Finetuning Latin BERT for Word Sense Disambiguation on the *Thesaurus Linguae Latinae*

Piroska Lendvai
Department for Digital Humanities Research & Development
Bavarian Academy of Sciences and Humanities
Munich, Germany
piroska.lendvai@badw.de

Claudia Wick
Thesaurus Linguae Latinae
Bavarian Academy of Sciences and Humanities
Munich, Germany
claudia.wick@thesaurus.badw.de

Abstract

The *Thesaurus Linguae Latinae* (TLL) is a comprehensive monolingual dictionary that records contextualized meanings and usages of Latin words in antique sources at an unprecedented scale. We created a new dataset based on a subset of sense representations in the TLL, with which we finetuned the Latin-BERT neural language model (Bamman and Burns, 2020) on a supervised Word Sense Disambiguation task. We observe that the contextualized BERT representations finetuned on TLL data score better than static embeddings used in a bidirectional LSTM classifier on the same dataset, and that our per-lemma BERT models achieve higher and more robust performance than reported by Bamman and Burns (2020) based on data from a bilingual Latin dictionary. We discuss the differences in sense organizational principles between these two lexical resources, and report about our dataset construction and improved evaluation methodology.

1 Introduction

In the field of Natural Language Processing (NLP), there is a growing amount of languages for which contextualized representation models are created. For Latin, a pretrained BERT model (cf. Devlin et al., 2018) was published by Bamman and Burns (2020), which they finetuned for four classical NLP tasks, among others for Word Sense Disambiguation (WSD). WSD, an area of computational semantics, has been approached in NLP by several machine learning setups (for an overview cf. Navigli, 2009 and Bevilacqua et al., 2021), and recent works (e.g. Scarlini et al., 2020) have also targeted the use of neural models and architectures in combination with lexical knowledge bases and encyclopaedic resources.

WSD is typically cast as supervised classification, where the learning task consists of predicting the appropriate sense label for one or more focus tokens in their context unit, e.g. within a sentence. Based on the application end task, sense labels can be defined in a variety of ways, e.g. aiming to distinguish coarse or fine granularity of senses, binary or multiple sense distinctions, etc. Creating labeled data for a supervised WSD application is nontrivial. Large, sense-annotated benchmark datasets are scarce, especially in languages other than English. A promising resource to be utilized for Latin WSD could be the Latin Wordnet[1]; for its evaluation and references cf. Franzini et al. (2019). Seeking proxy resources and methods to leverage WSD resources is important, since it is expensive to manually produce a sense labeled corpus from scratch that captures contextual information for several senses of a word. Therefore, our study aims to contribute insights into methods that use dictionaries for automatically assigning sense labels.

Bamman and Burns (2020) (henceforth: B&B) constructed WSD data for BERT, a transformer-based language model, by taking the textual examples (i.e., quotes from antique sources) inventorized for a particular headword (aka 'lemma') in the bilingual *Latin Dictionary* of Lewis and Short (1879): to each quote snippet in the first two sense groups of each lemma, they assigned its sense category (i.e., I or II) as gold standard label.

Inspired by this sense inventory creation (i.e., bi-

[1] https://latinwordnet.exeter.ac.uk

Proceedings of the Workshop on Cognitive Aspects of the Lexicon, pages 37–41

https://doi.org/10.26615/978-954-452-056-4_005

nary class labeling) method of B&B, we requested data for the same lemmas that B&B presented, from a currently proprietary resource: the *Thesaurus Linguae Latinae*[2] (TLL)[3]. The TLL is a comprehensive monolingual Latin dictionary that aims to record all meanings of all ancient Latin words, citing all (or a representative sample) of its seen attestations. The TLL is vast: it is estimated to comprise cca. 53k-56k entries as of now[4], so it likely holds a major part of the quotes that occur in L&S and thus in the B&B WSD dataset.

The prospect of comparing WSD performance across datasets constructed from two dictionaries – one bilingual, another monolingual – was intriguing in several scholarly respects, a.o. for gaining quantitative insights into dictionary structuring practices, or even for attempting to validate sense structuring in an empirical way. After inspecting the data, we realized that a direct comparison of machine learning performance based on data constructed from the TLL resp. from L&S would be methodologically flawed:

1. We made pilot analyses of the quotes across the B&B and TLL sense labeled data, and noted that sense categorization in TLL and L&S draws on very different semantic principles: for one and the same lemma, the subset of quotes labeled with sense I in B&B can be distributed across both sense class I and II in TLL, and/or vice versa.

