EvaLatin 2020

-

Guidelines

Version 1.0

December 10, 2019

Rachele Sprugnoli and Marco Passarotti
CIRCSE Research Centre, Università Cattolica del Sacro Cuore
Largo Agostino Gemelli 1, 20123 Milano
rachele.sprugnoli[AT]unicatt.it
marco.passarotti[AT]unicatt.it

Contents

Chapter 1

Introduction

EvaLatin 2020 is the first campaign totally devoted to the evaluation of Natural Language Processing (NLP) tools for the Latin language. The campaign is designed following a long tradition in NLP, see for example other campaigns such as MUC, SemEval, CoNLL, EVALITA, with the aim of answering two main questions:

- How can we promote the development of resources and language technologies for the Latin language?
- How can we foster collaboration among scholars working on Latin and attract researchers from different disciplines?

EvaLatin first edition has 2 tasks (i.e. Lemmatization and PoS tagging) each with 3 subtasks (i.e. Classical, Cross-Genre, Cross-Time). Shared data and a scorer are provided to the participants. Participants can choose to participate in either one or all tasks and subtasks. The organizers rely on the honesty of all participants who might have some prior knowledge of part of the data that will be used for evaluation, not to unfairly use such knowledge.
EvaLatin is organized within the "Workshop of Language Technologies for Historical and Ancient Languages" (LT4HALA), co-located at LREC 2020[1]. The workshop will be held in Marseille, France, on Tuesday 12 May 2020. EvaLatin is an initiative endorsed by the Italian association of Computational Linguistics (AILC)[2] and it is organized by the CIRCSE research centre at the Università Cattolica del Sacro Cuore in Milan, Italy, in the context of the *LiLa: Linking Latin* ERC project[3].
For any update, please check the LT4HALA website: `https://circse.github.io/LT4HALA/`.

[1] `https://lrec2020.lrec-conf.org/en/`
[2] `http://www.ai-lc.it/`
[3] `https://lila-erc.eu/`

6

Chapter 2

Data

The dataset of EvaLatin 2020 is made of texts taken from the Perseus Digital Library[1] [3]. Texts have been processed with author-specific UDPipe models [5] and then manually corrected by Latin language experts.
Automatic models were trained on "Opera Latina" [2], a corpus manually annotated since 1961 by the Laboratoire d'Analyse Statistique des Langues Anciennes (LASLA) at the University of Liège[2]. Based on an agreement with LASLA, the "Opera Latina" corpus cannot be released to the public but we can use it to create models for NLP tasks. Thus we converted the original space-separated format to CoNLL-u and we trained automatic models using the UDPipe pipeline[3]. Models were run on the Perseus files: we downloaded the files from the Perseus Github repository[4] and then we transformed them in raw texts removing punctuation and converting *v* into *u* (so that *vir* 'man' becomes *uir*). The output of this automatic annotation have been manually checked and corrected by two annotators. Any doubts have been resolved by a third Latin language expert.

2.1 Data Format

Training data are distributed in the CoNLL-U format[5]. Following such format, annotations are encoded in UTF-8 plain text files containing:

- Comment lines starting with an hashtag (`#`): one line indicates the sentence id (`# sent_id`) and another line reports the texts of the sentence (`# text`).

- Lines containing the annotation of a token in 10 fields separated by a tab. When an annotation is not available, an underscore (`_`) is used instead. The 10 fields of the CoNLL-u format are the following:

[1] `http://www.perseus.tufts.edu/`
[2] `http://web.philo.ulg.ac.be/lasla/textes-latins-traites/`
[3] `http://ufal.mff.cuni.cz/udpipe`
[4] `https://github.com/PerseusDL/canonical-latinLit`
[5] `https://universaldependencies.org/format.html`

1. ID: numeral identifier of each token, an integer starting at 1 for each new sentence;
2. FORM: word form;
3. LEMMA: lemma of word form;
4. UPOS: universal PoS tag;
5. XPOS: language-specific PoS;
6. FEATS: list of morphological features from the universal feature inventory; underscore if not available;
7. HEAD: identifier of the head of the current word;
8. DEPREL: universal dependency relation to the HEAD;
9. DEPS: list of HEAD-DEPREL pairs;
10. MISC: any other annotation.

