Combating Data Sparsity in Low Resource Languages and Specialized Domains

Introduction
Data sparsity is a complex problem that manifests in different forms depending on the task at hand. Modeling a low resource language (LRL) requires one to confront the traditional challenge of data sparsity in that there simply may not be enough training data available to cut through the noise and identify relevant, significant correlations. On the other hand, even in languages with as many linguistic resources as English, data sparsity issues arise when trying to process highly specialized domains which exhibit lexical or grammatical traits uncommon elsewhere. Our work presents a number of potential solutions for addressing the challenges of data sparsity on both of these fronts.

As linguists and historians, we are concerned with modeling ancient networks by compiling information on who wrote about whom, where, in what context, at what time, and in what language throughout history, and using that information to build a database presenting the global evolution of social networks from the point of view of the ancients who were part of them. The building blocks of such a database would be ethnonyms representing ancient peoples and groups. To accomplish this, we are using a Natural Language Processing technique called Named Entity Recognition (NER) to automatically identify the names of groups, people, and places mentioned in a wide variety of historical sources, ranging from formal prose, to poetry, to digitized records of inscriptions, graffiti and engravings. While many of these sources do exist in English translation, we have found that such translations represent enough of a specialized domain that we encounter data sparsity problems. Furthermore, not all of the material we mean to incorporate in our database has been translated into English. Thus, starting with Latin material and eventually incorporating other historical languages, our goal is to test the nature of data sparsity in a number of ways in order to develop solutions that enable us to effectively perform NER on the diverse source material we plan to include in our database.

This paper explores three major steps toward accomplishing this. First, we set about determining whether data sparsity introduced by the specialized domain of translations from a historical language to English is more detrimental or less detrimental than the data sparsity we face when conducting NER in an LRL like Latin. Next, we address specialized domains within the LRL (i.e. poetry or epitaphs) by outlining a supervised domain adaptation technique that allows gradient domains to distinguish general features from domain-specific features, thus maximizing the amount of useful information available for processing such domains while minimizing the interference of non-generalizable information. Last, we investigate some pioneering approaches to unsupervised learning which allow inferences to be made without the aid of annotated training data, thus minimizing the effect of data sparsity. For example, there is a lot of digitized text available in Latin and Ancient Greek, but very little of it is annotated. Unsupervised learning provides a means of teaching the model to leverage this data the way a child acquires its native language, by learning to comprehend the language without overt instruction on how to interpret each token. When complete, the NER system we develop will be capable of processing multiple different LRL’s over multiple different domains varying in style, register, and dialect, without requiring a lot of manually annotated training data. Such a system will be capable of recognizing named entities accurately enough over a broad enough range of texts that we will be able to reconstruct the evolution of these ancient networks we are looking to study.

Choosing to Work with LRL’s as Opposed to English Translations
To compare the issue of data sparsity in performing NER on English translations of historical texts with the same issue in performing NER on the original LRL, we considered a number of factors. Working with English translations would allow us to utilize better NLP resources like larger volumes of annotated data and gazetteers as well as better part-of-speech (POS) taggers. Furthermore, English's
morphological simplicity enables NER systems to perform accurately with only a small set of mostly lexical, type-based features (Finkel et al., 2005), yet if the NER system is frequently exposed to new word types in the test set, such a sparse feature set would have more difficulty determining named entity status than would a feature set utilizing more complex semantic and morphological features. We were particularly concerned about the complications of such specialized domain data sparsity as we suspected that each new historical text would likely mention enough esoteric named entities to confound classifiers relying on primarily type-based feature sets. To investigate this, we used the Stanford NER classifier to train one model on a portion of the English translation of Herodotos' Histories which David Smith had annotated with group names, and another model on the first 90% of the original Latin text of Caesar's Gallic Wars (GW) which we had annotated for person, group, and place. Then we tested both models on the corresponding version (English translation or Latin original) of the remaining 10% of GW. Of the 21 unique group names mentioned in the test set, the English model recognized 14 and spuriously generated 2, whereas the Latin model recognized 20 and spuriously generated 9.