2. Working with the methodology of B&B of constructing sense-balanced data would not allow unleashing the full potential of the TLL data size. As we chose not to discard quotes (i.e., did not match the amount of quotes in the smaller sense label set), our TLL dataset became orders of magnitude larger and sense-label-wise possibly more aggregative, thus likely coarser-grained.

Our aims in the current contribution were thus:

- Investigating methods and challenges for experimentally validating sense representations and their WSD distinction
- Giving account of joint work between the Humanities and the NLP communities that deliver complementary expertise
- Reusing a pretrained contextual representation model for Latin, released by Bamman and Burns (2020)
- Reproducing the WSD experiment of Bamman and Burns (2020) via the benchmark data, code, and baseline classifier they released[5]
- Repeating the WSD experiment by finetuning Latin BERT on new WSD data that we constructed from the TLL
- Observing sense organization principles and scale across the two datasets
- Improving experimental methodology by providing a detailed evaluation in terms of F-macro scoring in a per-lemma-WSD-setup.

The paper is structured as follows: first, a short exploratory analysis is given for the B&B resp. the TLL data in terms of the original resources and their construction principles. Afterwards we report on the finetuning experiments and we summarize the study with a conclusion section.

βres definitae:
① *specimina pauca ad illustrandas notiones selecta:*
Ⓐ *tolerandi, sustinendi* (*sc. fortiter sim.*):
① *in universum*: PLAVT. Men. 721 viduam esse mavelim, quam istaec flagitia tua -i (779 perpeti). 978 magis multo -or facilius verba: verbera ego odi. TER. Eun. 244 neque ridiculus esse neque plagas -i possum. PACVV. trag. 279 -or facile iniuriam, si est vacua a contumelia. CIC. Verr. II 3, 95 quem *contumeliae aculeum* -i ... viri boni difficillime possunt. 3, 201 si hoc vectigal aratio tolerare, hoc est Sicilia ferre ac -i potest. Phil. 6, 19 aliae nationes servitutem -i possunt, populi Romani est propria libertas. fin. 3, 42 si dolores eosdem tolerabilius -untur qui excipiunt eos pro patria quam qui leviore de causa (*item in philosophia*: 4, 23 *Panaetius* cum ... de dolore -endo scriberet. Tusc. 4, 60 qui non turbulente humana -antur. *sim. al.*). BRVT. Cic. ad Brut. 24, 6 W. servire et -i contumelias ... odero. VARRO rust. 2, 10, 3 *senes* callium difficultatem ac montium arduitatem ... *non* facile ferunt, quod -undum *est pastoribus. et passim.*
② *-untur qui quid incolumes, sine noxa sim. sustinent* (*exempla potiora; cf. p. 725, 31*): OV. trist. 3, 3, 7 nec caelum -or nec aquis adsuevimus istis. CELS. 2, 18, 3 pisces ..., qui salem non -untur. SEN. epist. 51, 10 quamlibet viam iumenta -untur, quorum durata ... ungula est. COLVM. 8, 17, 8 *mullus* raro ... *vivarii* claustra -itur. PLIN. nat. 31, 23 *fluvii cuiusdam* gurgitem periuri negantur -i velut flammam.
③ *-untur qui pondera corporea sustinent* (*proprie et in imagine; cf. e. g. p. 724, 7*): SEN. contr. 3 praef. 9 quidam *equi* melius equitem -antur, quidam iugum (*addas imagines vol. VII 2, 641, 70 sq.*; *aliter p. 722, 4*). suas. 2, 1 (*ironice*) insueta ... arma non -surae manus (STAT. Theb. 11, 551 [*Pofynices ad fratrem*] exercita ... membra vides *mea*; disce a. -i). SEN. Thy. 931 (*in imag.*) pondera regni non inflexa cervice -i (SIL. 14, 90).
Ⓑ *subeundi, experiendi* (*sc. mala, quibus quis afficitur neglecto respectu fortiter, laboriose sim. perpetiendi; bona v. sub B3*): CIC. rep. 3, 23 cum de tribus unum est optandum, aut facere iniuriam nec accipere, aut et facere et accipere, aut neutrum, optumum est facere impune ..., secundum nec facere nec -i, miserrimum digladiari semper tum faciendis tum accipiendis iniuriis. NIGID. Gell. 9, 12, 6 imminentia fraudis,

