MyMedia

Dynamic Personalization of Multimedia

Thesaurus based Keyword Extraction
TIR 2010 Bilbao 29 August 2010

Luit Gazendam (Novay)
Christian Wartena (Novay)
Rogier Brussee (Univ. of Applied Sciences Utrecht, presenter)

Problem description

- Keywords used for organising and retrieval of documents (including non textual ones)
- Problem:

> Determine keywords automatically

- Operational problem:
  - Define relevance measure of terms
  - Select collection of terms based on relevance
    - Here, just rank

Keywords, world knowledge, informativity

- Relevance of term as keyword depends on:
  - **Importance** of term for the *document*
  - **Discriminative power** of term within *document collection*
  - **A priori criteria**
    - in a thesaurus
    - right wordclass,
    - non stopword,
    - ...

Example : TF.IDF

- Consider TF.IDF
  - Input:
    - document collection $\{d_1 \ldots d_N\}$
    - Term collection $\{t_1 \ldots t_n\}$
    - Count's
      - $n(d,t)$ = # occurences of term t in document d
      - df(t) = # documents containing t (doc frequency)
      - N = # documents

$$\text{tf.idf}(d,t) = n(d,t)\ \log(N/\text{df}(t))$$

$n(d,t)$: weighs importance of term in the document

N/df(t) : weighs importance of term in the doc collection

World knowledge from thesaurus structure

- Problem: What can we do if we do not have access to large document collection ?
  - or there is no natural document collection
- Importance in the doc collection is really a proxy for importance of terms in “the world”.
  - Importance w.r.t. everything
    - ever written, English web, Information retrieval literature
- Thesauri are alternative sources of world knowledge
  - Also, required by many archives