In our dataset ID, FORM, LEMMA and UPOS fields are annotated: all the other fields are filled in with underscores.

- Blank lines marking boundaries between sentences.

An example of the data format is given in Figure 1. This format is used for the training data and participants are expected to produce the same format for the final evaluation.

```
# sent_id = 1
# text = Gallia est omnis diuisa in partes tres quarum unam incolunt Belgae
aliam Aquitani tertiam qui ipsorum lingua Celtae nostra Galli appellantur
1   Gallia      Gallia      PROPN   _   _   _   _   _   _
2   est         sum         AUX     _   _   _   _   _   _
3   omnis       omnis       DET     _   _   _   _   _   _
4   diuisa      diuido      VERB    _   _   _   _   _   _
5   in          in          ADP     _   _   _   _   _   _
6   partes      pars        NOUN    _   _   _   _   _   _
7   tres        tres        NUM     _   _   _   _   _   _
8   quarum      qui         PRON    _   _   _   _   _   _
9   unam        unus        DET     _   _   _   _   _   _
10  incolunt    incolo      VERB    _   _   _   _   _   _
11  Belgae      Belgae      PROPN   _   _   _   _   _   _
12  aliam       alius       DET     _   _   _   _   _   _
13  Aquitani    Aquitani    PROPN   _   _   _   _   _   _
14  tertiam     tertius     ADJ     _   _   _   _   _   _
15  qui         qui         PRON    _   _   _   _   _   _
16  ipsorum     ipse        DET     _   _   _   _   _   _
17  lingua      lingua      NOUN    _   _   _   _   _   _
18  Celtae      Celtae      PROPN   _   _   _   _   _   _
19  nostra      noster      DET     _   _   _   _   _   _
20  Galli       Galli       PROPN   _   _   _   _   _   _
21  appellantur appello     VERB    _   _   _   _   _   _
```

Figure 1: Example of the data format.

2.2 Training Data

Texts provided as training data are by 5 Classical authors: Caesar, Cicero, Seneca, Pliny the Younger and Tacitus. For every one of these authors we release around 50,000 annotated tokens in the format described in the previous Section, for a total of almost 260,000 tokens. Each author is represented by one specific text genre: treatises in the case of Caesar, Seneca and Tacitus, public speeches for Cicero, and letters for Pliny the Younger. Table 1 presents details about the training dataset of EvaLatin 2020.

AUTHORS	**TEXTS**	**# TOKENS**
Caesar	De Bello Gallico	44,818
Caesar	De Bello Civili (book II)	6,389
Cicero	Philippicae (books I-XIV)	52,563
Seneca	De Beneficiis	45,457
Seneca	De Clementia	8,172
Pliny the Younger	Epistulae (books I-VIII)	50,827
Tacitus	Historiae	51,420
TOTAL		259,646

Table 1: Texts distributed as training data in EvaLatin 2020.

2.3 Test Data

Tokenization is a central issue in evaluation and comparison because each system could apply different tokenization rules leading to different outputs. In order to avoid this problem, test data will be provided in tokenized format, one token per line, and with a white line separating each sentence. An example of test data format is given in Figure 2 Test data will contain only the tokenized words but not the correct tags, that have to be added by the participant systems to be submitted for the evaluation. For the Lemmatization task, the third field of the CoNLL-u format should be filled in (the others should be filled in with underscores); for the PoS tagging task, the fourth field should be filled in (the others should be filled in with underscores). The gold standard test data, that is the annotation used for the evaluation, will be provided to the participants after the evaluation.

```
# sent_id = 2
# text = Hi omnes lingua institutis legibus inter se differunt
1   Hi
2   omnes
3   lingua
4   institutis
5   legibus
6   inter
7   se
8   differunt

# sent_id = 3
# text = Gallos ab Aquitanis Garumna flumen a Belgis Matrona et Sequana
diuidit
1   Gallos
2   ab
3   Aquitanis
4   Garumna
5   flumen
6   a
7   Belgis
8   Matrona
9   et
10  Sequana
11  diuidit
```

Figure 2: Example of the test data format.

