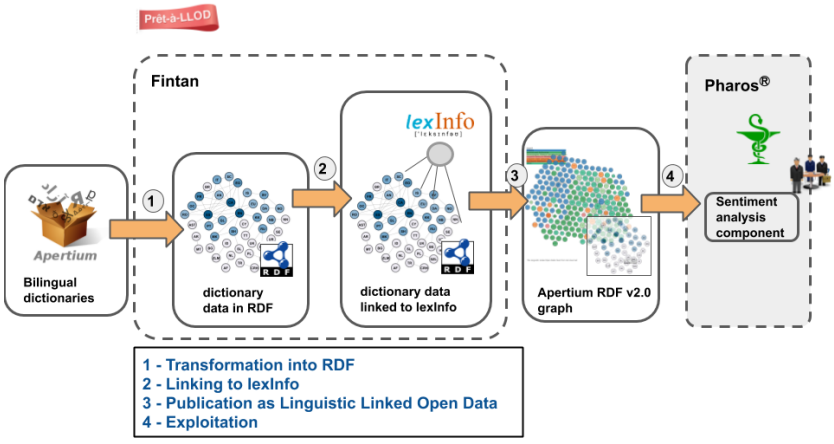

Fig. 1. Apertium RDF v2.0 pipeline: from source data to exploitation

for mapping the Apertium XML data as RDF.

Pharos® is marketed by Semalytix[14] as a Pharma Analytics Platform that provides actionable real-world evidence (RWE)[15] for customers from the global pharmaceutical industry. Since its inception in 2019, the platform has been adopted in more than 10 projects by pharma companies from three countries. RWE extraction requires to analyse large volumes of heterogeneous content, including subjective assessments of patients and medical experts, which is typically available as unstructured text in multiple languages. Underlying Pharos, there is a complex stack of NLP components and modules, comprising entity and concept recognition, relation extraction, sentiment analysis, among others. In this paper, we focus on cross-lingual transfer of an RWE-tailored sentiment model from English to Spanish using an LLOD-based transfer learning framework.

Our final goal is running the whole Apertium pipeline in a fully automated way, therefore periodically gathering updates in Apertium, running the transformation and linking scripts through Fintan, and serving the produced LD to Pharos®, in an automated manner. Manual intervention is only necessary if adjustements to the data model or the mapping of annotation schemes are required (see Sections 4.1 and 4.2). Then, the whole pipeline can be run automatically since such a modelling and mapping design is common for all the Apertium data and dictionaries.

[14] `https://www.semalytix.com`

[15] RWE is evidence for the effectiveness and safety of a drug product, gathered outside of the controlled settings of clinical trials, in order to demonstrate added value of a drug in terms of improvements in quality of life in specific patient populations.

4 Apertium data transformation and linking with Fintan

Some methodologies to convert multilingual language resources into LD can be found in the literature [18]. Particularly, the W3C Best Practices for Multilingual Linked Open Data (BPMLOD) community group proposed a guidelines document for the conversion of bilingual dictionaries, taking Apertium as a motivating example.[16] We followed the steps recommended in such guidelines, slightly adapted, that is: (i) vocabulary selection, (ii) data modelling, (iii) linking, (iv) generation, and (v) publication.

As for the first step, vocabulary selection, we chose the de-facto standard Ontolex-lemon for representing the lexical information contained in the Apertium dictionaries, jointly with its *vartrans* module to specify translation relations (see section 2). The part of speech (POS) information contained in Apertium is represented, in its RDF counterpart, by using LexInfo as reference model, which is a registry of linguistic categories widespread in the linked data community [3]. In the rest of this section we review the remaining steps for the RDF conversion.

4.1 Data modelling

Following the Apertium RDF v1.0 approach, three files are generated for each language pair in a source Apertium dictionary: one for each dictionary (source and target lexicons), and the third one for the translation relations between the corresponding lexical senses. Figure 2 shows the RDF representation of the translation relation between the senses of the entries *safety* in English and *seguridad* in Spanish based on the *vartrans* module of OntoLex-lemon.

