Towards a Tool for the Automatic Extraction of Canonical References

Matteo Romanello[1]

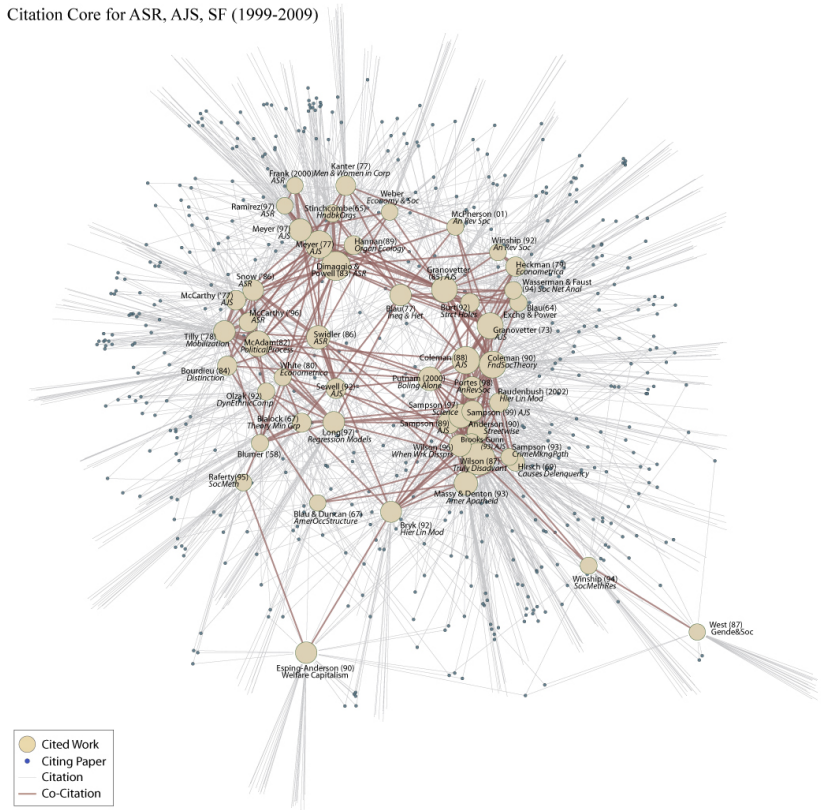
[1]Centre for Computing in the Humanities
King's College London

Digital Classicist and Institute of Classical Studies Seminar 2010
25th June 2010

Overview

- possible applications
- what's out there about canonical references?
- some Jargon explained
- (live) demo

Citation and Co-citation Networks



Caption: Data compiled from Web of Science. Citing papers include all those published in ASR, AJS, and SF since 1999 that cited one of the most-cited works. Cited works are the 53 pieces most cited by papers published in ASR, AJS or Social forces since 1999. Node size is proportaion to number of cites recieved. Edges without nodes are from papers that cite only the target node, all other citing papers cite at least two of the 53. Citation ties indicate that the source paper cites the cited work, co-citation links are the number of times two cited works are jointly cited by a single paper.