This led us to decide that pursuing NER in the original source language was the smarter enterprise for a number of reasons. Even though the different training sets suggest that the accuracy could have been higher for the English translation system had it been trained on GW as well, the standard features used by Stanford's NER software are designed to optimize performance in English, meaning that there is a higher ceiling for working in Latin if we develop language specific features. Furthermore, our small experiment suggests that English translations of historical texts do indeed represent somewhat specialized domains, meaning that data sparsity will be an issue regardless of the language we choose to work in. Our biggest reason though, for doing NER in the original LRL's, is that we are interested in reconstructing the social networks of ethnic groups as they were portrayed by writers of the era, not as they are portrayed by modern historians. It is thus more fitting that we use those writers' own words to do so and not risk losing information in a translation which inevitably bears the bias of its translator. That being said, using aligned translations to pull features from both the English and the original text might prove a worthwhile avenue for future work.

**Domain Adaptation**

In order to address the data sparsity stemming from specialized domains within an LRL (a yet ongoing endeavor), we started by identifying a source domain, Latin historiography, which exemplifies a sort of median style and content of all the domains from which we want to extract named entities. Within the source domain, we chose to annotate GW for training data because, for one, it includes a large number of named entities typical of the information we seek to include in our database, and two, Caesar's writing style has long been considered a model of Latin prose. Therefore, a model trained on GW could generalize to other Latin texts better than a model trained on a more obscure text. Even so, we still need to be able to tag named entities accurately in a variety of texts that systematically differ from GW along a number of dimensions. While the goal is to treat this variation as continuous, at the outset, the domain adaptation problem had to be addressed by grouping this continuous variation into specialized domains. Hence, while GW represents the source domain of Latin historiography, Ovid's Ars Amatoria would fall into a target domain of Classical Latin poetry. The problem then becomes how the model recognizes what domain it is dealing with and how it adapts what it has learned about the source domain to address the variation in the target domain without compromising accuracy any more than necessary.

Here we resort to the solution of Daume (2007) by manually determining the domain of each text and

---
1 We also used the off-the-shelf version of the Stanford NER classifier as a third model though it performed far worse than the other two presumably because, despite being large in volume, its base of annotated training data was very different from the test set in terms of content and style.
annotating a minimal amount in each target domain (we start with about 5,000 words). Then we augment the feature space such that each feature is listed twice – once as a general feature and once as a feature specific to its unique domain. Thus, if the model comes across the abbreviation “M.” (for the name, “Marcus”) as the first word in a sentence in a text that has been categorized as belonging to the target domain of epitaphs, then the model will read in the following features (among others): general-firstWordInSentence, epitaph-firstWordInSentence. Since the general-firstWordInSentence feature fires for every sentence-initial word in every domain, this feature will not be highly correlated with named entity status. However, since the epitaph-firstWordInSentence feature only fires for sentence-initial words in epitaphs and epitaphs are more likely to begin a sentence with a personal name, the model will learn to weight this feature more heavily, increasing the ceteris paribus probability that it will label sentence-initial “M.” as a personal name in the domain of epitaphs. More generally speaking, Daume (2007) allows us to distinguish which features are generally useful in NER in all Latin domains from those which are useful in poetry, from those useful in classical Latin prose and so on. Such a method further provides a foundation for a quantitative metric of stylometry by answering the question of how different the specific feature sets and corresponding optimized weights are from domain to domain (McClosky et al., 2010). This allows us to compute a sort of stylometric distance from say, Classical Latin poetry to epistulae. We can then use this metric to build from our original structure relying on predetermined, discrete domains to one that actively learns to adapt to continuous stylometric variation. To do this, we reimagine our previously defined domains as fixed-point anchors in multidimensional space, no longer housing any internal variation but representing both the average distribution of the occurrences of features in texts recently assigned to that domain, and the optimum feature-weighting scheme for such a distribution. Then, regardless of whether or not it had previously been assigned a domain membership, every text is assigned coordinates in the same multidimensional space according to its own distribution of features. Lastly, the feature-weighting scheme is optimized uniquely for each text, being most heavily influenced by whatever anchor it lies closest to and least influenced by the farthest anchor. Thus, if the epistulae of Pliny the Younger exhibit some qualities of classical poetry even though they mostly behave like other Latin epistulae, we can capture this non-discrete variation without having to create an entirely new target domain and annotate the requisite training data for it.