To represent the POS tags of Apertium as RDF, a URI in the Apertium namespace is associated to each tag, using the string value of every tag as its local name, e.g. `apertium:n` for the tag `n` (noun), and we assign it as a morphosyntactic property to the lexical entry: `:safety-n-en lexinfo:morphosyntacticProperty apertium:n.`

4.2 Mapping to LexInfo

As a part of the conversion explained in the previous section, a list of approx. 700 category abbreviations used for morphosyntactic description across the datasets in Apertium source files were extracted and gathered under the same namespace (e.g. `apertium:def` for `definite`, `apertium:dat` for `dative case`). However, the tags in Apertium to indicate POS and other morphosyntactic properties are not normalised and sometimes are not very informative. In order to allow for the integration of the Apertium dictionaries among themselves and with external resources, a normalisation process is necessary. To that end we have mapped the Apertium POS tags to LexInfo, resulting in a homogeneous tagging across all the Apertium dataset family and facilitating its querying and reuse.

[16] `https://www.w3.org/2015/09/bpmlod-reports/bilingual-dictionaries/`

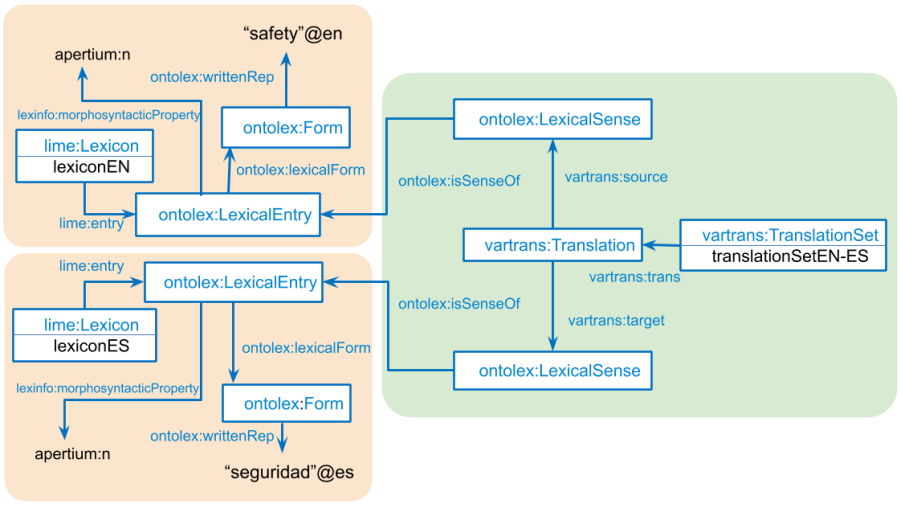

Fig. 2. Modelling example of the translation between "safety"@en and "seguridad"@es, prior to the linking to LexInfo.

Apertium tag	Lexinfo property	Lexinfo tag
apertium:adj	lexinfo:partOfSpeech	lexinfo:adjective
apertium:A	lexinfo:partOfSpeech	lexinfo:adjective
apertium:det	lexinfo:partOfSpeech	lexinfo:determiner

Table 1. Apertium-LexInfo mapping examples, for *adjective* and *determiner*.