http://orgtheory.wordpress.com/2009/08/14/sociologys-citation-core/

Co-citation Network Analysis of Religious Text

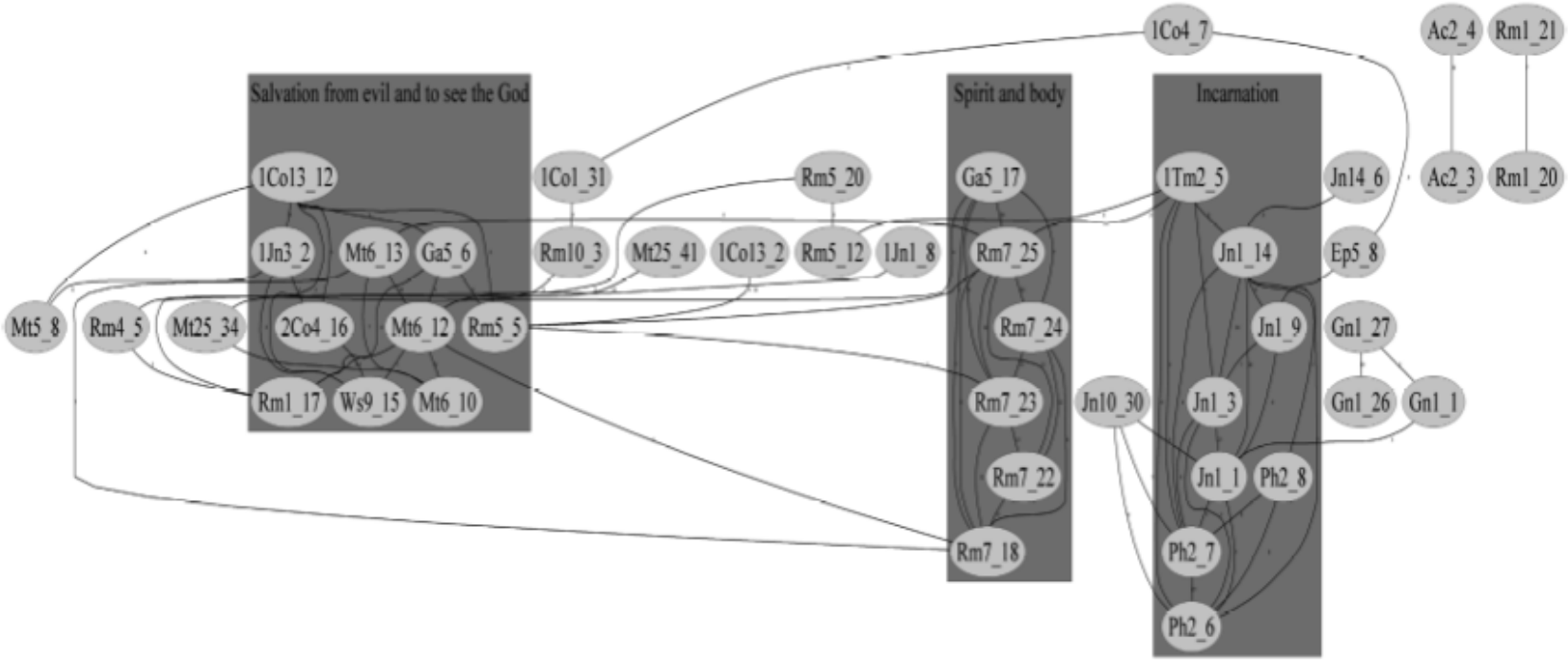


Fig. 4. Clustered Co-citation Network (Augustine)

Hajime Murai, Akifumi Tokosumi: Extracting Concepts from Religious Knowledge Resources and Constructing Classic Analysis Systems. LKR 2008: 51-58

Canonical References

Definition

- Abbreviations used to refer to **primary sources** (i.e. works of ancient authors)
- refer to the research object itself (i.e. text passage)
- logical instead of physical citation scheme (e.g., chapter/paragr vs. page)

Example

Hom. Il. XII 1
Aesch. 'Sept.' 565-67, 628-30; Ar. 'Arch.' 803
Hes. fr. 321 M.-W.
Callimaco, 'ep.' 28 Pf., 5-6

Problem definition

My Goal

Automatic identification of Canonical Rereferences within (plain) texts

1. extraction (manual | automatic)
2. semantic markup (TEI, HTML + microformats, OpenURL etc.)
3. linking to other resources (from citation to actionable link)

Related Work

- (TuftsU) Perseus DL's navigation tools
- (HarvardU) Canonical Text Services (CTS): example, project page
- (CornellU+APh) Classical Works Knowledge Base (CWKB): demo
- Reference Linking to primary sources as value added service for e-journals (screenshot)

Possible Encodings

- HTML + Microformat (Plut. Sol. XIX 1)

```
1  <a class="citation" target="_blank" href="[...]">
2  <cite class="ctref">
3  <abbr class="ctauthor" title="urn:cts:greekLit:tlg0007">Plut.</abbr>
4   <em>
5  <abbr class="ctwork" title="urn:cts:greekLit:tlg0007.tlg007">Sol.</abbr>
6  </em>
7  <abbr class="range" title="19.1">XIX 1</abbr>
8  </cite>
9   </a>
```

- OpenURL for (Aeschylus, Supplices 40-57)

```
1  http://cwkb.org/in/r.php?ctx_ver=Z39.88-2004&rft_val_fmt=info:ofi/fmt:kev:mtx:
       canonical_cit&rft.work-id=tlg0085.014&rft.auform1=Aeschylus&rft.
       titleform1=Supplices&rft.slevel1=40&rft.elevel1=57&rfr_id=info:sid/aph
```