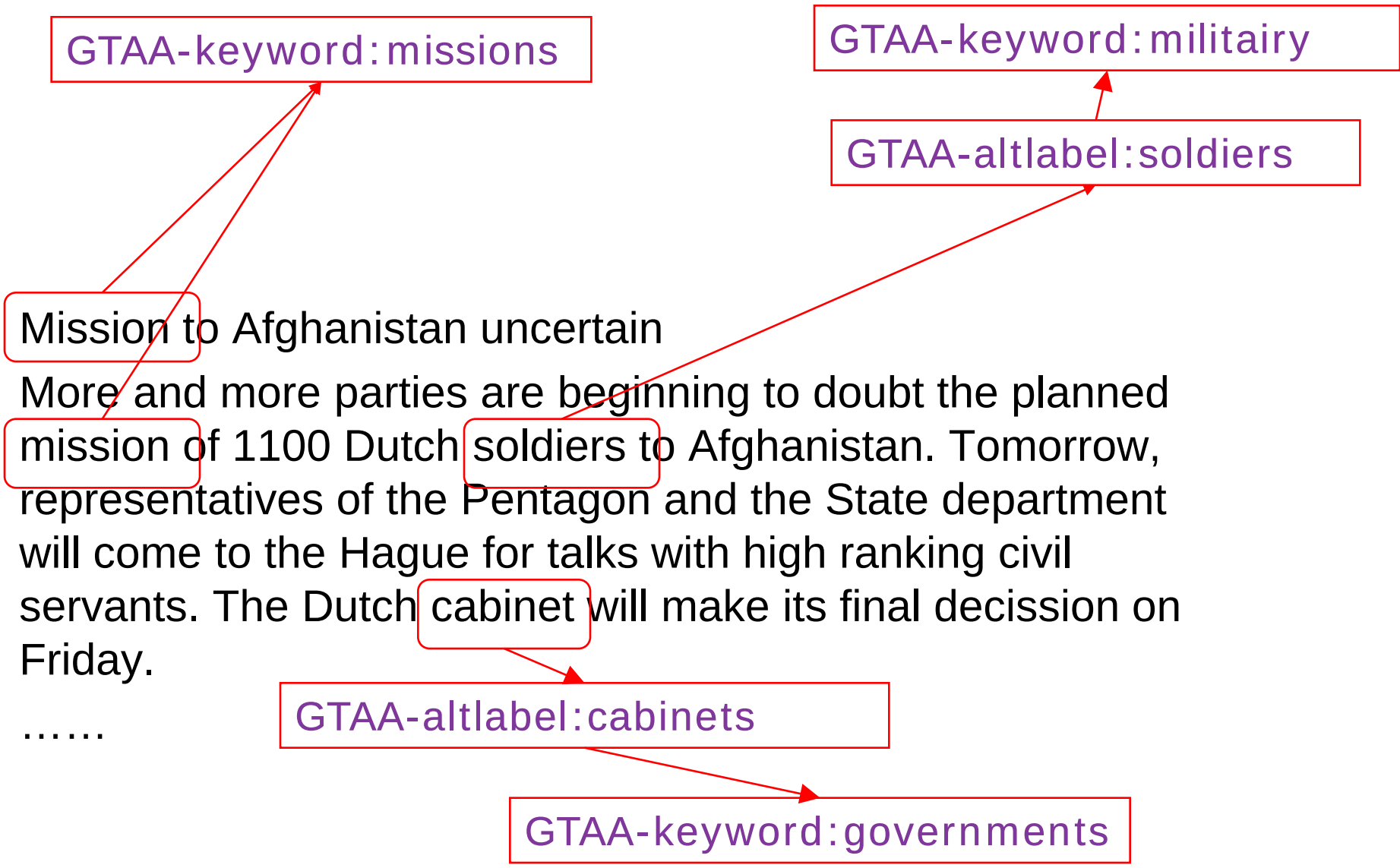
Document

Mission to Afghanistan uncertain

More and more parties are beginning to doubt the planned mission of 1100 Dutch soldiers to Afghanistan. Tomorrow, representatives of the Pentagon and the State department will come to the Hague for talks with high ranking civil servants. The Dutch cabinet will make its final decission on Friday.

Document analysis: find thesaurus terms

(Apolda semantic annotation tool)



Count lexical representations of Th-terms.

Prisons (1)

Camps (1)

Voting (1)

Missions (6)

Democratisation (1)

Prisoners of war (1)

Civil servants (1)

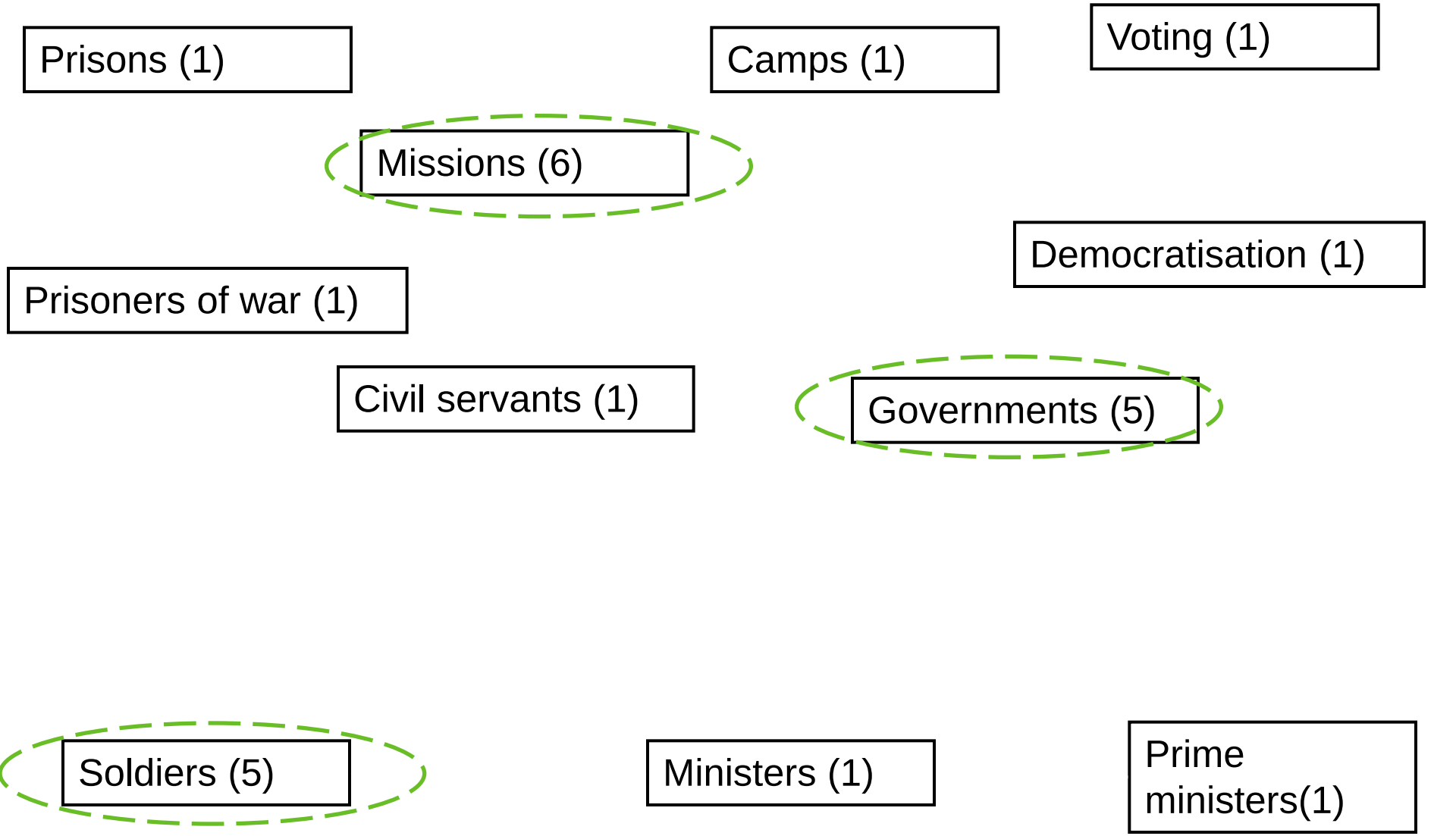
Governments (5)