We use `lexinfo:morphosyntacticProperty` to account for the Apertium POS individuals initially. The mapping between Apertium POS and LexInfo is defined as a CSV file, which provides `predicate - object` pairs for each of those Apertium tags acting as objects (e.g. `lexinfo:partOfSpeech`, `apertium:vblex`, `lexinfo:verb`). In total, the initial number of Apertium categories identified as POS was 104, which were mapped into 28 different LexInfo categories. Table 1 shows three mapping examples. The initial mapping between the POS Apertium tags and LexInfo was performed manually by the authors and made available online for the review and validation by the larger community of linguists and lexicographers.[17]

4.3 RDF generation

Prior to generate the RDF data, a URI naming strategy has to be defined. To that end, we follow the same approach as in Apertium v1.0, which follows

[17] The mapping is available as CSV and TSV in GitHub and open to comments and modification by the community. See `https://github.com/sid-unizar/apertium-lexinfo-mapping`

the ISA Action recommendations.[18] There are, however, some novelties, as for instance the addition of the `apertium:` namespace to document information in Apertium that could not be mirrored into LexInfo, with the aim of avoiding any loss of information during the transformation process.

As a basis for conversion, we take a shallow converter for Apertium, developed for the ACoLi Dictionary Graph [2]. In order to create a full transformation pipeline from Apertium data into OntoLex-lemon including the LexInfo tagset, we rely on the Fintan platform, which comprises a modular architecture allowing the integration of existing converters. We refer to the technical description of Fintan [8] for its implementation details. Modules of the following types that can be implemented in its pipeline:

- `Loader` modules consume uploaded files or input streams of a specific input format (in our case, the original Apertium data).
- `Splitter` modules are relevant for stream-based graph processing and divide input data into a stream of RDF data segments which can be processed independently, thus avoiding memory and performance limitations.
- `Update` modules consume a stream of RDF data segments and use multi-threading to process multiple segments in parallel.
  The transformation steps are rendered as SPARQL updates which are sequentially executed and optionally iterated to allow recursive operations.
- `Writer` modules export graph data into RDF serializations or other formats. Fintan currently supports the native export of TSV data, however, a custom `Writer` module can be integrated in the same way as a `Loader`.

All modules can be built into complex pipelines using a graphical workflow manager, or be directly called using a command-line interface. Figure 3 shows the Apertium pipeline within the Fintan workflow manager. The Apertium transformation pipeline in Fintan consists of the following steps:

1. The current Apertium repositories are checked out,
2. Morphological properties are extracted from all the source files,
3. With XSLT, each dictionary is transformed from XML to OntoLex-lemon,
4. Using the LexInfo mapping table, a dynamically built SPARQL update script inserts LexInfo morphological categories into the RDF, removing raw Apertium ones where possible,
5. The output is the Apertium RDF data in turtle and TSV formats, for the NLP application to choose the most suitable format for consuming the data.

Given the iterative nature of the conversion, Fintan is a suitable choice for making the workflow more reproducible, user-friendly and less resource-intensive, since some of the Apertium dictionaries are quite large and applying the update to the whole dataset can pose a bottleneck.

[18] `http://ec.europa.eu/isa/actions/01-trusted-information-exchange/1-1action_en.htm`