Figure 1: Excerpt from the nested structure of the TLL article for the lemma *patior*, meant for human reading.

```
77  patior  I   viduam esse mavelim , quam istaec flagitia tua   pati
78  patior  I   magis multo   patior  facilius verba : verbera ego odi
79  patior  I   neque ridiculus esse neque plagas pati   possum
80  patior  I     patior  facile iniuriam , si est vacua a contumelia
81  patior  I   quem contumeliae aculeum  pati    … viri boni difficillime possunt
82  patior  I   si hoc vectigal aratio tolerare , hoc est sicilia ferre ac    pati   potest
83  patior  I   aliae nationes servitutem pati    possunt , populi romani est propria libertas
84  patior  I   si dolores eosdem tolerabilius   patiuntur   qui excipiunt eos pro patria quam qui leviore
    de causa
85  patior  I   panaetius cum … de dolore patiendo    scriberet
86  patior  I   qui non turbulente humana patiantur
87  patior  I   servire et    pati    contumelias … odero
88  patior  I   senes callium difficultatem ac montium arduitatem … non facile ferunt , quod  patiundum
    est pastoribus . et passim
89  patior  I   nec caelum    patior  nec aquis adsuevimus istis
90  patior  I   pisces … , qui salem non  patiuntur
91  patior  I   quamlibet viam iumenta    patiuntur   , quorum durata … ungula est
92  patior  I   mullus raro … vivarii claustra    patitur
93  patior  I   fluvii cuiusdam gurgitem periuri negantur pati    velut flammam
94  patior  I   quidam equi melius equitem    patiantur   , quidam iugum
95  patior  I   insueta … arma non    passurae    manus
96  patior  I   exercita … membra vides mea ; disce a .   pati
97  patior  I   pondera regni non inflexa cervice pati
98  patior  I   cum de tribus unum est optandum , aut facere iniuriam nec accipere , aut et facere et
    accipere , aut neutrum , optumum est facere impune … , secundum nec facere nec  pati    , miserrimum
    digladiari semper tum faciendis tum accipiendis iniuriis
99  patior  I   imminentia fraudis , quam quis vel facturus cuipiam vel   passurus    est
```

Figure 2: Flattened, sense-labeled WSD data for BERT, derived from the TLL article and its sense inventory, for the lemma *patior*.

[2] https://tll.degruyter.com/about

[3] The Bavarian Academy of Sciences and Humanities plans to make the complete TLL data open source by 2030.

[4] Currently headwords are prepared till letter R.

[5] https://github.com/dbamman/latin-bert

2 Exploratory Data Analysis

2.1 B&B Data

The B&B dataset comprises 8,354 instances for a total of 201 dictionary headwords (lemmas). The source of B&B data is the bilingual dictionary of L&S[6] that is a translation of Freund's dictionary from the 19th century, reflecting edition techniques from 200 years ago.

2.2 TLL Data

The ongoing TLL compilation project started in 1900; its editorial principles have changed every once in a while[7]. Within each TLL article, a contrastive, nested (thus: semantically additive) structure is pursued that can descend as deep as 10+ levels.

Sense groups on the same level are aimed to be of the same sense granularity but to feature mutually exclusive parameters of syntactic or semantic nature or their combination. This implies that the TLL structure does not reflect sense distinctions that depend on their translatability to another language, but its goal is a dichotomic arrangement (which is not always limited to two sense groups on the highest level) of word attestations (quotes from Latin texts) while staying within the same language.

The TLL data was available to us in TEI XML format. Just like B&B, we generated the data from within a single dictionary entry, by definition excluding homonymy, and we only considered the first two main senses of a lemma, labeling all text snippets that are longer than 4 words with the corresponding highest-level sense label (see Figure 2), by recursively descending into – thus flattening – the nested structure of the printed article (see Figure 1). Our TLL data points correspond to 25,227 text snippets for the subset of 40 lemmas, whose part-of-speech distributions are: 40% verbs, 22.5% adjectives, 10% nouns, 27.5% others (adverb, pronoun, preposition, conjunction, particle).

Starting from letter C, in the articles a large amount of words – by definition the lemma onset itself, but also other tokens – are heavily and somewhat irregularly abbreviated, which we had to resolve by extensive human-in-the-loop procedures, e.g. by identifying patterns and writing replacement rules for omitted subword material in a per-lemma fashion. Reconstructing the omitted subword parts of the inflected lemma forms was mandatory for running meaningful WSD experiments because the lemma forms supply a core piece of information to the learning algorithms.