Soldiers (5)

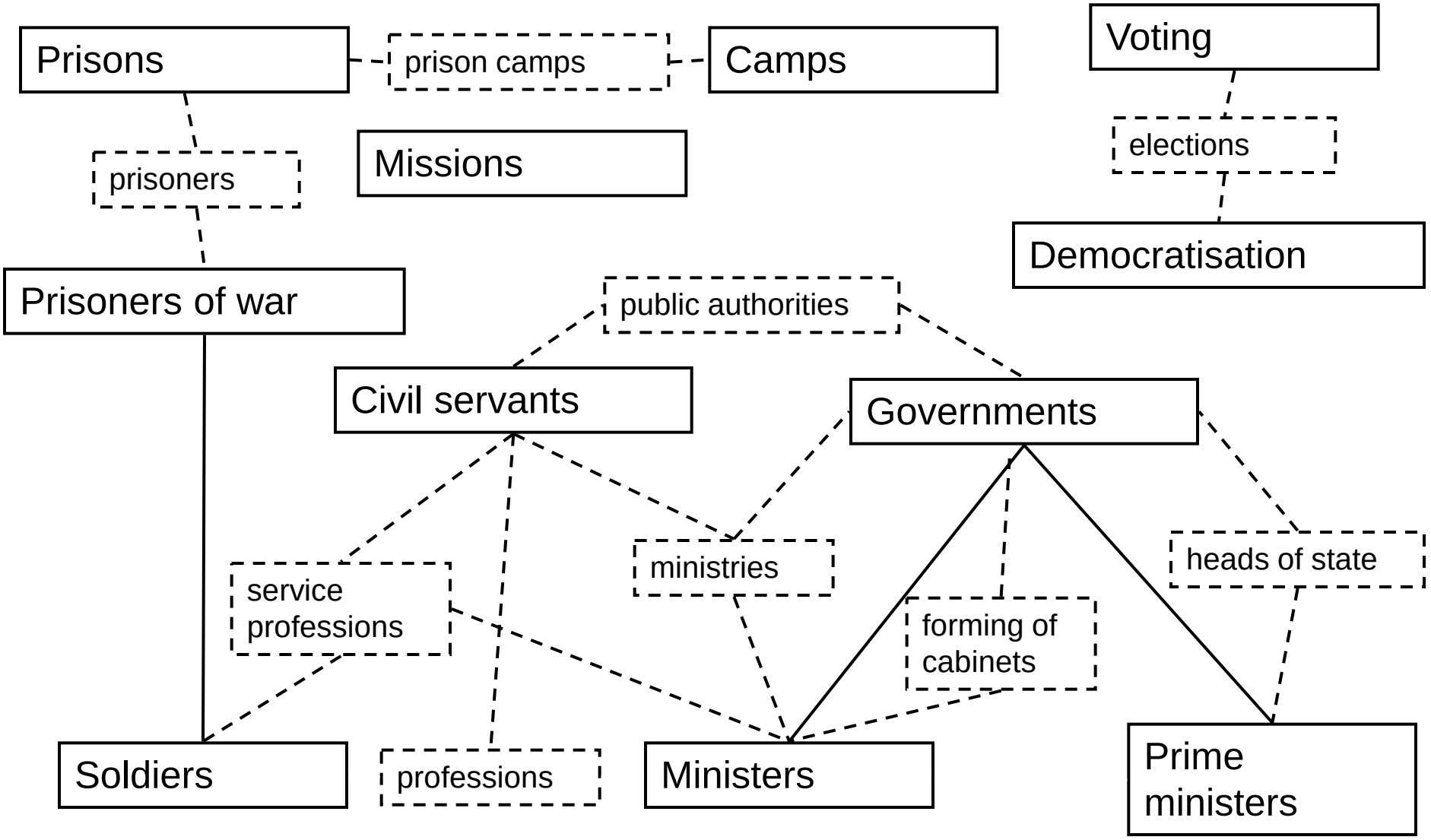
Ministers (1)