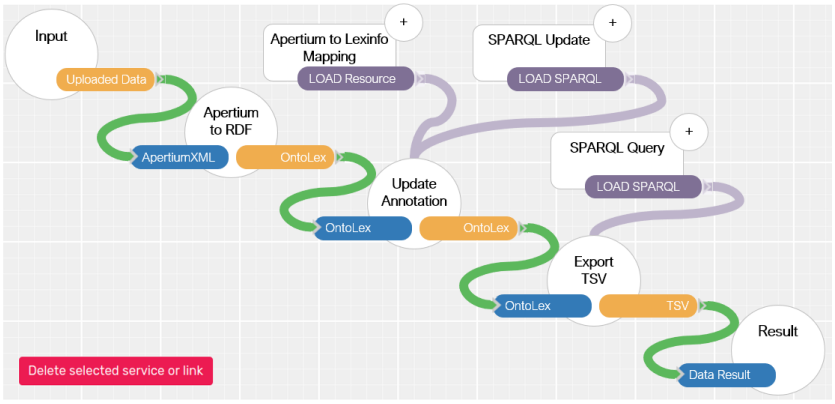

Fig. 3. Conversion pipeline for Apertium in the Fintan Workflow Manager

4.4 Publication

A result of the previously described pipeline, a graph of dictionary linked data, interconnected at the level of lexical entries and linked to an external resource such as LexInfo, has been created. It allows for a seamless exploration of the Apertium data, moving them beyond its original data silos (bilingual dictionaries in XML) and application domain (Machine Translation) to enable other usages like the one illustrated in this paper (sentiment analysis in a multilingual setting). Figure 4 illustrates the new Apertium RDF graph, covering 44 languages and 53 translation sets among them (compare with the 16 languages of the previous Apertium RDF version). It contains a total of 1,535,853 translations among different lexical entries and 1,838,295 links to LexInfo.

A preliminary version of the Apertium RDF v2.0 dictionaries, provided under GPL license (like the original data), is available via `https://github.com/acoli-repo/acoli-dicts`. The release contains the build scripts, such that the data can be locally re-built if new Apertium dictionaries are being published or existing dictionaries are being updated. The build scripts provide an implicit versioning via the time-stamp provided with every RDF dump they create.[19]

5 RDF exploitation

In this section, we demonstrate how the RDF workflow presented above can be exploited in a real-world industry use case. We address the problem of trans-

[19] Access to a testing SPARQL endpoint, as well as a number of example queries to the Apertium RDF v2.0 dataset, can be found at `https://doi.org/10.6084/m9.figshare.12355358`. A stable version of Apertium RDF v2.0 will be uploaded to `http://linguistic.linkeddata.es/apertium/` and hosted by Universidad Politécnica de Madrid (UPM) as part of the Prêt-à-LLOD project and documented through `https://lod-cloud.net/`

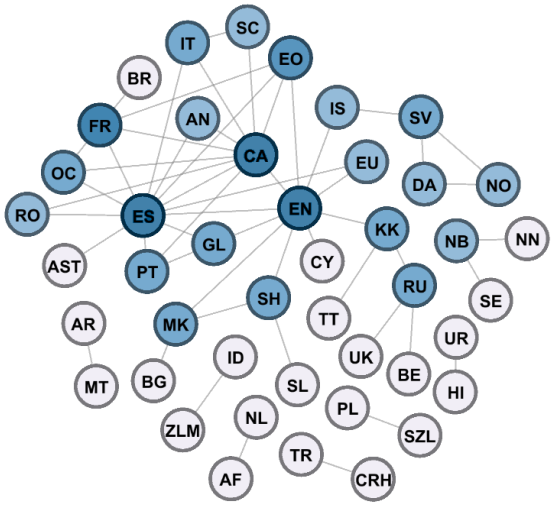

Fig. 4. New Apertium RDF graph. The nodes represent monolingual lexicons and edges the translation sets among them. Darker nodes correspond to more interconnected ones.

ferring a domain-specific model for sentiment prediction that exists for pharmaceutical text in a source language (here: English) to a target language (here: Spanish) for which no labeled training data is available. Our approach capitalizes on a deep learning transfer framework based on BLSE (*Bilingual Sentiment Embeddings*) [1].

In comparison to other approaches, BLSE is relatively parsimonious in terms of language data and resources required, as training signals for the learning algorithm need to be provided only in terms of ground-truth sentiment labels in the source language and a bilingual lexicon which contains translation pairs of words in both languages. In our use case scenario, we can assume that ground-truth labels are available in terms of manual annotations, whereas the selection of the most appropriate bilingual lexicon(s) is subject to empirical evaluation.

In the following, we describe the BLSE architecture, the lexical resources that are acquired using the RDF workflow presented above, and methods to combine such resources in order to increase their domain- and task specificity.