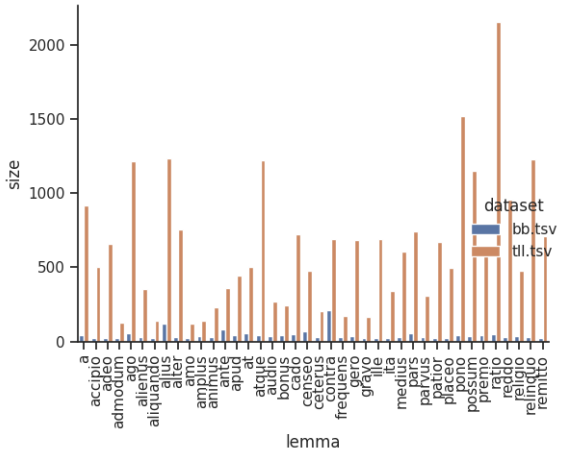

Figure 3: Data size per lemma per dataset. Blue: B&B data. Orange: TLL data.

2.3 Analysis of the Derived Sense Classes

We observed a number of important phenomena about sense classes as derived from the dictionaries.

1. Semantically motivated separation between senses Often, sense separation is mutually exclusive, e.g. the lemma *relinquo* demonstrates that out of the first two main TLL sense groups, I pertains to 'relocation in physical space', as opposed to II that describes 'movement in a figurative sense'.

2. Artificial dichotomy of senses The separation of senses can often be rather artificially constructed, e.g. in *ratio*, and L&S uses such separation practices, e.g. by container labels such as "in general" vs. "in particular", even though the latter split oftentimes does not yield a semantically or syntactically homogeneous group.

3. Lemma vs. Sublemma Classes can also be split on certain grammatical phenomena in L&S, e.g. on participle perfect used as an adjective (cf. *remitto* where this usage makes up class II for B&B), whereas the TLL renders such usage as a so-called sublemma and treats it structurally elsewhere than in the main article, thus the Latin quotes in it do not get extracted into the TLL WSD data.

4. Temporal and domain diversity The TLL has a uniquely wide temporal scope spanning nearly 1000 years from Old and Classical Latin till late antiquity and Christian Latin (cca. AD 700),

[6] http://www.perseus.tufts.edu/hopper/

[7] For an impression see https://publikationen.badw.de/de/thesaurus/lemmata

and encompasses genres beyond the domain of literature, such as legal and medical texts and inscriptions. Thereby, it delivers markedly different semantic representation proportions than (a) the pretrained BERT that saw texts spanning cca. 2000 years, seeing attestation from Middle Latin and Humanism, or (b) the B&B finetuned BERT that saw texts from cca. 200 years, focused on a subset of canonical classical authors. As an example: *religio* in the contemporary sense of *religion* as 'a dogmatic system of faith based on revelation' did not exist before the rise of Christianity; for "pagan" Romans, *religio* denoted 'feelings of awe, fear, respect towards the gods or strictly defined forms of (liturgical) worship'.[8]

5. Truthfulness to sources In both the B&B and the TLL data, their antique sources are not always literally cited, but the quotes are often edited. The TLL maintains more strictness, e.g. no syntactic changes are allowed. In the B&B (aka L&S) data, one regularly finds modified or artificially inserted constructions that diverge from the sources.

3 WSD Experiments

Finetuning BERT is a technique that takes its pretrained language model and explicitly trains it for the WSD task, i.e., in our case on Latin quote – typically on the subsentential level – that are labeled to have class I or class II, as assigned based on the TLL sense inventory. This yields a classification model that can distinguish exactly two meanings for the token that designates the focus lemma. This is certainly a simplified WSD setup, nevertheless helpful for pilot studies to assess the power of newly constructed data for disambiguating between two major senses (or usage contexts) of words. The finetuning task is in contrast with what already took place in the first phase of creating lexical representations, the so-called pretraining. There the task was that BERT's Latin language model learns as many senses of a word as possible.

3.1 Training and Testing Setup

The setup across our WSD experiments on a machine with GPU running Linux Ubuntu 18 is listed below. Splitting the data into partitions for training, development, and testing was done by the method and Github code of Bamman and Burns (2020).

[8]We aim to utilize TLL data for chronological analyses, characterizing and training the recognition of e.g. semantic drift, but this goes beyond the scope of the current paper.

Dataset	Model	mean F-macro	stdev
B&B	biLSTM	.613	.205
	BERT	.695	.213
TLL	biLSTM	.705	.132
	BERT	.794	.143

Table 1: Mean performance scores over 40 lemmas.