Prime ministers(1)

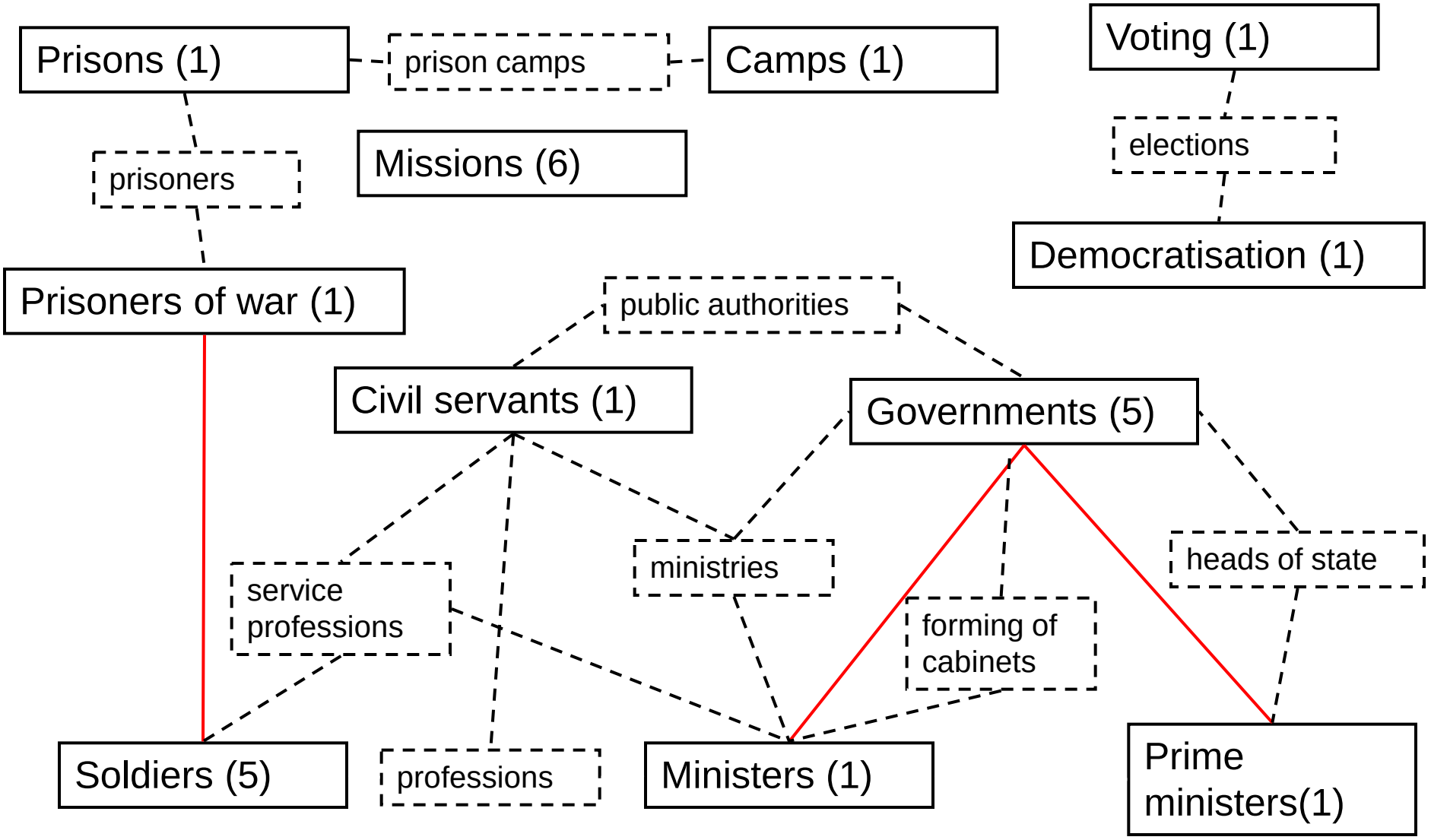
Ranking: frequency of terms (?).