5.1 BLSE Architecture

As can be seen from the high-level architecture displayed in Figure 5, BLSE requires (i) monolingual word embeddings in both the source and target language, (ii) ground-truth annotations in the source language, and (iii) a bilingual dictionary that maps words from the source language to their translations in the target language. These resources provide the foundation for learning mappings M and M' into a bilingual task-specific embedding space. The learning procedure is guided by a composite loss function based on the cross-entropy

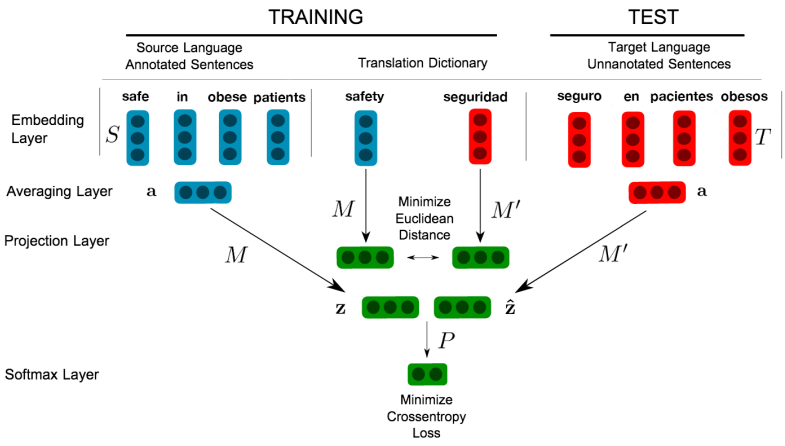

Fig. 5. Overview of BLSE Architecture (slightly adapted from [1])

between sentiment predictions and ground truth labels in the source language and the spatial proximity of source/target pairs from the bilingual dictionary in the bilingual embedding space. The latter part enables the model to tailor target-language embeddings such that they can be used as input to a softmax classification layer that returns target-language predictions without any direct supervision in this language being available (see [1] for more details of BLSE).

5.2 Lexical Resources used in BLSE

Monolingual Word Embeddings used in this study are selected along the two axes of *language* and *domain*: For English, we use *google*[20] open-domain and *PMC*[21] biomedical embeddings. For Spanish, we use *sg_300_es*[22] as open-domain embeddings, and *scielo_wiki*[23] as domain-specific representations. All embeddings were pre-trained on the respective corpus using word2vec [14].

Bilingual Dictionaries In order to inform the cross-lingual projection in BLSE, we apply three different lexicons that provide Spanish translations for English lexical entries. These lexicons were selected according to the criteria of *domain-* and *task specificity*.

[20] Trained on news text, available from `https://drive.google.com/open?id=1GpyF2h0j8K5TKT7y7Aj0OyPgpFc8pMNS`.
[21] Trained on the PubMed Central corpus, available from `http://bio.nlplab.org`.
[22] Trained on Wikipedia text, available from `https://drive.google.com/open?id=1GpyF2h0j8K5TKT7y7Aj0OyPgpFc8pMNS`.
[23] Trained on the concatenation of the Scielo corpus and a medical subset of Wikipedia text, available from `https://zenodo.org/record/2542722#.XeUOo5NKjUK`.

Apertium. For the purpose of a *broad-coverage, open-domain* lexicon, we use Apertium RDF v2.0, as introduced in Section 2.2. In particular, we use the EN-ES translation set, which contains 28,611 translations.

Pharma. As a source of *domain-specific* knowledge in order to render the bilingual embedding space resulting from BLSE training more sensitive to pharma-specific contents, 2,687 bilingual entity lexicalizations from the proprietary Semalytix Knowledge Graph were extracted. As a large repository of pharma-specific knowledge, the graph contains entity types such as diseases and symptoms, drug products and agents, drug manufacturers, therapy areas, among others.

BingLiu. As an *open-domain*, *task-specific* resource, we use the sentiment lexicon originally provided by [12] in its bilingual extension as generated by [1] via machine translation.[24] BingLiu contains 5,749 bilingual lexical entries; we do not make use of the polarity information that is provided alongside each entry.

