Exhaustive Entity Recognition for Coptic: Challenges and Solutions

Amir Zeldes
Georgetown University
amir.zeldes@georgetown.edu

Lance Martin
Catholic University of America
71martin@cua.edu

Sichang Tu
Georgetown University
st1018@georgetown.edu

Abstract

Entity recognition provides semantic access to ancient materials in the Digital Humanities: it exposes people and places of interest in texts that cannot be read exhaustively, facilitates linking resources and can provide a window into text contents, even for texts with no translations. In this paper we present entity recognition for Coptic, the language of Hellenistic era Egypt. We evaluate NLP approaches to the task and lay out difficulties in applying them to a low-resource, morphologically complex language. We present solutions for named and non-named nested entity recognition and semi-automatic entity linking to Wikipedia, relying on robust dependency parsing, feature-based CRF models, and hand-crafted knowledge base resources, enabling high accuracy NER with orders of magnitude less data than those used for high resource languages. The results suggest avenues for research on other languages in similar settings.

1 Introduction

Recent developments in high quality NLP have been likened to a tsunami (Manning, 2015), powered largely by Big Data for tasks such as Named Entity Recognition (NER), and continuous meaning representations in the form of word embeddings, for English and other languages (Upadhyay et al., 2016; Peters et al., 2018). Meanwhile, low resource and historical languages have not been able to take advantage of these advances for several reasons: 1. Gold datasets for most tasks are much smaller – English OntoNotes 5.0 NER (Hovy et al., 2006) has 1.7M words, more than all digitized text available in many ancient languages; 2. Work on embeddings often assumes at least ∼30 million words of training data (Cao et al., 2018);[1] 3. For historical and non-standardized languages, orthographic variation, regional differences, lacunae and other phenomena make learning from unlabeled data exceptionally hard.

In this paper, we report on a series of experiments, including successes and failures, in creating historical language resources for the Coptic language. The main contributions of this work are threefold:

1. We provide a new dataset annotated for entity types, named and non-named, nested entities, and entity linking, i.e. connecting spans of text to a table of specific people, places and other identifiers.
2. We evaluate recent NLP approaches to NER in a low resource setting and show that they do not perform adequately for the needs of (Digital) Humanists.
3. We present an alternative approach relying on dependency parsing, feature-based CRF models and a modest sized knowledge base, which are less labor intensive to produce than million-word training datasets, and offer a way forward for high accuracy NER in a morphologically complex, low resource historical language.

The Coptic language Coptic is the last stage of Ancient Egyptian, written in the first millenium CE in a modified Greek alphabet with additional letters derived from Egyptian Demotic script. Coptic writings

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/.

[1] This limitation has also been recognized in NER work targeting Latin (Erdmann et al., 2016), where models based on word embeddings did not yield competitive results.

Proceedings of LaTeCH-CLfL 2020, pages 19–28
Barcelona, Spain (Online), December 12, 2020.

are abundant, ranging from religious writings, such as hagiographies, homilies, magical texts, and ascetical treatises to social documents including letters, legal documents, and administrative records (Bagnall, 2009; Fournet, 2020). A vast amount of original compositions survive next to translated texts, especially of religious writing. Many early monastic centers wrote extensively in Coptic and many Manichaean and gnostic works survive only or primarily in Coptic. Although it was rarely the most common bureaucratic language, Coptic papyri provide crucial information about official and economic affairs in Egypt during Roman, Byzantine, and Arab rule. As such, Coptic literature is important to general history and the history of Christianity and other religions in and outside Egypt in antiquity.

Coptic grammar presents challenges to automated processing, including agglutinative morphology (fusion of multiple affixes to content words, as in (1)), incorporation (e.g. compound verbs which contain fused verbal arguments, as in (2)), and spelling variation which characterizes many ancient texts, including for widespread Greek loanwords (3).

(1) ⲛⲉ-ⲁ-ⲥ-ⲧⲣⲉ-ⲫ-ⲥⲱⲧⲙ *ne-a-s-tre-f-sōtm* 'she had made him hear'
PRET-PST-she-CAUS-he-hear

(2) ϩⲉⲧⲃ-ψⲩⲭⲏ *hetb-psychē* '(to) soul-kill' = *hōtb* 'kill' + *psychē* 'soul'

(3) ⲕⲟⲗⲗⲩⲃⲉ *kollyᵏ̯e* 'group', misspelled for *kollēgion* (Gr. form of Lat. *collegium*)

In (1), the sequence corresponding to 'she had made him hear' is spelled in Coptic without spaces (or hyphens), meaning segmentation is needed before individual words, and then entities, can be recognized. In (2), a compound verb fuses to 'soul', changing the form of the verb and again requiring splitting. In (3), spelling variation makes the Greek *kollēgion* hard to recognize.