Combating Data Sparsity with Morphological Features that Decrease Reliance on Supervised Training

In addition to building toward a discrete system for domain adaptation, we are working on other strategies to minimize the need for supervised learning requiring annotated training data. Because Latin is a morphologically complex language, any successful NER model will need a more sophisticated set of morphological features in order to process it than what is required for English. While there does exist a high quality Latin morphological analyzer (Whitaker, 1993) which provides every possible morphological analysis for a given input word, using every possible analysis as a feature introduces too much noise to our NER model to improve accuracy. Additionally, there are a few Latin part-of-speech (POS) taggers which attempt to use limited available training data to syntactically analyze sentences, though their accuracy is low enough that incorporating their output as features also hurts accuracy. However, taking inspiration from Farber et al (2008) which confronts a similar data sparsity issue in processing Arabic, another morphologically rich language, we are using a combination of different POS taggers and filtering their results by the allowable analyses offered by the morphological analyzer in order to reduce noise. Preliminary results show that this does indeed increase our accuracy on the GW test set. The next step then is to move away from relying on annotated data by incorporating information theoretic metrics for information density like those outlined in Levy & Jaeger (2007). Assuming information density is correlated with preferences over various morphological and syntactic structures, the differences in information densities between familiar domains and a new one can be used to anticipate the prevalence of such structures in the new domain, enabling the model to better recognize constructions likely to contain named entities. More work on this front is forthcoming.
Combating Data Sparsity with Semantic Features that Decrease Reliance on Supervised Training

Farber et al. (2008) also points out two lexical/semantic issues in performing NER on Arabic that are just as problematic in Latin. One involves modeling meaning and the other involves distinguishing boundaries of multiple word expressions (MWE). To address the first issue, they suggest that prefabricated gazetteers could capture the meanings of related words, however, we opt for an unsupervised method to minimize the effect of data sparsity. Leveraging unannotated data, we generate gazetteers automatically based on word co-occurrence by implementing Latent Semantic Indexing (LSI) as described in Manning & Schütze (1999), a method inherently adaptable to different domains. Thus, when the model comes across a morphologically unique word, it can still propose a gazetteer membership and use that along with the sequence of memberships in the preceding and following words to infer its named entity status.

As for the common failure to recognize named entities that span multiple words in their entirety, we address this by modifying collocation extraction methods from Bouma (2009) to again leverage unannotated data to identify multi-word expressions and use them in identifying the boundaries of named entities. Bouma (2009) also enables us to utilize semantic information to recognize MWE's that are not entirely comprised of a named entity, but still contain one. For example, consider the mis-tagging of *Pictonum* (the genitive plural of the name of the group known as the Pictones) as a person. This occurred because the word type never appeared in the training data, so we have no type-based features for it, and it confused our morphological features enough that they failed to parse out the exponent *-um* as being indicative of plurality – a morpho-syntactic property more highly correlated with groups than individual persons or places. Even so, one might use the two words preceding *Pictonum*, which are *in fines*, “in the territory”, to infer that a group is more likely to follow than a person, yet introducing previous word as a feature overwhelsms the model with noise given our dearth of annotated training data. Bouma (2009) however enables the model to recognize that while *in fines* *Pictonum* may not itself be a significant collocation, *in fines* does exist as the first two words in a lot of tri-gram collocations where the third word is a group, and therefore, *Pictonum* is likely a group as well. Like LSI, collocation extraction enables the model to make sophisticated generalizations from abstract information, even in the presence of data sparsity, thereby improving its ability to make conditioned statistical inferences.

**Conclusion**

Historians may argue that even in overcoming the challenges of data sparsity, some degree of error will always be present in automated classification relying on statistical inference – that such error is unforgivable and qualitative analysis is the only way to construct the type of database we wish to build. To that we say that yes, there will be some loss of precision in using automated NER, but if we attain a certain level of accuracy, our errors will become more insightful than problematic. Consider the mis-tagging of *Commio Atrebate*, “Commius the Atrebatan” as a first-name-last-name sequence. To what extent is this error harmful to our database and to what extent is it actually indicative of how group names are capable of functioning as personal names in Latin historiography? Our errors thus enable us to ask the linguistic question of what behaves like ethnohistory and not just simply what is ethnohistory. In addition to Linguistics, the database we build will impact a wide range of fields including stylometry, anthropology, history, and archaeology. Our goal is not to replace the need for qualitative analysis in these disciplines, but to illuminate previously under-investigated historical interactions among ethnic groups and invite thorough, qualitative research to improve our understanding of them.
References