Before being used in BLSE, each resource undergoes a procedure of (i) *de-duplication* (removing duplicate entries), (ii) *disambiguation* (in case of translation ambiguities, selecting the translation candidate that occurs most frequently in the target-language corpus) and (iii) *filtering* (removing all entries with translations that do not occur in the target-language corpus). This results in 5,084 processed entries for Apertium, 277 for Pharma, and 1,362 for BingLiu.

Lexicon Extension Procedures In order to exploit complementarities in the lexical content of the previously described lexical resources,[25] we generate three extensions as summarized in Table 2. For each extension, the individual source lexicons are composed successively in the given order by either adding novel entries or overwriting existing ones (in case of conflicting translations in the source lexicons). After composition, each extension undergoes the same post-processing procedure described above.

	Source Lexicons	#Entries (original)	#Entries processed
Domain Extension	Apertium + Pharma	31,192	5,307
Task Extension	Apertium + BingLiu	34,254	5,799
Full Extension	Apertium + Pharma + BingLiu	36,941	6,018

Table 2. Overview of extensions generated by composing individual source lexicons, with numbers of original and processed entries (i.e., translation pairs) per extension

[24] Available from `https://github.com/jbarnesspain/blse/tree/master/lexicons/bingliu`

[25] The overlap between these resources amounts to 647 processed entries between Apertium and BingLiu, but only 54 between Apertium and Pharma, and only 12 between Pharma and BingLiu.

6 Experiment: Impact of Lexical Resources on Cross-lingual Transfer of Sentiment Detection Models

In this section, we report on an experimental evaluation of the different configurations of lexical resources as regards their performance in the cross-lingual sentiment projection task.

6.1 Corpus

The corpus used in our experiments consists of a non-parallel sample of comparable English and Spanish transcripts of summaries of conversations between pharma representatives and medical experts. In these conversations, the medical experts are asked to state their opinions and assessments about particular aspects of medical treatments (e.g., safety and effectiveness of a drug, among others). The following examples denote positive and negative assessments of safety and effectiveness, respectively:

(1) DRUG can be safely used in elderly patients with renal failure. – `SAFETY; positive`

(2) No effect on glycaemic control when using DRUG as add-on. – `EFFECTIVENESS; negative`

A collection of 21,400 English summaries is manually annotated with binary sentiment labels at the document level (11,069 positives vs. 10,331 negatives) and subsequently used for training the cross-lingual transfer model in a cross-validation setting. A set of 1,001 Spanish summaries is annotated likewise (559 positives, 442 negatives) in order to provide a test set in the target language which is used for evaluation purposes only.

6.2 Results

	Monolingual Embeddings	Target Language Accuracy
Apertium	google; scielo_wiki	**0.768**
Pharma	google; sg_es_300	0.434
BingLiu	PMC; sg_300_es	0.711
Apertium & Pharma	google; scielo_wiki	0.763
Apertium & BingLiu	google; scielo_wiki	**0.773**
Apertium & Pharma & Bing Liu	google; scielo_wiki	0.767

Table 3. Accuracy scores in the target language obtained from BLSE when different lexicons are used in isolation (upper part) or in combination (lower part).

Performance of Individual Lexicons The upper part of Table 3 shows the results of cross-lingual projection using BLSE when each of the lexicons introduced in Section 5.2 above is injected into the BLSE framework as the only source of bilingual information. We clearly observe that the best accuracy[26] in the target language is due to Apertium (Acc=0.768). The considerable margin over Pharma and BingLiu confirms the status of Apertium as a linguistically rich, general-purpose source of bilingual lexical knowledge. Even though the underlying data set is highly pharma-specific, the relative individual performance of Pharma and BingLiu suggests that task-specific sentiment information benefits cross-lingual projection approaches more than technical domain knowledge.