- CTS URN for (Ath., Deipn. I) link

```
1  urn:cts:greekLit:tlg0008.tlg001.fhg01:1.1.1-1.1.9
```

Possible approaches to the problem

1. Rule-based parsing (e.g. RegExps)

Example

```
// tokenizer written in ANTLR syntax

/* Lexicon definition */
SPACE : ('␣'|'\n')+ {skip();};
NON_SPACE :(~('␣'|'\n'))+;

/* Grammar definition */
text : (SPACE|NON_SPACE)+;
```

2. Machine Learning-based

Machine Learning (ML) Jargon

- Machine Learning
- Training / Test set
- instances
- Gold standard
- Sequence Labelling
- Baseline

Named Entity Recognition

- shared tasks: CoNLL2002 and 2003 on Language Indipendent NER
- IOB format for annotated texts/corpora

Statistical NER

Use of statistical model as an approach to this task

Performance Measures

- Accuracy $acc = \frac{tp+tn}{tp+fp+tn+fn}$
- Recall $r = \frac{tp}{tp+fn}$
- Precision $p = \frac{tp}{tp+fp}$
- F-score (aka F-measure) $fscore = 2\frac{accuracy*recall}{accuracy+recall}$

Example

```
list=['a',1,'b','c',10,2,4]
# filter integers only
out=[10,2]
# P = 1 (100%), R = 0.5 (50%), A = 0.71 (71%), F1 = 0.66
    (66%)
```

My experiment

CRefEX

- JSTOR data
- Data for Research (DFR) API: pros and cons
- python program CRefEx (code on github)
- CRF++ (C++ library) to build the statistical model

Annotating

```
# Original line: Hom. Il. 1, 477; 24, 788;
Hom. B-CRF
Il. I-CRF
1, I-CRF
477; I-CRF
24, I-CRF
788; I-CRF

Si O
veda O
anche O
Prop. B-CRF
2, I-CRF
18a, I-CRF
7 I-CRF
ss. I-CRF
```

Feature extraction

- case
- punctuation
- number
- first and last 4 characters
- currently *no dictionaries*!

Example

```
Hom. FINAL_DOT OTHERS INIT_CAPS NO_DIGITS hom 3 H Ho Hom Hom. . m. om. Hom. B-CRF
II. FINAL_DOT OTHERS INIT_CAPS NO_DIGITS il 2 I II II. II. . I. II. . I-CRF
1, CONTINUING_PUNCTUATION OTHERS OTHERS NUMBER 1 1 1 1, 1, 1, , 1, , 1, I-CRF
477; CONTINUING_PUNCTUATION OTHERS OTHERS NUMBER 477 3 4 47 477 477; ; 7; 77; 477; I-
      CRF
24, CONTINUING_PUNCTUATION OTHERS OTHERS NUMBER 24 2 2 24 24, 24, , 4, 24, , I-CRF
788; CONTINUING_PUNCTUATION OTHERS OTHERS NUMBER 788 3 7 78 788 788; ; 8; 88; 788; I-
      CRF
```

Feature extraction

```
1  Hom. –> GT_Label "B–CRF" : Label "B–CRF" alpha: 3.944701 beta: 25.339320 p: 0.998706
2  II. –> GT_Label "I–CRF" : Label "I–CRF" alpha: 6.791799 beta: 19.480477 p: 0.992910
3  1, –> GT_Label "I–CRF" : Label "I–CRF" alpha: 12.547932 beta: 17.041941 p: 0.999661
4  477; –> GT_Label "I–CRF" : Label "I–CRF" alpha: 17.573822 beta: 11.290089 p: 0.997075
5  24, –> GT_Label "I–CRF" : Label "I–CRF" alpha: 21.537220 beta: 6.219392 p: 0.953548
6  788; –> GT_Label "I–CRF" : Label "I–CRF" alpha: 25.219915 beta: 2.182355 p: 0.886300
```

Evaluation

```
1  Average fscore: 0.929243
2  Average accuracy: 0.951830
3  Average precision: 0.940909
4  Average recall: 0.929243
```

What's next?

- share code, share training data, etc.
- train the system for specific corpora
- train the system to extract abbreviations for other materials (manuscripts, inscriptions, coins etc.)
- exploit the emerging net of citations/references between primary and secondary sources

Thanks for your attention!

matteo.romanello@kcl.ac.uk