- 100 epochs (training rounds) per lemma
- Training and testing performed per lemma
- B&B used cross entropy loss without class weights for training. Since in our data the two classes are imbalanced per lemma, we calculated the weights for each class for the cross entropy loss function
- Performance was evaluated in terms of the unweighted macro F1-score per lemma using Pedregosa et al. (2011). Accuracy would be suboptimal to use as it does not transparently express how well we perform on the two classes and it does not correct for class imbalance
- For each epoch, macro F1 was calculated on the development set
- For each lemma, the best performing development epoch's parameters were used to measure macro F1 on the heldout test set
- As baseline model we used from B&B[9] 200-dimensional static word2vec embeddings (Mikolov et al., 2013) in a biLSTM classifier
- Enclitica were not separated from words since BERT's wordpiece tokenizer[10] was assumed to account for these.

3.2 Evaluation

B&B Dataset We reproduced the B&B WSD study with a similar accuracy score as they report (.737). Next, we derived from the B&B aggregated dataset a per-lemma dataset, on which we trained both classifier models, using the B&B code that we amended with the settings listed in Section 3.1. The results are shown in Table 1. We observe that the B&B per-lemma data are small (cf. Figure 3) and yield statistically unreliable results as standard deviation values are large; this variability is also illustrated by the whiskers of the boxplot (cf. Figure 4). While Table 1 reports the means and the standard deviations, the boxplots show the median.

[9] https://github.com/dbamman/latin-bert/blob/cd6bea9f7ff84ff4b18c172f3d5719d1d3198e69/case_studies/pos_tagging/scripts/download_static_vectors.sh

[10] https://ai.googleblog.com/2021/12/a-fast-wordpiece-tokenization-system.html

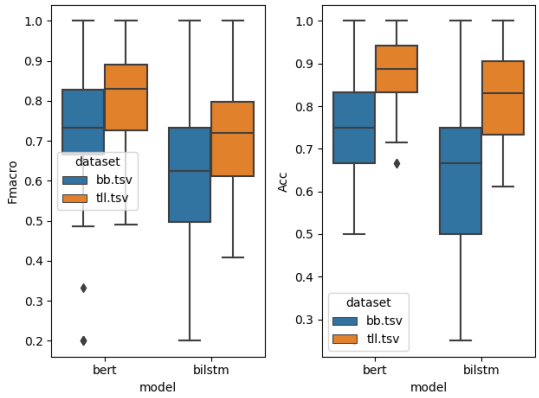

Figure 4: Performance distribution boxplots: F-macro and accuracy across lemmas per dataset per model.

TLL Dataset Due to the data preparation overhead, thus far we processed a subset of 40 lemmas. The WSD performance scores on TLL data are also listed in Table 1: BERT attains a nearly .80 F-score and outperforms the baseline biLSTM model with a large margin (for both datasets). Figure 4 also indicates that the median of the scores for TLL data is higher than for B&B data.

4 Summary and Conclusion

Our study aimed to confirm the impact of Latin BERT (Bamman and Burns, 2020) and to point out an important new Latin WSD resource. We constructed a large dataset from the TLL that holds quotes labeled with the first two highest-level senses of a headword. These likely incorporate senses that the B&B dataset did not include. We experimentally validated that the nested dictionary structure of the TLL is able to deliver WSD data for finetuning contextual representations in a transformer architecture. The WSD models yielded a large improvement above the static embeddings baseline, when evaluated on held-out data from our new, TLL-based dataset. We plan to scale up this study and to release a benchmark dataset and trained models for Latin WSD in future work.

References

David Bamman and Patrick J. Burns. 2020. Latin BERT: A Contextual Language Model for Classical Philology. *CoRR*, abs/2009.10053.

Michele Bevilacqua, Tommaso Pasini, Alessandro Raganato, Roberto Navigli, et al. 2021. Recent trends in word sense disambiguation: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. International Joint Conference on Artificial Intelligence, Inc.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Greta Franzini, Andrea Peverelli, Paolo Ruffolo, Marco Passarotti, Helena Sanna, Edoardo Signoroni, Viviana Ventura, and Federica Zampedri. 2019. Nunc est aestimandum: Towards an evaluation of the Latin WordNet. In *CLiC-it*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Roberto Navigli. 2009. Word Sense Disambiguation: A Survey. *ACM Computing Surveys*, 41(2).

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020. SENSEMBERT: Context-enhanced sense embeddings for multilingual word sense disambiguation. In *Proceedings of The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)*.