These challenges are not the focus of this paper, and our results below will be based on normalized, gold segmented text. However they mean that in practice, entity recognition will degrade when applied to automatically segmented and normalized text. Since humanists often have very high standards of accuracy, we therefore require robust solutions, as well as possibilities of incorporating semi-automatic correction steps, which we discuss below.

Entity recognition Entity annotation for DH studies can encompass three related but distinct tasks:

- Narrow NER, which identifies words referring to *named* entities, and classifies them as PERson, PLACE, ORGanization etc. Named entities are often assumed not to overlap, possibly creating awkward spans (e.g. [UK]$_{\text{PLACE}}$ [Prime Minister Boris Johnson]$_{\text{PER}}$, in which the place span is outside of the person span, despite belonging to the same syntactic phrase).
- The more exhaustive task of Non-named/Nested Entity Recognition (NNER), including all spans referring to an entity, including overlapping entities (e.g. [[[*LA*]$_{\text{PLACE}}$ *police*]$_{\text{ORG}}$ *chief*]$_{\text{PER}}$).
- Entity Linking, sometimes referred to as Wikification, in which notable entities are associated with a unique identifier, often a corresponding article about the entity in Wikipedia.

For Coptic, we are interested in all three, since texts include unnamed people and places of interest (unspecified 'monks', unnamed locations such as 'the monastery') which scholars may want to find, count and compare in texts. For named entities (esp. people/places), identities are important but difficult to find with string searches, since names repeat frequently ('Johannes' can be John the Baptist, St. John the Evangelist, Apa Johannes the Archimandrite, etc.), and marking up unique identities allows linking resources to data from other projects and languages, increasing their utility and discoverability. In the following, we present a new dataset for Coptic (N)NER and Wikification (Section 2), test out-of-the-box and custom approaches to our tasks (Section 3) and discuss some applications for humanities research using entities (Section 4). Section 5 draws the conclusion and suggests directions for further study.

2 Data

Entity annotation As an underlying dataset we chose the Coptic Treebank (Zeldes and Abrams, 2018) from the Universal Dependencies project (V2.6, `https://universaldependencies.org/`),

which contains close to 50,000 genre-balanced tokens (30K train, and 10K each dev/test) annotated for gold dependency syntax, POS tags and lemmatization, covering both native and translated texts. For automatic annotation using the tools in the next section we use the larger dataset made available by Coptic Scriptorium (approx. 1M words, `http://copticscriptorium.org`) which we aim to make searchable for entity information. We annotate 10 entity types, shown in Table 1 with their proportions in the corpus. Although some of the types, such as events, plants, organizations and animals, are comparatively rare, they are nevertheless distinct and potentially very interesting; for example, events cover turning points in stories, such as a person's death, a war or conquest, famine, etc., while organizations often identify factions in theological and military conflicts (the Catholic Church, Diocletian's army, etc.). When NER is applied at scale, we expect them to be useful in conducting research on the underlying entity types.

entity type	%	examples	entity type	%	examples
ABSTRACT	28.72	'humility'	PERSON	39.92	'all angels'
ANIMAL	1.09	'200 horses'	PLACE	10.87	'Alexandria'
EVENT	2.00	'his death'	PLANT	1.00	'wheat'
OBJECT	9.79	'bottles'	SUBSTANCE	1.43	'water'
ORGANIZATION	0.86	'the army'	TIME	4.31	'ten years'

Table 1: Entity types in our data with examples and percentages.

As an annotation interface we use the version controlled online editor GitDox (Zhang and Zeldes, 2017) , shown in Figure 1, which visualizes entity spans as boxes, color-coded for entity type, and enforces strict entity nesting (entities only overlap fully contained entities), as well as no crossing of sentence boundaries (based on the Coptic Treebank's sentence splits).

Figure 1: Annotation interface.

To assess the reliability of our annotations, we carried out an inter-annotator agreement study by double annotating 1,162 tokens, containing 147 entities after adjudication. The annotators both had college level training in Sahidic Coptic and previous experience annotating Coptic corpora for other categories, such as part-of-speech tagging. Since measuring agreement on nested entities is non-trivial due to overlapping entities, we compute several metrics:

1. Cohen's Kappa, based on a single gold label per token, such that each token reflects the category of its deepest nested span (e.g. "for the army of Diocletian" becomes [O, B-org, I-org, I-org, B-per], since an organization begins at word 2, and continues until a person begins at the last word). This results in 21 possible tags (2 * entity types + 'O').
2. Micro-averaged mutual precision, recall and f-score, taking each annotator as a control for the other. Precision and recall take the perspective of annotator #1, but can equally be reversed.
3. Head accuracy: ignoring spans, what proportion of head nouns are assigned correct types.