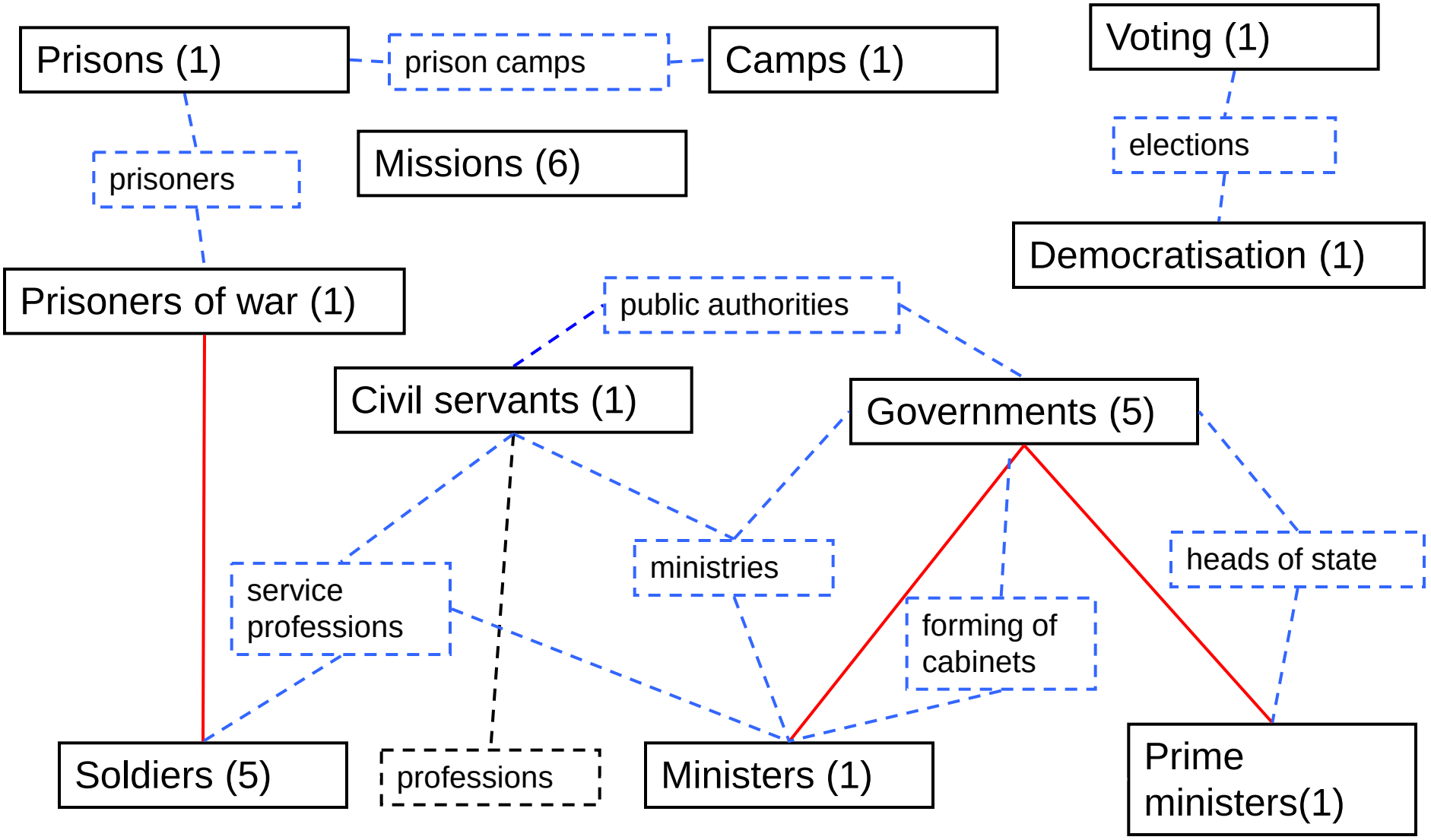
Thesaurus structure (USE, BT, NT, RT).



