Hybrid and Supervised NER for Old English: Methods, Results, and Replication

Javier Martín Arista
Universidad de La Rioja, Logroño, Spain

Darío Metola Rodríguez
Universidad de La Rioja, Logroño, Spain

Daniel B. Morris
Universidad de La Rioja, Logroño, Spain

Abstract

We study NER for Old English, a setting where spelling drifts, inflection and compounding are pervasive, and annotated material is limited, thus disrupting tokenisation, lemmatisation and span decisions (Pettersson, 2016; Piotrowski, 2012). Using the UD-conformant OEDT resource (≈50k tokens) with BIO splits, we keep the token segmentation and evaluate a unified tagset for five types: PERSON, PLACE, ORGANISATION, RELIGIOUS, and OBJECT. Scoring uses exact-span precision, recall, and F1; token-level F1 is reported.

We compare two systems. The supervised baseline is a Stanza sequence tagger that couples word representations with character-level context and trains directly on the BIO files with fixed UD segmentation (Qi et al., 2020; Lample et al., 2016; Ma & Hovy, 2016). The hybrid method fuses three sources: morphology- and syntax-sensitive priors from UD features and dependencies; a zero-shot pass from a contemporary English transformer NER model with subword-to-token aggregation (Devlin et al., 2019); and a cautious ecclesiastical lexicon targeting religious names and offices. Spans are merged by confidence with a mild preference for longer, coherent mentions and mapped to the unified label set.

The Stanza model achieves 89.69 entity-level F1 on the test set (precision 91.53, recall 87.91) and 89.29 token-level F1. By class, RELIGIOUS is near saturation, PERSON is robust, PLACE lags due to historical toponyms and compounding, and ORGANISATION/OBJECT are constrained by sparsity. The hybrid covers 85% of gold entities with the correct type; among covered items, 78% receive exact boundaries, yielding 66% exact-span recall overall. The morphology-aware cues promote tighter spans, the transformer adds recall but favours short segments, and the ecclesiastical component contributes high-precision matches within scope.

We provide a UD-aligned, reproducible workflow, a strong baseline for Old English NER, and a compact hybrid that improves coverage and boundary fidelity. Remaining gaps cluster in toponyms and institutional names. We outline extensions to byte- or character-level transformers, multi-task learning with morphology, document-level coherence via linking, and broader ParCorOEv3 coverage. We conclude that lightweight morphological priors and conservative domain knowledge complement neural taggers in early English.