The results are shown in Table 2, including scores for span agreement without entity types. Agreement is far above chance, with a typed and untyped kappa of .85 and .9 respectively, falling in the so-called

'perfect' bracket of 0.8–1.0 (Landis and Koch, 1977). Precision, recall and f-scores are harder to interpret, since neither annotator corresponds to a 'ground truth'; however, since we use the same metrics to evaluate systems below, these values help give a ceiling for automatic accuracy in Section 3.

	kappa	F1	precision	recall	head acc
typed	0.859	0.807	0.785	0.829	0.903
untyped	0.902	0.883	0.859	0.908	N/A

Table 2: Results of the inter-annotator agreement experiment.

Comparing typed and untyped metrics, we also see annotators agree much better on spans than on entity types. To understand why, we examine cases like (4)–(5).

(4) mn [hah n-kah] haro-u 'it has no [soil]$_{\text{OBJ/SUBST}}$ under them'
not much of-earth below-them

(5) [nē] ne nt-a-u-jo-ou ejm-p-kah et-nanou-f '[those]$_{\text{PER/PLANT}}$ sown on the good ground'
those are REL-PST-they-sow-them upon-the-earth REL-good-it

In (4), annotators disagreed whether 'earth' is a SUBSTANCE or concrete OBJECT, which is murky in context. In (5), as part of a parable in which Jesus likens people to plants, the PLANT in one analysis, is resolved as a PERSON in the other. It seems likely that NLP tools will err in assigning the more common category to ambiguous words across such contexts. At the same time, the head-based metrics shows that annotators mostly agree on entity types when exact spans are ignored (about 90% of the time). Most disagreements are due to entities omitted by one of the annotators, as in (6), where one annotator ignored 'sky' in 'birds of the sky' as non-referential, while another treated it as a PLACE.

(6) n-halate n-[t-pe] 'The birds of [the sky]$_{\text{PLACE}}$'
the-birds of-the-sky

While not perfect, the analysis suggests that human-quality scores of around 90% would provide a good basis for humanists' work using the entity annotations, with at least some of the disagreements revolving around weakly referential cases, which may be of less interest to researchers (see Section 4).

Entity Linking For entity linking, we followed the broadly used approach of linking mentions of named entities to their corresponding Wikipedia articles (Milne and Witten, 2008; Shnayderman et al., 2019), i.e. Wikification. Using Wikipedia as a table or authorities brings a number of advantages and disadvantages, though we feel the former outweigh the latter substantially. The main advantages are obtaining an existing high quality table of authorities, and the wide range of other projects using Wikipedia identifiers, including resources linked in multiple languages (McNamee et al., 2011). Many notable Coptic person entities are not indexed in other relevant inventories, but do have Wikipedia pages. At the same time, Wikipedia identifiers are also available in the largest subset of projects, including broad projects on antiquity, such as Pleiades, and targeted ones adjacent to ours, such as Syriaca.[2]

Due to the high coverage of Wikipedia (especially for people) and the desirability of re-using common, existing identifiers, we opted to annotate our gold entity dataset with Wikipedia identifiers, which included 610 named entity types, of which 441 were found to have Wikipedia articles, amounting to 104 unique identities (i.e. distinct articles). The remainder consisted of minor entities or unknown/unidentified people mentioned in texts, such as 'Bibrus', an unidentified minor character in the Dormition of John, or Mahlon in the Book of Ruth. We evaluate the feasibility of using these seed annotations for automatic Wikification in Section 3, and give plans for wikifying more data in Section 5.

Availability Our annotations are made freely available online under a Creative Commons Attribution (CC-BY) 4.0 license, matching the license of the Coptic Treebank. To facilitate re-use and interoperability, data is versioned on GitHub in UD's CoNLL-U format, Corpus Workbench format (Christ, 1994), PAULA stand-off XML (Dipper, 2005) and TEI XML (`https://tei-c.org/`).

[2]For example `http://syriaca.org/place/572.html` and `https://pleiades.stoa.org/places/727070` are both aligned with the Wikipedia entry for Alexandria, meaning their entries can easily be aligned with ours.