Chapter 3

Tasks and Sub-tasks

Participants can choose to participate in either one or all tasks and subtasks described in this Chapter.

3.1 Tasks

This Section provides details on the two tasks included in EvaLatin 2020.

3.1.1 Lemmatization

Lemmatization is the process of transforming each word form into its corresponding base form found in the dictionary (i.e. lemma). The rules we followed are summarized below:

- verbs are lemmatized under the first person, singular, present, active (or passive, in case of deponent verbs), indicative form: e.g., *accingere* → *accingo*;
- abbreviations are expanded: e.g., token: *L.* → lemma: *Lucius*; token: *s.* → lemma: *salus*;
- the lemma of roman numerals (e.g., *ccc*, *XVII*) is *numerus_romanus*;
- the lemma of Greek words (e.g., Θρασσός) is *uox_greca*;
- the lemma associated to lacunae (e.g. *p..*) is *uox_lacunosa*;
- multi-word expressions are not combined into a single token: e.g. *res publica* is made of two tokens with two different lemmas and PoS tags;
- clitics are not separated from the token: e.g. token: *exercitumque* → lemma:*exercitus* → PoS:*NOUN*.

3.1.2 PoS tagging

In the Part-of-Speech (PoS) Tagging task, systems are required to assign a lexical category (PoS tag) to each token. The universal POS tags[1] used in our corpus are the following:

- ADJ: adjective. They modify nouns and specify their properties or attributes, e.g.: *inopinantes*, *album*. They are distinguished from determiners (see the DET tag) and from cardinal numbers (see the NUM tag). Ordinal numbers, such as *tertia*, can be annotated as ADJ or ADV depending on the context.

- ADP: adposition. Adposition is a cover term for prepositions. Examples: *post*, *in*, *trans*.

- ADV: adverb. Adverbs modify other words in the sentence, especially verbs, providing information about manner, degree, cause, place, or time. Examples: *semper*, *paulatim*, *simul*.

- AUX: auxiliary. Auxiliary verbs are verbs that modify another verb, often to change the tense. Latin is a synthetic language thus it tends to express functional meaning with affixes, not with auxiliary verbs. In our dataset there are only two auxiliary verbs, i.e. *sum* ("to be") and *eo* ("to go"). This second auxiliary is used in periphrastic future passive infinitive, e.g. *ad castra iri oportere*.

- CCONJ: coordinating conjunction. Coordinating conjunctions are words that link constituents without syntactically subordinating one to the other. Examples: *et*, *atque*, *uel*.

- DET: determiner. A determiner is a word that occurs together with a noun or noun phrase expressing the reference of that noun or noun phrase in the context. Examples: possessive determiners, *nostros* and demonstrative determiners, *hoc*.

- INTJ: interjection. Interjections are words used as exclamations, thus expressing an emotional reaction: they are not syntactically related to the rest of the sentence. Examples: *mehercule*, *agedum*.

- NOUN: noun. This tag is used for common nouns typically denoting a person, place, thing, animal or idea. Examples: *mater*, *senatus*, *bellum*, *dignitatem*, *auis*. Gerunds and infinitives functioning as nouns are always annotated with the VERB tag.

- NUM: cardinal numerals. Example: *milia*, *XVIIII*.

- PART: particle. Particles are function words associated with another word or phrase. In our dataset they encode the grammatical category of negation . Examples: *non*, *haud*, *ne*.