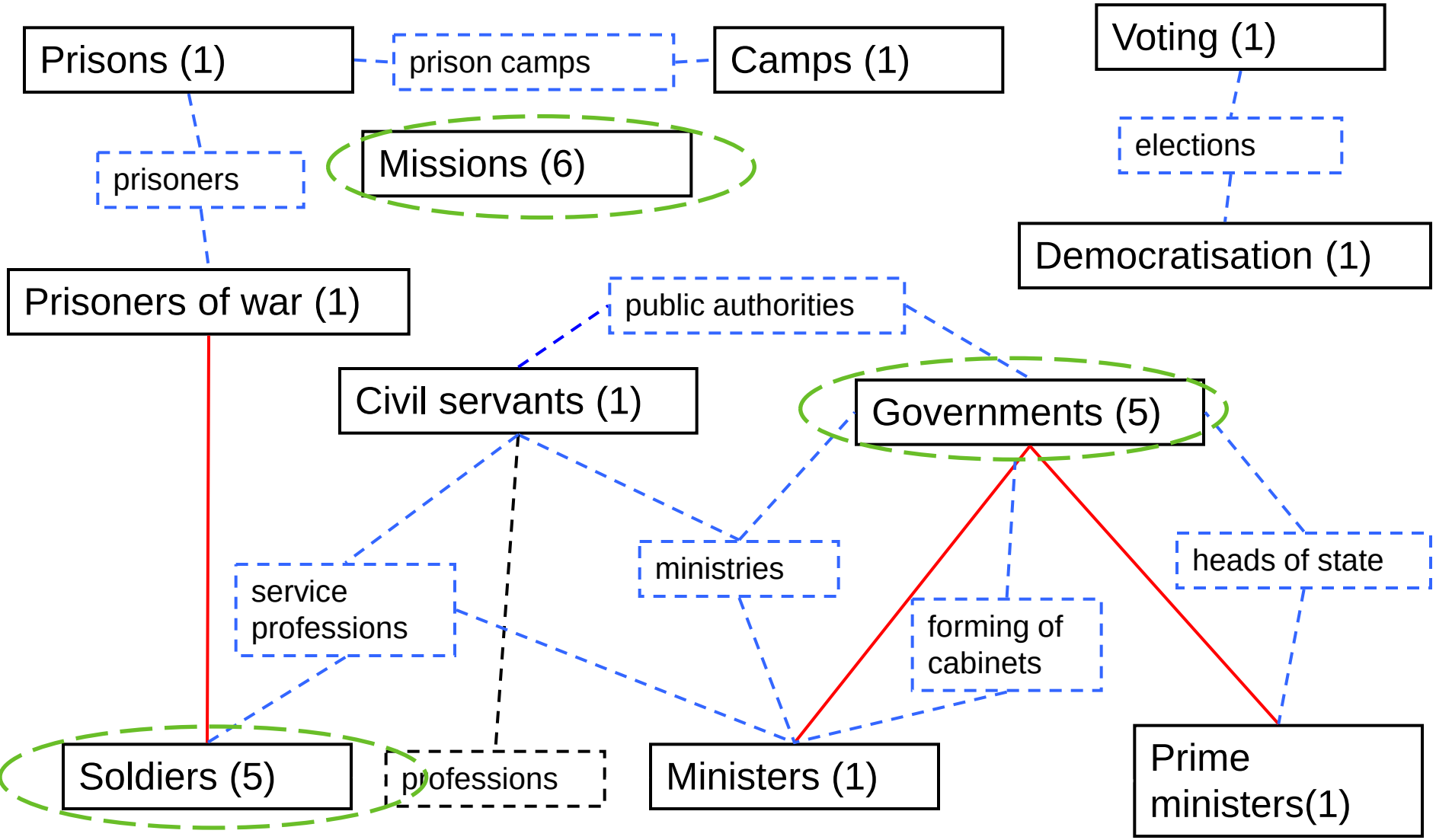
Centrality: realised relations in doc (order = 1)



