D3.3

LEXICAL-SEMANTIC ANALYTICS FOR NLP: DOMAIN LABELING (SOFTWARE)

Author(s): Cesare Campagnano, Federico Martelli, Roberto Navigli, Paola Velardi

Date: 29. 1. 2021

H2020-INFRAIA-2016-2017

Grant Agreement No. 731015

ELEXIS - European Lexicographic Infrastructure

D3.3 LEXICAL-SEMANTIC ANALYTICS FOR NLP: DOMAIN LABELING (SOFTWARE)

Deliverable Number:	D3.3
Dissemination Level:	Public
Delivery Date:	31. 1. 2021
Version:	1.0
Author(s):	Cesare Campagnano, Federico Martelli, Roberto Navigli, Paola Velardi

Project Acronym:	ELEXIS
Project Full Title:	European Lexicographic Infrastructure
Grant Agreement No.:	731015

Deliverable/Document Information

Project Acronym:	ELEXIS
Project Full Title:	European Lexicographic Infrastructure
Grant Agreement No.:	731015

Document History

Version Date	Changes/Approval	Author(s)/Approved by
22. 1. 2021	Initial draft	Cesare Campagnano, Federico Martelli, Roberto Navigli, Paola

		Velardi
29. 1. 2021	Revised version	Cesare Campagnano, Federico Martelli, Roberto Navigli, Paola Velardi

Table of Contents

1 Introduction

This document accompanies and describes the software that is released as deliverable D3.3 (Lexical-semantic analytics for NLP: domain labeling) as part of task T3.3 (B) in work package WP3 (JRA Lexicographic Data for NLP).

The aim of task T3.3 (B) is twofold: (i) on the one hand, it seeks to provide a means to **automatically label dictionary glosses with domain tags** so as to analyze, create and standardize contents across lexicographic data in arbitrary languages, whereas, (ii) on the other hand, it seeks to demonstrate how the quality of the **lexicographic data made available as part of the ELEXIS network of resources enhances the performance of domain labeling models** and, particularly, fosters state-of-the-art results for **low-resourced languages**.

The remainder of this deliverable is structured as follows: in Section 1.1 we briefly introduce the task of domain labeling; in Section 2 we provide details concerning the architecture employed by the software whose usage we fully report in Section 3. Finally, in Section 4, we describe the experimental setup and results which testify to the quality of the software we make available for domain labeling.

1.1 Domain Labeling

We define "domain labeling" as the task of performing classification by means of tagging **dictionary glosses** with **domains of knowledge labels**. Given that domain labeling entails the process of **assigning categories** (from a finite inventory of many) **to raw text**, we can identify domain labeling as belonging to the wide spectrum of applications included within the broader task of text classification (Aggarwal and Zhai, 2012), together with, inter alia, sentiment analysis or intent detection.

Moreover, it can be argued that the sheer size of the text to be labeled determines whether the task of text classification can be better referred to as "sentence classification". Hence, for our purposes, we will consider domain labeling as a subtask in the broader scope of sentence classification.

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

2 Architecture

In the past few years, Natural Language Processing witnessed the introduction of several neural architectures, bridging the gap with **human-level performance** in many tasks, and enhancing the representational power of neural models. One of the most important additions was the **Transformer architecture** (Vaswani et al., 2017), followed by the first pre-trained multilingual contextualized encoder, multilingual BERT[1] (Devlin et al., 2019), which works in more than 100 languages and is able to encode sentences in different languages in the same shared multilingual space.

In the implementation of task T3.3 (B), **we exploited the representational power of multilingual BERT** (M-BERT) and followed its authors in using a simple yet efficient **Sequence Classification architecture**, composed of a **multilingual BERT encoder**, which encodes the input, and a **linear layer**, to perform the classification on top of the [CLS] token.

Considering that we are interested in tagging glosses according to domain labels, we needed to exploit the wealth of knowledge contained within lexicographic resources that feature these data in order to train our system. To do so, we had our model classify sentences formed from “gold” triples <*d, l, g*>, where *l* and *g* are, respectively, the lemma and gloss which are fed as input to the encoder, and *d* is the domain label the model is fine-tuned to predict.

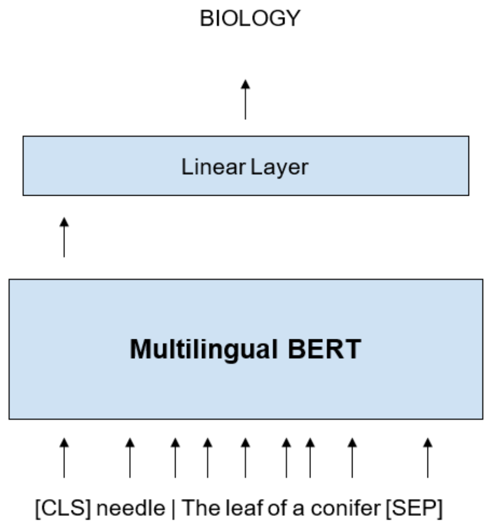

Figure 1: Architecture diagram for our domain labeling system when the lemma l is “needle”, the gloss g is “The leaf of a conifer”, and the model is trained to predict the domain label d “BIOLOGY”.

