Multilingual Term Extraction

Kassius Vargas Prestes

Schedule

1 Task

2 Objective and Motivation

3 Term Extraction

- 3.1 Preprocessing
- 3.2 Methods

4 Experiments

5 Future Work

Task

- Term extraction from specific domain corpora
- Focus on multiword terms
- Specific domain
  - Software Enineering
  - Medicine, Molecular Biology - GENIA

Objective and Motivation

- Build multilingual dictionaries and ontologies to help automatic translators
- Geographically distributed software development teams
- They need to communicate with specific vocabulary
- Automatic translation can be improved with domain terminology

Preprocessing

- Tokenize the text
  - Regular Expressions
  - Can be different for each corpus
    - Genia terms contain characters like +-[ ]()
      - ca2+-modulating cyclophilin ligand
      - v-(d)-j recombinase activity

- Extract all n-grams from text
  - n from 2 to specified parameter

Preprocessing

- Filter extracted pairs by POS
  - N N
  - A N
  - N A
  - N N N
  - N N A
  - N A N
  - A N N
  - ...
- This step allow us to discard candidates like:
  - of the, in the, ...
  - basically, expressions with articles and prepositions that are very frequent and not interesting

Methods

- Actually implemented
  - Frequency
  - c/nc-value (FRANTZI; ANANIADOU; TSUJII, 1998)
  - Association Measures
    - Pointwise Mutual Information
    - Log-Likelihood Ratio
    - Poisson Stirling Measure
    - Mutual Information

- To be implemented
  - Contrastive Weight (BASILI et al., 2001)

Frequency

- Simple count of number of occurrences in the corpora

C-value

- Try to assign better scores to maximal term candidates, i. e., candidates that are not contained in another candidate
  - score of *real time clock* > *real time*

$$C - value = \log_2 |a| f_a - \frac{1}{|T_a|} \sum_{b \in T_a} f_b$$

a is a term

f_a is the frequency of a

T_a is the set of all terms containing a

C-value

- Try to assign better scores to maximal term candidates, i. e., candidates that are not contained in another candidate
  - score of *real time clock* > *real time*

$$C - value = \log_2 |a| f_a - \frac{1}{|T_a|} \sum_{b \in T_a} f_b$$

For maximal terms, this component will be 0

a is a term

f_a is the frequency of a

T_a is the set of all terms containing a

NC-value

- Used to boost c-value score
- Try to look the words that occur together with terms - "adjacent words"
- Compute a score for "adjacent words"
- Boost c-value with better scores to term candidates that occur with higher scored "adjacent words"

NC-value

- Adjacent Words score

$$weight(w) = t(w) / T$$

where:
w is an adjacent word
t(w) is the number of terms that occur with w
T is the total number of terms analysed

NC-value

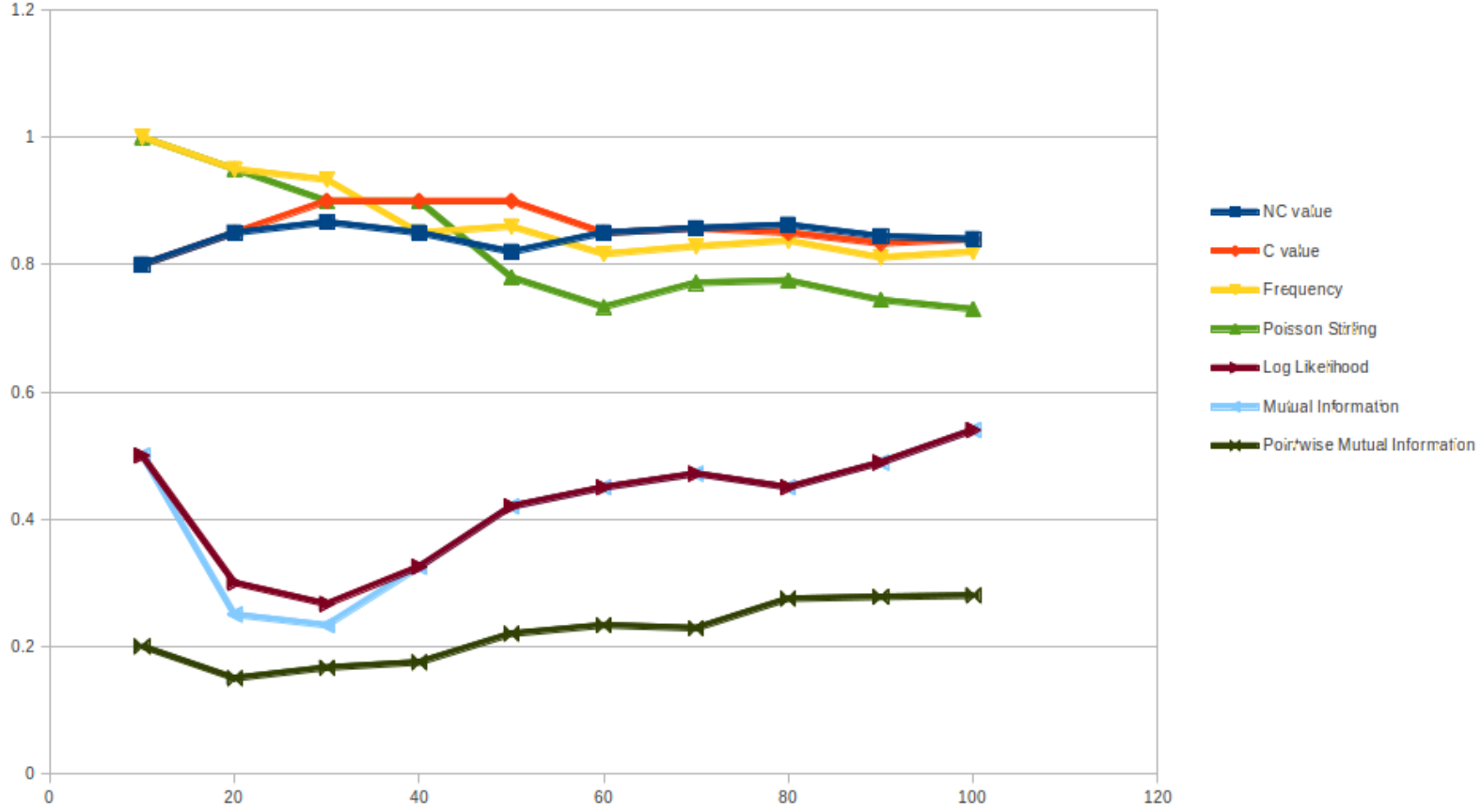
$$NC - value = 0,8 * C - value(a) + 0,2 * \sum_{b \in C_a} f_a b * weight(b)$$

