footnotes\_parser

Lit Review Notes

# John Unsworth

## Scholarly Primitives: what methods do humanities researchers have in common, and how might our tools reflect this? (2000)

* Linking/referring as a scholarly primitive
* Used additively alongside other parts of the research workflow (including, but not limited to discovery, sampling, selection, representation, etc.
* Importance of stability in reference
* Questions/Comments:
  + This paper captures some of the contemporary concerns about citation/referencing – it needs to be stable, transparent, etc. In the process, it makes explicit certain parts of research that historical documents especially tend to ignore/not visibilise. It helps us see what’s missing in the colonial citation chain.

# Matteo Romanello

## Exploring Citation Networks to Study Intertextuality in Classics (2016)

* Annotating entities **but also** relations between entities:  
  Citation as a relation existing between two entities, where one must be the indication of the citation’s scope
* Method to extract citations:
  + Extracting named entities (REFAUWORK, REFSCOPE, AAUTHOR, AWORK) through ML approach
  + Extracting relations between named entities (SCOPE) through rule-based approach
  + Disambiguation of extracted named entities and relations using unique identifiers (CTS URNs) + normalisation of citation scopes  
    Question: Was this disambiguation done manually?
* Method to create network visualisations:
  + Visualisation-1  
    Three-level network (macro: ancient authors; meso: ancient works; micro: text passage) with bipartite nodes (APh works + ancient authors/works/passages) and directed edges (citing document to cited document, i.e., APh work to ancient author/work/passage) arranged using a force-layout algorithm (repulsion-gravity holding nodes within view; attraction between connected nodes)
  + Visualisation 2  
    One-mode (only ancient authors/works) network with undirected edges (connecting authors/works cited in the same APh document)
  + Visualisation 3  
    One-mode (only APh documents) network with undirected edges (connecting documents that cite the same ancient text passage)

## Creating an Annotated Corpus for Extracting Canonical Citations from Classics-Related Texts by Using Active Annotation (2013)

* Zotero 20, 21
* Footnotes parser 23, 24
* Method for annotation: Active Annotation
  + Active Learning paradigm: the more informative the training examples, the higher the model’s performance  
    (Useful when there is lots of unlabelled data)
  + Uncertainty sampling method based on least confident strategy used to identify those instances that are hardest to classify for a statistical model  
    An instance is added to the training set when the Confidence Interval for one or more of its tokens is above a given threshold; CI is the difference between probabilities of two best labels for a given token, so it tells you how confident the model is about the assigned label (large difference means the best label is very probable)  
    Alternatives: entropy
  + Active Annotation is an iterative process and stops when performance does not improve between two subsequent iterations  
    Iteration performance is compared using F1 measure (before and after addition of a set of candidates to the training set)
  + 100 seed instances (partly random/partly manual): 6.4k/330k tokens in training set  
    2k tokens in test set  
    ↓  
    CI ≤ 0.2  
    Each round of active annotation:  
    Records containing token(s) with CI under the threshold were added to candidate set, which is de-duplicated and 30 highest scoring records are sent for manual correction of automatic annotation before adding to training set  
    No IAA recording  
    Test set size also increased at regular intervals using same method to avoid overfitting  
    ↓  
    9 rounds of Active Annotation  
    Re-tokenisation (initially whitespace tokenised because abbreviations were getting messy; Abbreviations were then extracted from corpus and provided to tokeniser to allow for POS-tagging-friendly tokenising)
* Method for training:
  + Supervised training using Conditional Random Fields model (conditional probability instead of joint probability) using CRF++ (C++) implementation  
    Model alternatives: Support Vector Machines (SVM), Maximum Entropy  
    Implementation alternatives: Wapiti (Python), MALLET (Java)
  + 2 sub-tasks: NE classification and NE resolutions
  + Feature set for training:
    - Linguistic features (3 words on either side of token)
    - Orthographic features (different kinds of punctuation, bracketing, case, number, sequence pattern)
    - Semantic features (tokens were compared to classical authors and works dictionary which was converted into a Python suffix array for this purpose)
    - POS tagging was done at the end once the model was built because of tokenisation choices

# Andrew Piper, Chad Wellmon, and Mohammed Cheriet

## The Page Image: Towards a Visual History of Digital Documents

* Possible method (or model even?) to extract footnotes from historical documents