3 Experiments in Automatic Entity Recognition

Mention detection Before we can evaluate entity classification and linking, we must first find entity span candidates in texts. As baselines we consider: a. using all and only nouns as entity spans (NOUN); and b. using all and only sequences of words attested as an entity in the training data (LOOKUP). The NOUN strategy will only recall single token mentions, and include incorrect non-referring expressions. LOOKUP should have few false positives, but low recall, since novel strings in the test data will be missed.

As competitive solutions we consider two families of methods: 1. In PARSE, entities are assumed to cover the spans of phrases headed by nouns, as identified by the syntax tree; 2. SEQUENCE: sequences of words in the corpus are scanned and classified as entities using a neural sequence tagger. Since we have reference syntax trees for our data, we can evaluate performance with gold and predicted trees from an automatic parser.

As a 'sequence' based system, we train a state of the art NNER system (Yu et al., 2020), which relies on a bidirectional recurrent neural network (RNN) with biaffine attention. The system scores spans based on start and end token indices, considering all possible spans, including nested mentions, and outputs a probability for each span category, or 'no-entity' (='O(outside)' in BIO encoding) for spans not predicted as mentions. The system relies on word embeddings, which we provide using Word2Vec (Mikolov et al., 2013) based on ∼1 million tokens of unannotated, automatically segmented Coptic text from Coptic Scriptorium (Zeldes and Schroeder, 2015) with a vocabulary size of ∼11,000 types and 50 dimensional representations.

	exact span match			**fuzzy head span**		
method	**R**	**P**	**F1**	**R**	**P**	**F1**
LOOKUP	0.386	0.555	0.455	0.591	0.849	0.697
NOUN *(gold tags)*	0.123	0.111	0.117	0.855	0.773	0.812
NOUN *(pred tags)*	0.121	0.107	0.113	0.853	0.756	0.802
PARSE *(gold parse)*	0.879	0.862	**0.870**	0.948	0.929	**0.938**
PARSE *(predicted)*	0.831	0.815	**0.823**	0.941	0.922	**0.931**
SEQUENCE *(10)*	0.463	0.651	0.541	0.611	0.859	0.714
SEQUENCE *(binary)*	0.653	0.732	0.690	0.793	0.725	0.757

Table 3: Results for automatic entity mention span detection, exact span match on the left and fuzzy match containing entity heads on the right.

Table 3 gives scores for baselines (NOUN and LOOKUP) and the competitive approaches (PARSE and SEQUENCE). For NOUN and PARSE we provide separate scores for gold versus predicted POS and trees, using Marmot (Müller et al., 2013) as the tagger, and MaltParser (Nivre, 2009) for parsing. For SEQUENCE, we tested two scenarios: 10-way classification (typed entities) and binary classification (entity/non-entity), which should be easier to learn given the limited data. In all cases, metrics evaluate correct/incorrect boundary detection, ignoring entity types. Fuzzy span scores are more lenient, matching spans that include the entity's lexical head word, even if exact boundaries are incorrect. The results confirm the suspicion that training data and/or word embedding representations are insufficient for high quality results using the neural system. The binary RNN does better by 15%, suggesting training data sparseness may be the main issue. PARSE methods are promising, indicating that investment in a higher quality parser may help. For fuzzy span scores, the small degradation of the parse-based strategy and NOUN using automatic NLP is due to the fact that tagging nouns in Coptic is comparatively easy, especially if we ignore the common vs. proper nouns distinction (both of which usually indicate mentions equally).

Entity classification As spans for entity classification evaluation, we use automatic parser output for all strategies, except for SEQUENCE, which uses the RNN's predicted spans. Scores are assigned for both exact match (span and entity type) and fuzzy match (minimal span containing the head receiving the correct entity type). As a baseline, we select the majority class ABSTRACT for all spans identified by the parser. As a third approach, we apply Knowledge Base (KB) lookup. To create our KB, we annotated

the most frequent 2,700 nouns from Coptic Scriptorium with possible entity types regardless of context; however to keep the experimental setup fair, we use a version of the KB with only those lexical items which are attested in the training set (about 1,300 entries). We note this approach cannot handle novel words which are not included in the KB (for these we guess the majority ABSTRACT), and has no way of disambiguating ambiguous entries (for which we guess the attested majority category from training).

Finally, we test two feature-based approaches, in which a conditional random fields (CRF) model is adopted, using scikit-learn's CRF Suite.[3] The model takes selected features as input, and outputs predicted entity type for each entity's head token position only, taking into account the most probable path of labels through each sentence. Three categories of features are extracted from the input data:

- **Grammatical features:** 1. first/last 2-3 characters of each token, giving access to some morphological affixes (e.g. initial *mnt-* forms abstracts, like English '-ness'); 2. POS tags and dependency functions; 3. syntactic parent: for example subjects of verbs like 'say' are likely to be a PERSON.
- **Numerical features:** 1. descendent span length, i.e. how many words are dependent on the current token, directly or indirectly. 2. percentile position in sentence: humans are often mentioned earlier in sentences, whereas inanimate modifiers tend to occur late; 3. sentence length.
- **Context features:** previous and next tokens and their POS tags and dependency functions.