[1] `https://universaldependencies.org/u/pos/index.html`

- PRON: pronouns. For example: personal pronouns (*ego*), reflexive pronouns (*sibi*), relative pronouns (*quibus*). Possessives pronouns are instead annotated as DET.
- PROPN: proper noun. Proper nouns are nouns that identifies single entities: they are the name (or part of the name) of a specific individual, place, deity. Examples: *Lucilio*, *Lugundum*, *Mosella*, *Uenus*. Proper nouns are often present in the dataset in abbreviated forms: e.g., *G.* is the abbreviation of *Gaius*, *Tib.* is the abbreviation of *Tiberius*.
- SCONJ: subordinating conjunction. Subordinating conjunctions are conjunctions that link constructions by making one of them a constituent of the other. Examples: *postquam*, *dum*.
- VERB: verb. Verbs convey actions, occurrences, or states of being. Examples: *rapiebat*, *scire*, *potest*.
- X: other. This tag is used for words that cannot be assigned a PoS category. In the EvaLatin dataset the tag X is used for Greek words (thus with lemma *uox_greca*) and for lacunae (thus with lemma *uox_lacunosa*).

Please note that the tags PUNCT (punctuaction) e SYM (symbol), included in the UD PoS tags, are not used in the EvaLatin dataset.

3.2 Sub-tasks

Each of the aforementioned tasks has three sub-tasks:

1. **Classical**: test data will be of the same genre and period of the training data;
2. **Cross-genre**: test data will be of a different genre compared to the ones included in the training data;
3. **Cross-time**: test data will be of a different period compared to the ones included in the training data.

Through these sub-tasks, we aim to enhance the study of the portability of NLP tools for Latin across different genres and temporal periods analysing the impact of genre-specific and diachronic features.

Chapter 4

Evaluation

Each participating team will initially have access only to the training data. Later, the unlabelled test data will also be released. After the assessment, the labels for the test data will also be released.

The scorer employed for EvaLatin is a modified version of the one developed for the "CoNLL18 Shared Task on Multilingual Parsing from Raw Text to Universal Dependencies"[1]. An example of the output of the scorer is given in Figure 3. The evaluation starts by aligning the system-produced words to the gold standard ones; given that we provide test data already tokenized and sentence splitted, the alignment for tokens, sentences and words should be perfect (that is 100.00). Then, UPOS tags and lemmas are evaluated: precision, recall, F1 and accuracy are calculated. The final ranking will be based on accuracy.

Metric	Precision	Recall	F1 Score	AlignAcc
Tokens	100.00	100.00	100.00	
Sentences	100.00	100.00	100.00	
Words	100.00	100.00	100.00	
UPOS	85.94	85.94	85.94	85.94
Lemmas	89.06	89.06	89.06	89.06

Figure 3: Example of scorer output.

As a baseline, we will provide the accuracy obtained on test data using UDPipe [4] trained with the model based on the Perseus Universal Dependencies Latin Treebank[2] [1], as available also in the web interface of the tool[3].

[1] `https://universaldependencies.org/conll18/evaluation.html`
[2] `https://github.com/UniversalDependencies/UD_Latin-Perseus/`
[3] `http://lindat.mff.cuni.cz/services/udpipe/`

16

Chapter 5

How to Participate

Participants will be required to submit their runs and to provide a technical report for each task (with all the related sub-tasks) they participated in.

5.1 Submitting Runs

Each participant can submit runs for each subtask within each task. A run should be produced according to the 'closed modality': the only annotated data to be used for training and tuning the system are those distributed by the organizers. Other non-annotated resources, e.g. word embeddings, are instead allowed. The second run will be produced according to the 'open modality': annotated external data, such as the Latin datasets of the Universal Dependencies initiative, can be also employed. All external resources are expected to be described in the systems' reports. The closed run is compulsory, while the open run is optional.
Once the system has produced the results for the task over the test set, participants have to follow these instructions for completing your submission:

- name the runs with the following filename format:
  task_subtask_teamName_systemID_modality.conllu.
  For example: *pos_classical_unicatt_1_closed.conllu* would be the first run of a team called *unicatt* using the closed modality for the PoS tagging task and the Classical subtask. *lemma_cross-genre_unicatt_2_open.conllu* would be the second run of a team called *unicatt* using the open modality for the lemmatization tagging task and the Cross-genre subtask.