[1] https://github.com/google-research/bert/blob/master/multilingual.md

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

3 Software Usage

Our domain labeling system requires lexicographic data to be trained with, containing information about lemmas, glosses and domain labels. Once the model is trained, the user can employ it to automatically tag new and unseen gloss instances according to the same domain labels used in the training set.

3.1 ELEXIS GitHub repository

The code is available at https://github.com/elexis-eu/D3.3.
Following are: (i) the README file that can be found at the same address and (ii) the detailed instructions contained in the USAGE.md file.

```
-----------------------------------------------------------------
README - Domain Labeling

Sapienza NLP Group
Sapienza University of Rome
http://nlp.uniroma1.it
-----------------------------------------------------------------

This package contains a software whose function is to
automatically tag dictionary glosses with domains of knowledge
labels.

The software exploits the representational power of
multilingual BERT (M-BERT) and makes use of a simple yet
efficient Sequence Classification architecture, composed of a
multilingual BERT encoder -- which encodes the input -- and a
linear layer, to perform the classification.


--------
CONTENTS
--------


This package contains the following components:

├── config
│   └── custom.jsonnet  # base jsonnet configuration file
├── data            # folder for data
├── LICENSE
```

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

```
├── models
│   ├── released       # folder for released models
│   └── trained        # folder for trained models
├── src                # source code
│   ├── allen_elements   # allennlp-related code
│   ├── serve.py       # simple cli interactive demo
│   └── main.py        # training entry point
├── requirements.txt
├── README
└── USAGE.md

------------
REQUIREMENTS
------------

A python 3.7 installation (possibly in an environment manager).

------------
INSTALLATION
------------

Before running the following command, it is advised to create a new
environment
(e.g., with conda) to avoid conflicts with your current one.

Install requirements via pip install -r requirements.txt.

The code is based off of AllenNLP's library. See usage instructions at
USAGE.md.

-------
AUTHORS
-------

Roberto Navigli, Sapienza University of Rome
(navigli@di.uniroma1.it)

Federico Martelli, Sapienza University of Rome
(martelli@di.uniroma1.it)

Acknowledgments go to Niccolò Campolungo, Babelscape
(campolungo@babelscape.com), for his contribution to the project.

---------
COPYRIGHT
---------

This software is licensed under a Creative Commons Attribution-
Noncommercial-
Share Alike 4.0 License. See the LICENSE file for details.
```

```
---------------
ACKNOWLEDGMENTS
---------------

This software is an output of the ELEXIS project (https://elex.is). This
project has
received funding from the European Union's Horizon 2020 research and
innovation
programme under grant agreement No. 731015.
```

```
USAGE - Domain Labeling

# Training a model

In order to train a model, place train.tsv, dev.tsv, test.tsv files under a
subfolder of data/ (e.g., data/my-folder/train.tsv etc)
The file structure is pretty simple: every line must contain the label and
the lemma/gloss pair to be fed to the model, separated by the tab
character. For example:

POLITICS_GOVERNMENT_AND_NOBILITY [TAB] royal family | Royal persons
collectively

Note that the usage of lemmas is not mandatory, since the model works even
with glosses only. In the example above, lemma and gloss are separated by
means of a pipe character, otherwise, the example would look like the
following:

POLITICS_GOVERNMENT_AND_NOBILITY [TAB] Royal persons collectively

Once the training data is ready, you can start training by executing:

PYTHONPATH=src/ python src/main.py <data-folder-name>

Where, in case your dataset files were placed under data/my-dataset, <data-
folder-name> is my-dataset.

## Running a demo of the trained model

Once the model has finished training, all the training files will be saved
under models/trained/<data-folder-name>. In case you wish to try an
interactive version of the trained model, you can launch it via python
src/serve.py trained/<data-folder-name>.

# Using a released / trained model
```

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

Download the WordNet-based model at https://drive.google.com/drive/folders/1v_sIGDcdx-KT6szYEo2DAeNq9NqM0ABs?usp=sharing and place it under the models/released/ folder.

To run a simple interactive command-line demo, run the following command:

```
python src/serve.py released/wn
```

To tag a file in the tsv format described above (in case you only have raw sentences, simply prepend NODOMAIN\t to every line of the file), run the following command:

```
allennlp predict models/released/wn.tar.gz <path/to/file.tsv> --output-file
<path/to/output.jsonl> --batch-size <batch-size> --cuda-device 0 --use-
dataset-reader --include-package src.allen_elements --silent
```

For further information about the command, check out the [AllenNLP Documentation page](https://docs.allennlp.org/v1.3.0/api/commands/predict/) or type allennlp predict --help.