With respect to the monolingual word embeddings involved, a clear pattern of complementarity can be observed: Apertium benefits most[27] from domain-specific embeddings in the target language, whereas the domain-specific Pharma lexicon is best complemented by open-domain embeddings in both the source and the target language. For BingLiu, using sentiment-specific knowledge from the lexicon and domain knowledge from the (source) embeddings works best.

Performance of Extended Lexicons The impact of lexicon extensions as generated through the procedure described in Section 5.2 can be seen from the lower part of Table 3. In comparison to using Apertium as the only source of bilingual information, we find that lexicon composition in individual configurations can be effective in generating richer bilingual lexical representations that result in more accurate cross-lingual projection of sentiment classifiers. Apparently, this is due to a certain degree of complementarity among the original lexicons, given that extending the general-purpose lexicon Apertium by task-specific knowledge from BingLiu yields the best performance overall (Acc=0.773).

6.3 Discussion

The present case study clearly demonstrates the value of Apertium as an example of a bilingual LLOD resource for cross-lingual transfer of NLP models in practical application scenarios. In our experiments on sentiment projection from English to Spanish in the pharmaceutical domain, we found Apertium to excel both in terms of its individual linguistic richness (being the most informative source of bilingual lexical information among the different resources compared) and resource interoperability (facilitating additional performance gains when being combined with complementary task-specific resources).

The good results are not inherent to the LD nature of the data, but illustrate how high quality resources can be gathered from the LLOD cloud and dynamically combined with other data sources and plugged into NLP pipelines.

[26] Accuracy is defined as the proportion of correct labels in all labels predicted by the model on the test set.

[27] For Apertium, Pharma, and Bing Liu, Table 3 displays only the best-performing configurations of monolingual embeddings.

The induced sentiment model meets an excellent trade-off on the cost-effectiveness spectrum: Without any supervision being required in the target language, its predictive performance is (i) reasonably close to the one of a supervised source language classifier,[28] and (ii) largely superior to a sequential machine translation pipeline, as reported in our previous work [11].

7 Conclusions and future work

This paper outlines the strong potential of LLOD resources to be used as catalyzers of cross-lingual transfer of NLP models in deep learning frameworks. We were able to demonstrate this with the Apertium RDF data processing pipeline for a real-world industry use case involving cross-lingual transfer of a sentiment model in the pharmaceutical domain.

In our experiments, we observed a beneficial effect of composing different lexical resources in order to achieve optimal transfer performance. This underlines the great potential of LLOD-based pipelines for setting up flexible and fully automated transfer workflows which could exploit regular updates of the underlying lexical resources in a dynamic manner.

Despite these encouraging findings, we believe that our approach has not exhausted its full potential and that some challenges remain. For instance, the extension to other language pairs will need additional validation. Further, we plan to extend the current workflow into a fully automated pipeline in order to (i) exploit bilingual lexical information in a fully dynamic manner, thus benefiting from regular updates and extensions of the Apertium data, and (ii) rapidly scale the transfer approach to numerous other languages already available as LLOD.

8 Acknowledgements

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. This work is also based upon work from COST Action CA18209 – NexusLinguarum "European network for Web-centred linguistic data science", supported by COST (European Cooperation in Science and Technology). It has been also partially supported by the Spanish projects TIN2016-78011-C4-3-R (AEI/FEDER, UE) and DGA/FEDER 2014-2020.

References

1. Barnes, J., Klinger, R., Schulte im Walde, S.: Bilingual sentiment embeddings: Joint projection of sentiment across languages. In: Proc. of ACL (2018)

[28] Despite not being exactly comparable due to non-parallel evaluation data, the classifiers resulting from the Task Extension setting differ by only 4.3 points in source vs. target language accuracy (0.816 vs. 0.773, respectively).