$f_a b$ is the frequency that b occurs with a

C_a is the set of adjacent words of term a

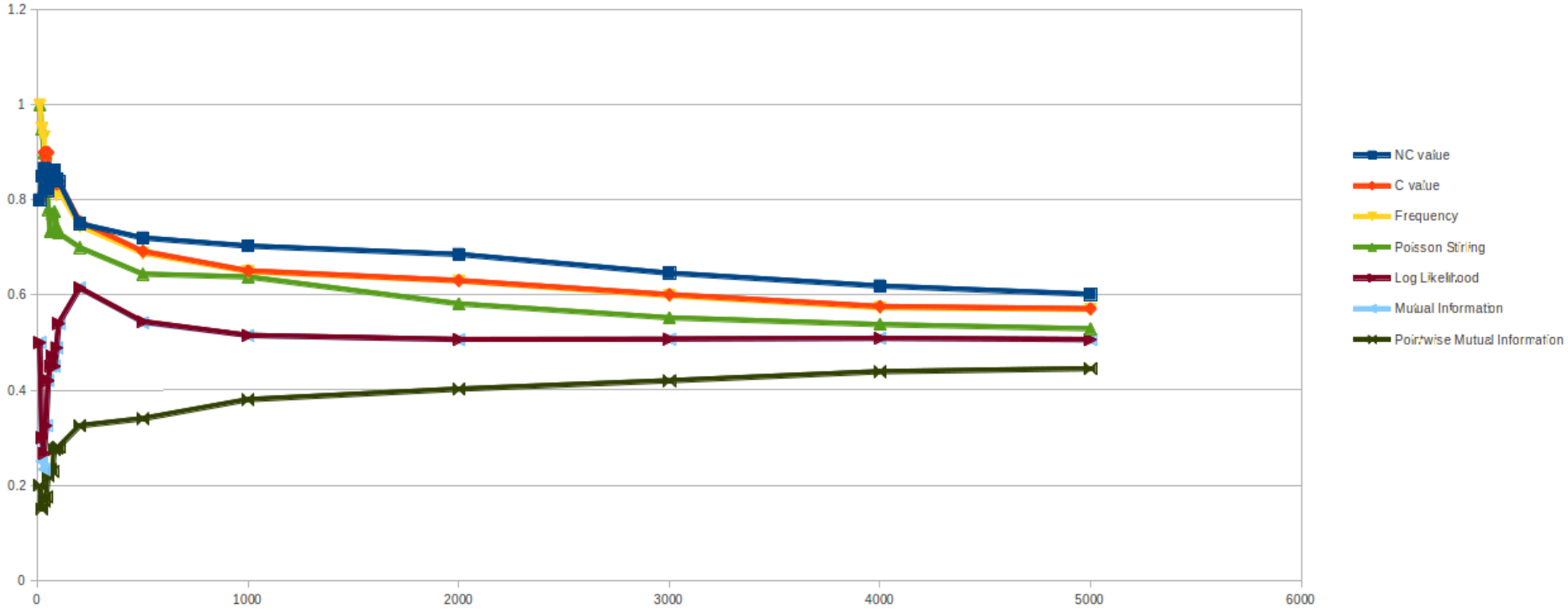
First Experiments

- GENIA corpus
  - p@10 - p@100



First Experiments

- GENIA corpus
  - p@10 - p@5000



Future Work

- Evaluate extraction of terms of software engineering manuals. (ISO, IEEE)
- Glossary as gold standard
- Algorithms of term Alignment

Multilingual Term Extraction

Kassius Vargas Prestes
kassiusvargasprestes@gmail.com