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

4 Experimental Setup

In this Section we briefly describe the experimental setup and results for **two distinct settings** in which we assessed the quality of our automatic domain labeling system, also, with respect to the improvements brought about by lexicographic data made available in the context of the ELEXIS project.

We used **BabelNet 4.0**[2] as our reference inventory to gather the data needed in order to conduct our experiments. Particularly, we retrieved from BabelNet the subset of WordNet 3.0 nominal synsets. Given that each of these synsets *s* has been manually labeled in BabelNet according to one or more of **37 domain labels** *d** (see Section 6.1 in the Appendix for the complete list of domains), we collected a set *S* of 81,975 <*d*, s*> pairs. Also, considering that each synset in BabelNet subsumes the lexicalizations (lemmas) *l* and glosses *g* which define its meaning in more than 200 languages, we have consequently been able to gather triples <*d, l, g*> in any of the supported languages to feed our system with (see also Section 2). Note that, in the case of multiple domain labels associated with a given synset, we duplicated <*d, l, g*> triples so as to make every triple a unique instance. Also note that the system is able to exploit triples in which *l* is not defined.

As a means to assess the improvement in terms of performance when ELEXIS lexicographic data is taken into account (henceforth, **ELEXIS data**), we collected the set of resources made available in the context of the ELEXIS project. We inspected each resource *r* to check whether <*d, l, g*> triples were featured and could have been extracted, then proceeded to parse each *r* and manually map domain labels for each distinct *r* to BabelNet's.

For a full breakdown of the resources contained in ELEXIS data, along with the triples we managed to extract for each of them (i.e. 18,112), see Table 1. For an excerpt of domain label mappings, see instead Table 4 in the Appendix of this document.

[2] https://babelnet.org/

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

institution	lexicographic data	lang	#triples
IBL	BG Dictionary New Words + Explanatory Dictionary	BG	1,031
IHJJ	School dictionary	HR	3,960
RAE	Dictionary RAE 22	ES	11,084
UB FMG	GeollSSterm1	RS	1,562
ZRC SAZU	SNB	SI	475

Table 1: Breakdown of the ELEXIS resources we employed for domain labeling purposes

For each model in each distinct setting we report, as experimental results: model name (model), number of unique triples in the training set (#train), development set (#dev) and test set (#test), respectively, as well as macro-F1 and weighted-F1 scores as computed on the relative test sets.

4.1 English domain labeling with ELEXIS data

With this setting, we aimed to analyze the impact carried on domain labeling performance by ELEXIS data when these are used in a monolingual environment. Particularly, we selected all triples <*d, l, g*> for all synsets in *S* (see Section 4), considering only lexicalizations and glosses in English which come from the WordNet resource in BabelNet. We split training, development and test set preserving, for each, the domain distribution of the whole dataset. We allowed instances tagged with multiple domain labels to appear in the training set only.

We then simply injected the ELEXIS data on top of the training data for this baseline. Results in Table 2 testify to the **high quality of the resources in the ELEXIS data**. First, the improvement of 1.5 points in macro-F1 shows **ELEXIS data can provide a means to level performances in the least covered domains**. Secondly, this also entails that **the high quality of ELEXIS data fits well the representational power of M-BERT**, to the point of raising performances on an English-only test set, even though the additional training data features only languages other than English.

model	#train	#dev	#test	macro-F1	weighted-F1
EN (base)	83,892	2,377	4,890	65.8	75.8
EN+ELEXIS data	102,004	2,377	4,890	**67.3**	**76.2**

Table 2: Results on the English domain labeling setting

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

4.2 Low-resourced languages domain labeling with ELEXIS data

With this setting, we aimed to witness a performance boost when adding high-quality ELEXIS data -- which include low-resourced languages triples -- on top of a baseline containing low-resourced languages triples too, but taken from open source resources only. To build this setting we selected all triples <*d, l, g*> for all synsets in *S* (see Section 4), considering only lexicalizations and glosses in the same languages covered by our ELEXIS data (Bulgarian, Croatian, Spanish, Serbian and Slovenian) which come from one or more of the following resources in BabelNet: WordNet, Wikipedia, Wikidata and Wikidis.[3] Once more, we split training, development and test set preserving, for each, the language and domain distribution of the whole dataset. We allow instances tagged with multiple domain labels to appear in the training set only.

As in Section 4.2, we injected ELEXIS data on top of the baseline's training set.