2. Chiarcos, C., Fäth, C., Ionov, M.: The ACoLi dictionary graph. In: Proc. of LREC. pp. 3281–3290. ELRA, Marseille, France (2020)
3. Cimiano, P., Buitelaar, P., McCrae, J., Sintek, M.: LexInfo: A declarative model for the lexicon-ontology interface. Journal of Web Semantics **9**(1), 29–51 (2011)
4. Cimiano, P., Chiarcos, C., McCrae, J.P., Gracia, J.: Linguistic Linked Data: Representation, Generation and Applications. Springer International Publishing (2020)
5. Feng, Y., Wan, X.: Learning bilingual sentiment-specific word embeddings without cross-lingual supervision. In: Proc. of NAACL:HLT (2019)
6. Forcada, M.L., Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., Pérez-Ortiz, J.A., Sánchez-Martínez, F., Ramírez-Sánchez, G., Tyers, F.: Apertium: a free/open-source platform for rule-based machine translation. Machine Translation **25**(2), 127–144 (2011)
7. Francopoulo, G., Bel, N., George, M., Calzolari, N., Monachini, M., Pet, M., Soria, C.: Lexical Markup Framework (LMF) for NLP Multilingual Resources. In: Proc. of the Workshop on Multilingual Language Resources and Interoperability. pp. 1–8. Sydney, Australia (2006)
8. Fäth, C., Chiarcos, C., Ebbrecht, B., Ionov, M.: Fintan - flexible, integrated transformation and annotation engineering. In: Proc. of LREC (2020)
9. Gracia, J., Montiel-Ponsoda, E., Vila-Suero, D., Aguado-de Cea, G.: Enabling language resources to expose translations as linked data on the web. In: Proc. of LREC. pp. 409–413 (2014)
10. Gracia, J., Villegas, M., Gómez-Pérez, A., Bel, N.: The apertium bilingual dictionaries on the web of data. Semantic Web **9**(2), 231–240 (2018)
11. Hartung, M., Orlikowski, M., Veríssimo, S.: Evaluating the impact of bilingual lexical resources on cross-lingual sentiment projection in the pharmaceutical domain. http://doi.org/10.5281/zenodo.3707940 (2020)
12. Hu, M., Liu, B.: Mining and summarizing customer reviews. In: Proc. of KDD. pp. 168–177 (2004)
13. McCrae, J.P., Bosque-Gil, J., Gracia, J., Buitelaar, P., Cimiano, P.: The OntoLex-Lemon Model: Development and Applications. In: Electronic lexicography in the 21st century. Proc. of eLex 2017. pp. 587–597 (2017)
14. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representationsin vector space (January 2013)
15. Mogadala, A., Rettinger, A.: Bilingual word embeddings from parallel and non-parallel corpora for cross-language text classification. In: Proc. of NAACL:HLT (2016)
16. SRIA-Editorial-Team: Strategic Research and Innovation Agenda for the Multilingual Digital Single Market. Tech. rep., Cracking the Language Barrier initiative (2016)
17. Søgaard, A., Vulic, I., Ruder, S., Faruqui, M.: Cross-lingual word embeddings. Morgan Claypool (2019)
18. Vila-Suero, D., Gómez-Pérez, A., Montiel-Ponsoda, E., Gracia, J., Aguado-de Cea, G.: Publishing Linked Data on the Web: The Multilingual Dimension. In: Cimiano, P., Buitelaar, P. (eds.) Towards the Multilingual Semantic Web, pp. 101–117. Springer Berlin Heidelberg, Berlin, Heidelberg (2014)
19. Yarowsky, D., Ngai, G., Wicentowski, R.: Inducing multilingual text analysis tools via robust projection across aligned corpora. In: Proc. of HLT (2001)
20. Zennaki, O., Semmar, N., Besacier, L.: Inducing multilingual text analysis tools using bidirectional recurrent neural networks. In: Proc. of COLING (2016)
21. Zhou, X., Wan, X., Xiao, J.: Cross-lingual sentiment classification with bilingual document representation learning. In: Proc. of ACL (2016)