Beyond testing the CRF classifier as a standalone solution, we also combine it with the KB resources. In this setup, input items found in the KB are classified according to their entries, and the CRF classifier is only consulted in three scenarios:

1. when there are out of vocabulary (OOV) tokens for the KB (i.e. unknown words)
2. when an item has multiple KB entries, we choose the one with the higher CRF classifier score
3. when the CRF classifier is highly confident that an item is non-referential (i.e. predicting the 'O' class with probability $>95\%$), the entity candidate is discarded

	span match			**head match**		
method	**R**	**P**	**F1**	**R**	**P**	**F1**
MAJORITY	0.213	0.209	0.211	0.235	0.230	0.232
SEQUENCE	0.476	0.614	0.536	0.527	0.757	0.621
KB	0.681	0.660	0.670	0.728	0.705	0.717
CRF	0.805	0.778	0.791	0.861	0.831	0.846
CRF+KB	**0.827**	**0.810**	**0.818**	**0.889**	**0.869**	**0.879**

Table 4: Scores for entity type identification. All methods except RNN use spans predicted by the best method for mention detection.

Table 4 provides the results of all models. The RNN model (SEQUENCE) is not competitive, probably due to the limited size of training data and word embeddings. KB gains approx. 14%, showing it covers many more cases. For the CRF model, F1 scores rise to 0.79 and 0.81, indicating the feature-based model is promising. The hybrid approach (CRF+KB) performs best, due to the ability of the CRF classifier to disambiguate uncertain cases and the KB's power to capture rare and unambiguous items (e.g. less frequent categories such as PLANT or EVENT, which the CRF dismisses as unlikely).

[3] A reviewer has asked why a CRF classifier is useful when we cannot use BIO encoding for nested span detection. In fact, in our experiments CRFs outperformed other word-wise classifiers, such as Random Forest and Gradient Boosting, since they can constrain transitions between labels which are helpful even for head word classification. Adjacent words may have plausible but incompatible tags (e.g. for the saint 'Apa Shenoute', both words can be PERSON, but only one should be labeled as the head), and certain transitions, such as inanimate object followed by an animate possessor, can be captured by CRFs.

Wikification For entity linking, our data is too small to use neural approaches with word embedding inputs. Due to the limited training data, many relevant identities will not appear in our data, meaning a large part of the target values for linking are unknown for our system. At present we therefore use a rudimentary semi-automatic strategy, offering possible links to human annotators, who can accept or reject suggestions, and enter new links for entities that appear for the first time. Our lookup strategy uses a heuristically ordered cascade applied to all minimal spans containing a proper noun:

- If the exact entity text is known in other documents in the same corpus, prefer the most frequent link associated with it (e.g. 'John who gives baptism' in the Gospel of Mark is 'John the Baptist')[4]
- Otherwise, if the entity text has appeared elsewhere in the corpus, prefer its most frequent link (exact match for 'John who gives baptism' is better, even if the corpus has other more frequent Johns)
- Else, if the entity's head noun is known in this corpus, prefer its most frequent link ('John' in the Gospel of Mark is most often 'John the Baptist')
- Else, if the entity's head is known anywhere, prefer its most frequent link ('John' might overall most frequently refer to 'John the Apostle' in all sub-corpora)

This cascaded heuristic can only work for proper nouns that appear somewhere in the training data, and is susceptible to a majority bias (i.e. it always guesses that a 'John' in a new corpus is the most frequent John in our data).

To evaluate our strategy, we compare it with two baselines: exact match majority choice (most frequent link matching the entire entity string) and head match (most frequent entity associated with a head noun). We use both the train and dev partitions to build our lookup table, while the test set remains the same, containing some 100 identifiable entity mentions belonging to 53 distinct types.

method	**acc**	**cov**	**no_err**
exact	0.227	0.273	0.953
head	0.433	0.500	0.933
cascade	**0.460**	**0.500**	**0.960**

Table 5: Wikification scores – accuracy (% correct links), coverage (% entities for which a response is retrieved), and % entities with no false links (correct link matched, or entity not covered).