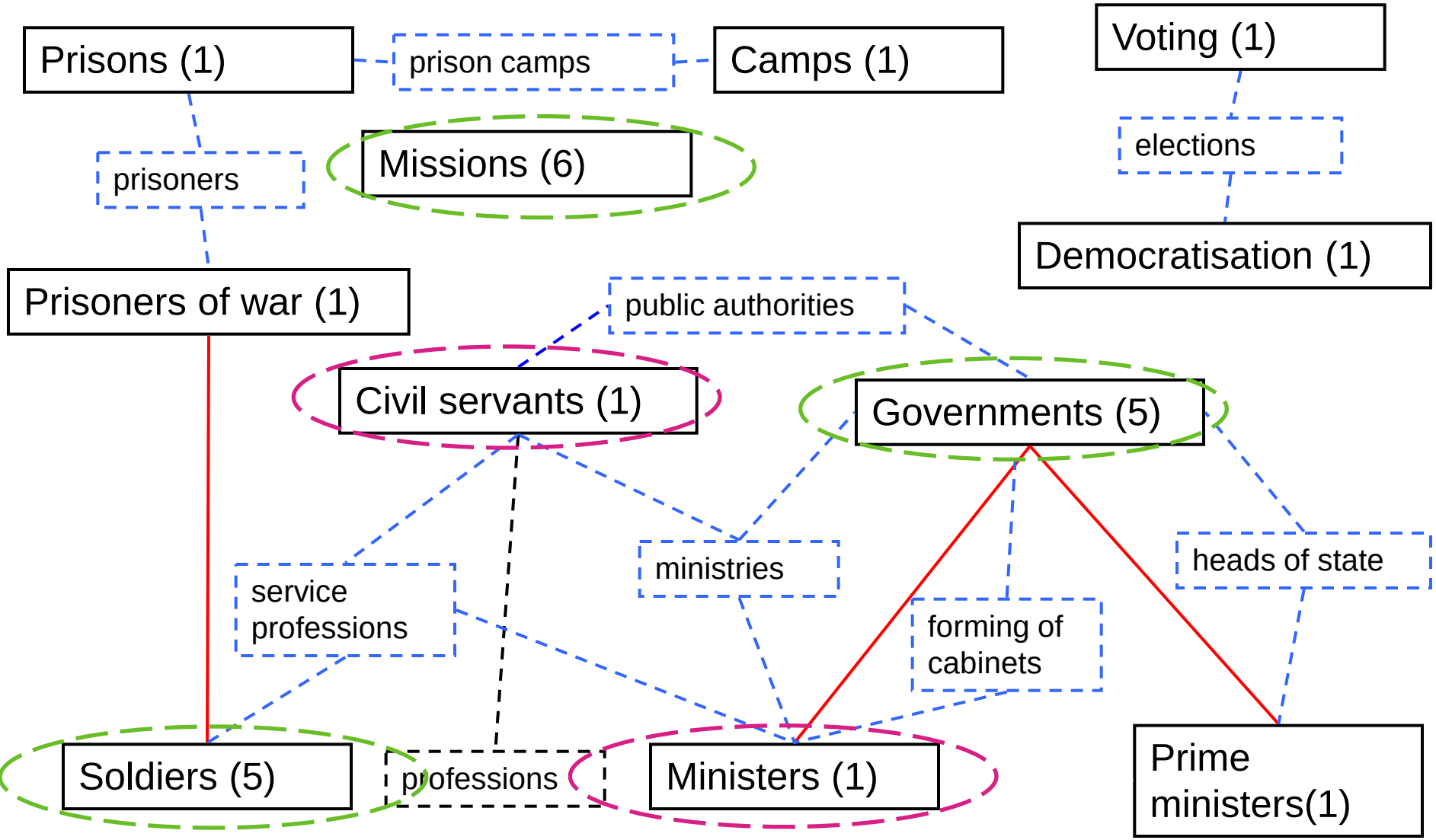
Centrality: realised relations in doc (order ≤ 2)



Ranking: frequency of terms (????)



Ranking: connectedness of terms !



TF.RR: Term frequency, Realised relations

- Select words in text that are concepts in the thesaurus
- Determine weight of (key)words by
  - Frequency
  - Number of thesaurus relations to other words in the text: central words in the text become higher weights