Results in Table 3 once more show an overall boost in both macro and weighted-F1 which provides further evidence of the quality of the ELEXIS data. In general, this setting is particularly valuable since it shows how the "gold" **ELEXIS lexicographic data contributes to an increase in performance when added on top of low-resourced languages data retrieved automatically from open sources** (e.g. via BabelNet).

model	#train	#dev	#test	macro-F1	weighted-F1
LR (base)	109,677	8,454	17,752	55.0	70.9
LR+ELEXIS data	127,789	8,454	17,752	**56.8**	**71.6**

Table 3: Results on the low-resourced languages domain labeling setting

4.3 Remarks

Results shown in Table 2 and Table 3 reflect the accuracy with which the models are able to correctly guess the answers established in the gold test sets. Still, **there is reason to believe that the actual performance of the aforementioned models is way higher**, especially when it comes to human fruition of the domain labeling system. In fact, analyzing the confusion matrix for our models, it is easy to identify areas in which two perfectly plausible domains of knowledge could have been chosen as equally valid.

[3] Wikidis contains glosses found in Wikipedia disambiguation pages associated with some synset.

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

As cases in point, we report two instances taken from the model described in Section 4.1, in which the model predictions have been considered wrong with respect to the test sets annotations, but which are arguably appropriate:

1)

Lemma: boat

Gloss: A small vessel for travel on water

Gold domain: TRANSPORT_AND_TRAVEL

Model predicted: NAVIGATION_AND_AVIATION

2)

Lemma: commemoration

Gloss: A recognition of meritorious service

Gold domain: HERALDRY_HONORS_AND_VEXILLOLOGY

Model predicted: WARFARE_VIOLENCE_AND_DEFENSE

Moreover, we observe that ELEXIS lexicographic data is still being processed, which unavoidably affects the performance boost shown in the previous sections. It is legitimate to assume that, while more data is made available within ELEXIS (especially with respect to data which can be exploited for domain labeling), **performances in all reported settings will continue to grow accordingly**.

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

5 References

Aggarwal, C. C., & Zhai, C. *(2012). A survey of text classification algorithms. In Mining text data (pp. 163-222). Springer, Boston, MA.*

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. *(2019, June). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 4171-4186).*

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. & Polosukhin, I. *(2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).*

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

6 Appendix

6.1 Mapping examples

In Table 4 we report examples of manual domain mapping for a few ELEXIS lexicographic resources (lexicographic data), where "source domain" is the original domain name (as used in the source data), and "mapped domain" is the domain resulting from the mapping.

lexicographic data (inst.)	lang	source domain	mapped domain (BabelNet)
BG Dictionary New Words (IBL)	BG	Мед.	HEALTH_AND_MEDICINE
Dictionary RAE 22 (RAE)	ES	Danza	MUSIC_SOUND_AND_DANCING
School dictionary (IHJJ)	HR	tehn.	CRAFT_ENGINEERING_TECHNOLOGY
SNB (ZRC SAZU)	SI	tekst.	TEXTILE_FASHION_AND_CLOTHING

Table 4: Excerpt of manual domain mapping for ELEXIS resources

6.2 Domains list

Following, is the full list of 37 BabelNet domains we used in our whole experimental setup:

ART_ARCHITECTURE_AND_ARCHAEOLOGY; BIOLOGY; BUSINESS_INDUSTRY_AND_FINANCE; CHEMISTRY_AND_MINERALOGY; COMMUNICATION_AND_TELECOMMUNICATION; COMPUTING; CRAFT_ENGINEERING_TECHNOLOGY; CULTURE_ANTHROPOLOGY_AND_SOCIETY; EDUCATION_AND_SCIENCE; EMOTIONS_AND_FEELINGS; ENVIRONMENT_AND_METEOROLOGY; FARMING_FISHING_AND_HUNTING; FOOD_DRINK_AND_TASTE; GEOGRAPHY_GEOLOGY_AND_PLACES; HEALTH_AND_MEDICINE; HERALDRY_HONORS_AND_VEXILLOLOGY; HISTORY; LANGUAGE_AND_LINGUISTICS; LAW_AND_CRIME; LITERATURE_AND_THEATRE; MATHEMATICS_AND_STATISTICS; MEDIA; MUSIC_SOUND_AND_DANCING; NAVIGATION_AND_AVIATION; NODOMAIN; NUMISMATICS_AND_CURRENCIES; PHILOSOPHY_PSYCHOLOGY_AND_BEHAVIOR; PHYSICS_AND_ASTRONOMY; POLITICS_GOVERNMENT_AND_NOBILITY; RELIGION_MYSTICISM_AND_MYTHOLOGY; SEX; SPORT_GAMES_AND_RECREATION; TEXTILE_FASHION_AND_CLOTHING; TIME; TRANSPORT_AND_TRAVEL; VISUAL; WARFARE_VIOLENCE_AND_DEFENSE

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.