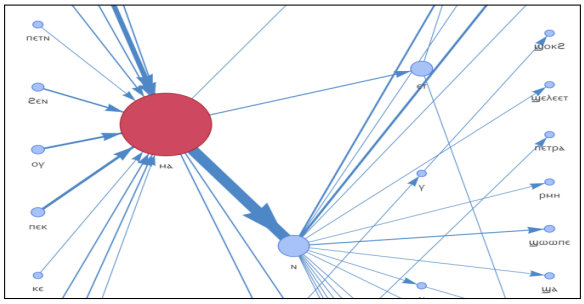
Table 5 shows that the cascade improves on the baselines, and that, while coverage is limited (only 50%), it rarely misclassifies an example. Errors arise due to single word names shared between multiple entities, such as John (the Apostle or the Baptist) or Paul (the Apostle, or Paul of Thebes), and context differences, such as 'Israel', which is linked to 'Kingdom of Israel (united monarchy)' when referring to King David's kingdom, but to 'Israelites' when referring to Israel as a people. We currently feed such predictions to human annotators for disambiguation.

4 Applications

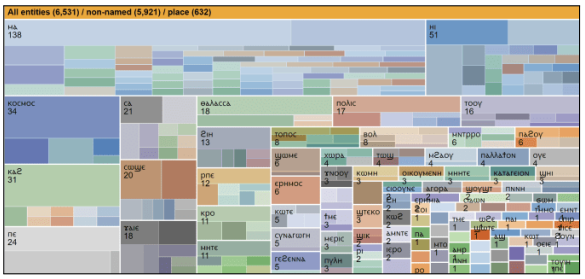
Distant reading Entity information can help address a variety of research questions, as well as making data more accessible and easier to discover in the larger DH ecosystem (Schroeder, 2020). As a first way of looking at entities in Coptic text, we can consider how to visualize the global picture of entity frequencies in our data as a kind of distant reading (Moretti, 2013). Two interactive visualizations we can use for this purpose are entity term networks and recursive TreeMaps (Shneiderman, 1992).

Entity term networks visualize a head word's relationships with other words in its entity spans. Figure 2a captures part of the network for ⲙⲁ *ma* 'place'. Larger nodes represent more frequently recurring terms, and broad arrows correspond to frequent transitions. The network for ⲙⲁ gives us a clearer idea of its potential semantic relationships: often preceded by ⲡⲉⲕ/ⲡⲉⲧⲛ *pek/petn* 'your', almost always followed by ⲛ *n* 'of', continuing to nouns indicating purpose ('place of dwelling', or 'lavatory' with ⲣⲙⲏ *rmē* 'urination'), events (ϣⲉⲗⲉⲉⲧ *šeleet* 'wedding'), directions (ϣⲁ *ša* 'East') and more. Similarly, the

[4]This is relevant e.g. because the Treebank includes only chapters 1–9 of Mark, but we want to annotate entire works.



(a) Entity Network for ⲙⲁ *ma* 'place.'



(b) TreeMap of non-named place entities

Figure 2: Distant reading visualizations.

TreeMap concisely depicts the most mentioned entities and headwords in the corpora. Following our annotation guidelines, the TreeMap initially separates named and non-named entities. These are divided into the ten entity types, and then by unique entity head words. Figure 2b shows the view of non-named place entities. The TreeMap shows that places headed by ⲙⲁ (top left, 138 cases) are most frequent, followed by ⲏⲓ *ēi* 'house' (51) and ⲕⲟⲥⲙⲟⲥ *kosmos* 'cosmos, world' (34). In lower ranks, the data evinces the importance of the desert in our corpus. Many texts center on monks who have left civilization, and their setting is often the ⲭⲁⲓⲉ *jaie* 'desert' (18). Consequently, we see our texts mention ⲭⲁⲓⲉ more than ⲡⲟⲗⲓⲥ *polis* 'city' (17), and even more so if we include Greek synonyms such as ⲉⲣⲏⲙⲟⲥ *erēmos* 'desert' (6). These visualizations allow easy interactive exploration and show patterns that may be missed when reading individual texts. Without entity annotation, such phrases cannot be trivially extracted and categorized for comparison.

Entity type proportions Another comparable quantity across texts is the proportion of named vs. non-named entities, or proportions of entity types. For the latter, we observe that sermons in the treebank (e.g. Pseudo-Athanasius, Letters of Besa) have much more abstract entities than narratives, since they concentrate on instruction, using abstractions to communicate their message and mentioning people less often. Narratives mention people more frequently, with a person/abstract ratio of 2:1 or higher, compared to ∼1:1 in sermons. Drilling down in more detail, we find exceptions such as The Life of Onnophrius and The Dormition of John: these have person/abstract ratios close to 1:1, mainly due to homiletic speeches delivered by the main characters, echoing Coptic sermons as they instruct disciples to avoid sin. This quantitative finding foregrounds an interesting commonality between seemingly unrelated texts and differences between texts from one genre.