- send the file to the following email address: rachele.sprugnoli[AT]unicatt.it, using the subject "EvaLatin Submission: task - teamName", where the "task" is either *PoS* or *Lemma*.

5.2 Writing the Technical Report

Technical reports will be included in the proceedings of the Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) as short papers and will be published along the LREC 2020 proceedings. Reports must be submitted through the START platform (URL available soon). All the reports must meet the following requirements:

- they must be written in English;
- they must be formatted according to the LREC 2020 conference style[1];
- the maximum length is 4 pages (excluding references);
- they should contain (at least) the following sections: description of the system, results, discussion, references.

Reports will receive a light review: we will check for the correctness of the format, the exactness of results and ranking, and overall exposition. If needed we will contact the authors asking for corrections.

[1] `https://lrec2020.lrec-conf.org/en/submission2020/authors-kit/`

Appendix A

Tokens Modified in Perseus Files

Bellum Civile (Liber II)

- sent_id 63, token 7: aminis → laminis
- sent_id 235, token 26: neubi → necubi

Bellum Gallicum

- sent_id 317, token 16: ego → agi

Historiae

- sent_id 611, token 14: tecgmen → tegmen

Letters

- sent_id 293, token 13: ualcas → ualeas
- sent_id 324, token 20: mira → mora
- sent_id 616, token 12: Lepcitanorum → Leptitanorum
- sent_id 628, token 14: carcere → carere
- sent_id 645, token 20: primi → primis
- sent_id 711, token 12: rectis → tectis
- sent_id 1495, token 14: asulescentulus → adulescentulus
- sent_id 1588, token 7: sc → sed
- sent_id 1782, token 3: hae → hac
- sent_id 1811, token 10: acre → aere
- sent_id 1881, token 1: Unodeuicensimo → Undeuicensimo

- sent_id 1901, token 17: sc → se
- sent_id 1943, token 52: passurus → passurum
- sent_id 1993, token 2: testinata → destinata
- sent_id 2125, token 10: scierint → scirent
- sent_id 2300, token 20: cos → eos
- sent_id 3183, token 29: atquc → atque

Appendix A

Selection of Resources for Latin

- Lemma embeddings: https://embeddings.lila-erc.eu/
- Latin texts and embeddings: http://www.cs.cmu.edu/~dbamman/latin.html
- Word embeddings: https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-1989
- CLTK: http://cltk.org/
- UD Latin PROIEL: https://github.com/UniversalDependencies/UD_Latin-PROIEL
- UD Latin ITTB: https://github.com/UniversalDependencies/UD_Latin-ITTB
- UD Latin Perseus: https://github.com/UniversalDependencies/UD_Latin-Perseus
- Latin texts: https://github.com/PerseusDL
- Collatinus: https://outils.biblissima.fr/en/collatinus/index.php
- LEMLAT v.3: https://github.com/CIRCSE/LEMLAT3
- Treetagger: https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
- Glossaria: https://glossaria.eu/outils/lemmatisation/#page-content
- Late Latin Charter Treebank: https://zenodo.org/record/3522868#.Xe-rRtEo85k
- Word Formation Latin (WFL) lexicon: http://wfl.marginalia.it/

Bibliography

[1] David Bamman and Gregory Crane. The ancient greek and latin dependency treebanks. In *Language technology for cultural heritage*, pages 79–98. Springer, 2011.

[2] Joseph Denooz. Opera Latina: une base de données sur internet. *Euphrosyne*, 32:79–88, 2004.

[3] David A Smith, Jeffrey A Rydberg-Cox, and Gregory R Crane. The Perseus Project: A digital library for the humanities. *Literary and Linguistic Computing*, 15(1):15–25, 2000.

[4] Milan Straka, Jan Hajic, and Jana Straková. Udpipe: trainable pipeline for processing conll-u files performing tokenization, morphological analysis, pos tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297, 2016.

[5] Milan Straka and Jana Straková. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada, August 2017. Association for Computational Linguistics.