A Better Latin Transducer for Apertium

Emily Schalk
Bryn Mawr College

Introduction

Apertium's current Latin transducer lacks the tools to fully tag most Latin word forms, especially:

- most verb forms outside of the present indicative tense and most noun forms outside of the first and second declensions
- the third declension of Latin nouns (only implemented as an irregular form, although this paradigm is highly regular)
- many common Latin stems (the stems present in the Latin transducer at the beginning of this project were few and of limited use in the analysis of most Latin texts)

Identified solutions:

1. Automate stem generation
2. Add more stems
3. Implement more paradigms

A Python script parsed vocabulary lists from Haverford's The Bridge and generated stems

The Bridge's vocabulary is exported as tsv → Python csv module parses file → POS info helps identify forms → Lemmas, saved as strings, are analyzed for familiar patterns → Paradigms are identified based on these patterns → Failed analyses use filler tags for paradigm info → A second script removes entries with filler tags*

*not yet implemented

Evaluation

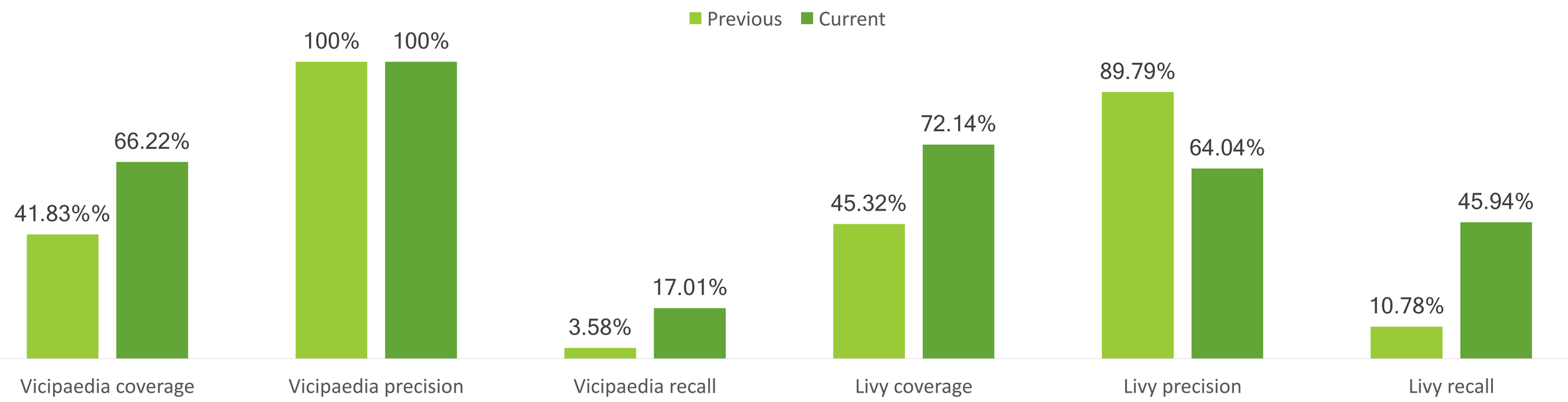
Due to the unusual community of Latin language users, this project's applications are evaluated based on the different ways Latin is used: in neo-Latin literature, and in ancient Latin texts

Neo-Latin

- The Neo-Latin wiki, Vicipaedia, was used as a large corpus on which to test the transducer
- Neo-Latin grammar is identical to classical Latin grammar, but its lexicon is made up of many modern neologisms
- The transducer struggled to analyze the Vicipaedia corpus

Ancient Latin

- The transducer was also tested on an ancient Latin text, Livy's *Ab Urbe Condita*
- Ancient Latin lacks a modern language community, but modern learners may still wish to translate ancient texts
- The transducer was more successful with this corpus



Comparison of evaluations of the old and new Apertium Latin transducers on corpora from Vicipaedia and Livy. It's not yet clear why the precision of the transducer on Livy decreased after this project.

All together, this project added over 17,187 total stems to the Latin transducer (with an estimated 30% error rate, the number of correctly generated stems is still 11,458) and 37 paradigms, increasing coverage by an average of 25%. The expansion of the Latin transducer has brought it closer to functioning as a mature tagger and has made it a more useful and versatile resource for future linguists and translators.