We imagine many more findings will emerge from Coptic entity annotations, which complement detailed philological, literary, and historical inquiries. Data aggregation and visualization of different subsets of texts enable analyses based on the quantity, proportion and dispersion of entity types which we are only beginning to explore. They abstract away from individual ways of phrasing references to people and places, while linking mentions of named entities across datasets, projects, and DH tools.

5 Summary and conclusion

In this paper we presented a new annotated data set for Coptic entity classification and linking, using ten entity types, including named and non-named, potentially nested entities, and attaching named entities to corresponding Wikipedia articles, i.e. Wikification. Our annotated data represents a wide range of genres, including translated and autochthonous texts, and is freely available in a number of popular formats, including the CoNLL and TEI XML formats, under an open license, as are our tools.[5]

From a technical perspective, our results demonstrate the difficulty in applying state-of-the-art neural frameworks to nested NER in languages with modest data sizes. At the same time, the lack of availability of many millions of words for training word embeddings limits the utility of RNN architectures relying on highly informative, context sensitive information. Instead, we are able to show that a syntax-based

[5]Available at: `https://github.com/CopticScriptorium/corpora`, including automatically annotated silver data. Gold entity annotations have also been merged into the release of the UD Coptic Treebank, at `https://github.com/UniversalDependencies/UD_Coptic-Scriptorium`.

approach using dependency trees to identify nested noun phrases is viable and substantially more accurate for data in the several 10K-100K tokens range. Although our approach relies on the existence of such syntactically annotated data to train a parser, the size of the data set is likely to be a more realistic target for projects in similar settings to ours than the size of datasets underlying standard approaches to (N)NER in modern languages. For comparison, the Universal Dependencies treebanks for Ancient Greek and Latin include over 400,000 and 800,000 tokens respectively, meaning that the treebank used here is very modestly sized; and even with larger treebanks, those languages too lack highly expressive word embeddings based on hundreds of millions, or even billions of words, which are commonly available for modern European languages. Our methods therefore show promise for mention detection in annotating other resources for classical languages, such as freely available Ancient Greek and Latin texts available in the Perseus Digital Library (Smith et al., 2000).

For entity classification, our approach shows the robustness of feature based classifiers (Section 3), while also revealing the added value of a knowledge base (KB) architecture. A KB derived strictly from the training data is already helpful for the hybrid approach taken here (KB+CRF in Section 3), while an even broader coverage KB can be constructed with relatively low effort, which makes it possible to capture some of the rarer, but often lexically unambiguous entity types, such as names of animals or plants. For entity linking, we show modest results in terms of recall, but quite good precision which can facilitate larger manual annotation efforts by offering reliable suggestions for human review.

In Section 4 we outlined some of the applications that wide-coverage entity annotations can bring for humanists, including style and genre studies, highlighting differences between documents that are otherwise similar from the text type perspective, and bird's eye-view visualizations which allow us to examine how texts talk about entities, and how often. The data used for the example studies in this paper comes from the high quality, but small, manually annotated corpus prepared for this work. We plan to publish a much larger automatically annotated corpus of Coptic texts using the tools described here, which we expect to deliver much more comprehensive capabilities in exploring the contents of Coptic texts, many of which still have no translations that might allow exploring the content in English. Such larger scale data could also be used to link to other projects featuring entity identifiers (in Coptic or otherwise), and to lexicographic projects, such as the Coptic Dictionary Online (Feder et al., 2018), or the Database and Dictionary of Greek Loanwords in Coptic (DDGLC) (Almond et al., 2013).

6 Acknowledgments

We would like to thank Coptic Scriptorium contributors Mitchell Abrams, Elizabeth Davidson, Rebecca Krawiec, Christine Luckritz Marquis, Elizabeth Platte, Dana Robinson and Caroline T. Schroeder for their work on the annotated data, as well as the anonymous reviewers for their feedback. This work was supported by a Stage III Digital Humanities Advancement Grant from the National Endowment for the Humanities (grant HAA-261271-18).

References

Mathew Almond, Joost Hagen, Katrin John, Tonio Sebastian Richter, and Vincent Walter. 2013. Kontaktinduzierter Sprachwandel des Ägyptisch-Koptischen: Lehnwort-Lexikographie im Projekt Database and Dictionary of Greek Loanwords in Coptic (DDGLC). In Ingelore Hafemann, editor, *Perspektiven einer corpusbasierten historischen Linguistik und Philologie*, pages 283–315, Berlin. BBAW.