$$tf.rr(t,d) = tf(t,d)rr(t,d)$$

$$tf(t,d) = 1 + \log(n(t,d))$$

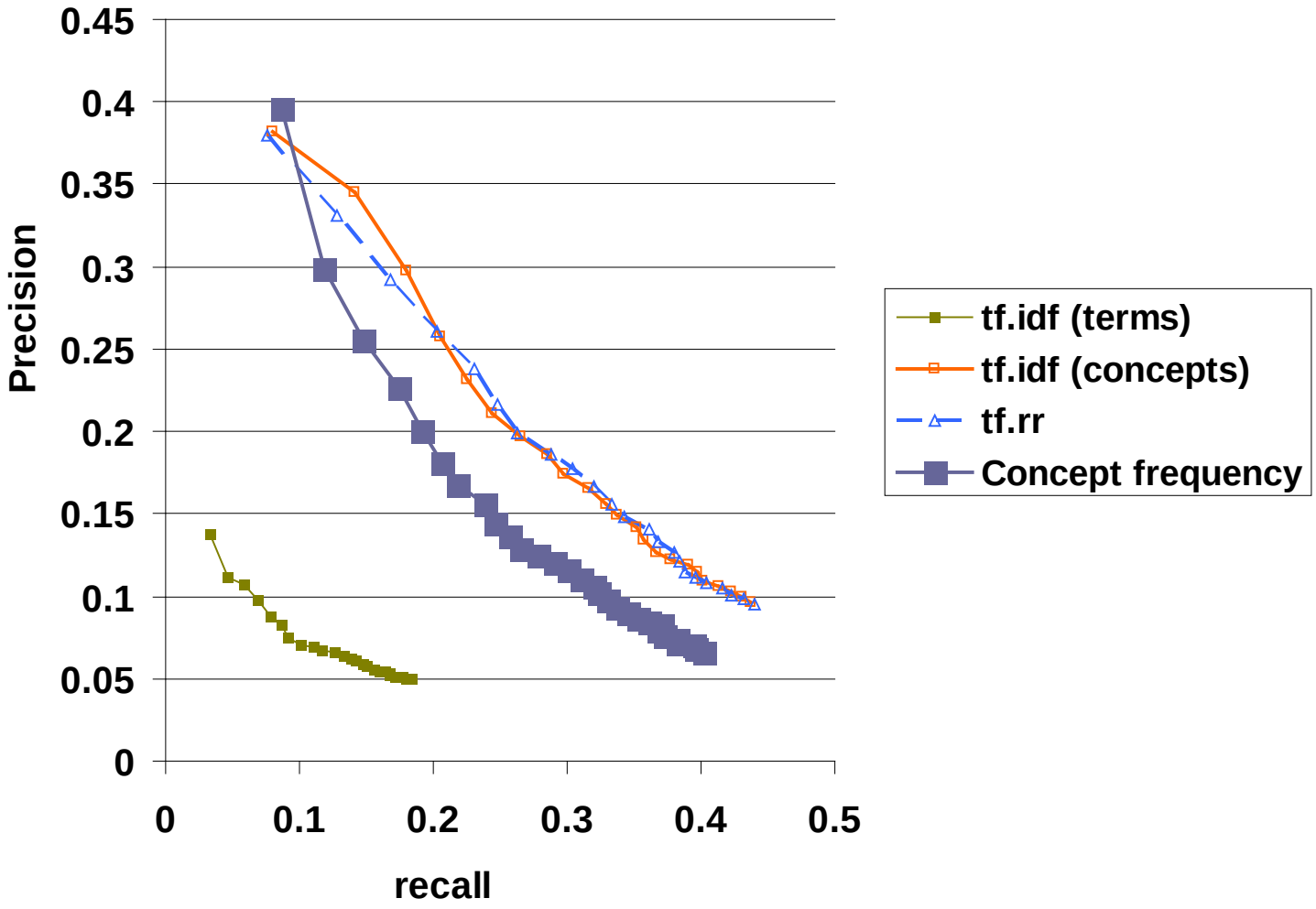
$$rr(t,d) = 1 + \mu r_1(t,d) + \mu^2 r_2(t,d)$$

  - with *n(t, d)* the number of occurrences of *t* in *d*,
  - *μ* = *α/avlinks* where *avlinks* is the average out degree of the thesaurus
    - average number of relations a term has in the thesaurus
  - We need 0 < *α* < *1 , we set α* = ½

Evaluation

- Generate and rank keyword suggestions for TV-programs from contextual resources
  - 258 TVbroadcasts
  - 362 context documents,
    - Length = 25 -- 7000 words, av = 1000
  - Thesaurus, so called GTAA (Common Thesaurus Audiovisual Archives)
    - #keywords = 3860, #relations = 20 591
  - Manual annotation by Dutch Sound and Vision Institute.
    - #keywords = 1 -- 15, av = 5.7
  - Manual keywords ground truth for evaluation
    - Inter-annotator consistency 13% --77%, av = 44%

Results


0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
Precision
0
0.1
0.2
0.3
0.4
0.5
recall
tf.idf (terms)
tf.idf (concepts)
tf.rr
Concept frequency

Conclusion

- Using thesaurus relations improves on just counting concepts/syn-sets
- Results comparable to using a corpus. i.e.

A *good* thesaurus is a reasonable alternative to having access to representative corpus