Roger S. Bagnall. 2009. *Early Christian Books in Egypt*. Princeton University Press, Princeton, NJ.

Yixin Cao, Lei Hou, and Juanzi Li Zhiyuan Liu. 2018. Neural collective entity linking. In *Proceedings of COLING 2018*, pages 675–686, Santa Fe, NM.

Oliver Christ. 1994. A modular and flexible architecture for an integrated corpus query system. In *Proceedings of Complex 94. 3rd Conference on Computational Lexicography and Text Research*, pages 23–32, Budapest.

Stefanie Dipper. 2005. XML-based stand-off representation and exploitation of multi-level linguistic annotation. In *Proceedings of Berliner XML Tage 2005 (BXML 2005)*, pages 39–50, Berlin, Germany.

Alexander Erdmann, Christopher Brown, Brian Joseph, Mark Janse, Petra Ajaka, Micha Elsner, and Marie-Catherine de Marneffe. 2016. Challenges and solutions for Latin named entity recognition. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 85–93, Osaka, Japan.

Frank Feder, Maxim Kupreyev, Emma Manning, Caroline T. Schroeder, and Amir Zeldes. 2018. A linked Coptic Dictionary Online. In *Proceedings of LaTeCH 2018 - The 11th SIGHUM Workshop at COLING2018*, pages 12–21, Santa Fe, NM.

Jean-Luc Fournet. 2020. *The Rise of Coptic: Egyptian Versus Greek in Late Antiquity*. Princeton, Princeton, NJ.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% solution. In *Proceedings of NAACL 2006, Companion Volume: Short Papers*, pages 57–60, New York.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Christopher D. Manning. 2015. Computational linguistics and deep learning. *Computational Linguistics*, 41(4):701–707.

Paul McNamee, James Mayfield, Dawn Lawrie, Douglas Oard, and David Doermann. 2011. Cross-language entity linking. In *Proceedings of IJCNLP 2011*, pages 255–263, Chiang Mai, Thailand.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations (ICLR) Workshop*, Scottsdale, AZ.

David Milne and Ian H. Witten. 2008. Learning to link with Wikipedia. In *Proceedings of the 17th ACM conference on Information and Knowledge Management*, pages 509–518, Napa Valley, CA.

Franco Moretti. 2013. *Distant Reading*. Verso, London.

Thomas Müller, Helmut Schmid, and Hinrich Schütze. 2013. Efficient higher-order CRFs for morphological tagging. In *Proceedings of EMNLP 2013*, pages 322–332, Seattle, WA.

Joakim Nivre. 2009. Non-projective dependency parsing in expected linear time. In *Proceedings of ACL 2009*, pages 351–359.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL 2018*, pages 2227–2237, New Orleans, LA.

Caroline Schroeder. 2020. Coptic literature in context (4th-13th cent.): Cultural landscape, literary production, and manuscript archaeology. In *Understanding Space and Place through Digital Text Analysis*, pages 229–242. Edizioni Quasar, Rome.

Ilya Shnayderman, Liat Ein-Dor, Yosi Mass, Alon Halfon, Benjamin Sznajder, Artem Spector, Yoav Katz, Dafna Sheinwald, Ranit Aharonov, and Noam Slonim. 2019. Fast end-to-end Wikification. arXiv:1908.06785[cs.CL].

Ben Shneiderman. 1992. Tree visualization with tree-maps: 2-d space-filling approach. *ACM Transactions on Graphics*, 11:92–99.

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

Shyam Upadhyay, Manaal Faruqui, Chris Dyer, and Dan Roth. 2016. Cross-lingual models of word embeddings: An empirical comparison. In *Proceedings of ACL 2016*, pages 1661–1670, Berlin.

Juntao Yu, Bernd Bohnet, and Massimo Poesio. 2020. Named entity recognition as dependency parsing. In *Proceedings of ACL 2020*, Seattle, WA.

Amir Zeldes and Mitchell Abrams. 2018. The Coptic Universal Dependency Treebank. In *Proceedings of the Universal Dependencies Workshop 2018*, pages 192–201, Brussels.

Amir Zeldes and Caroline T. Schroeder. 2015. Computational methods for Coptic: Developing and using part-of-speech tagging for digital scholarship in the humanities. *Digital Scholarship in the Humanities*, 30(1):164–176.

Shuo Zhang and Amir Zeldes. 2017. GitDOX: A linked version controlled online XML editor for manuscript transcription. In *Proceedings of FLAIRS-30*, pages 619–623, Marco Island, FL.