Title Bi-Gram and Title SigNoun features

Title Bi-Gram: bi-gram in noun phrases (NPs) in the Title.
Title SigNoun: significant nouns in a head NP and nouns in prepositional phrases in the Title.

Sample Parse Tree of a Title

Title Bi-Gram: ‘Named Entities’, ‘Clinical Notes’.
Title SigNoun: ‘Named’, ‘Entities’, ‘in + Notes’.

DATA

Collect technical papers in proceedings of ACL from 2000-2011.
The categories based on the category list for paper submission in the Language Resources and Evaluation Conference.
• Ex. Machine Translation, Information Extraction, Grammar and Syntax, Dialogue and so on.
Categories of each paper are manually annotated.
The total number of documents: 1972 papers.
The total number of categories: 38 categories.

FEATURES

Basic Feature
Bag-of-words in a paper (a set of lemmatized content words).

Feature Selection by Text Segmentation
Papers are segmented as follows by keyword matching:

Segments of Technical Papers
Only words in Title, Abstract, Introduction, Conclusion are used as feature.

EXPERIMENTAL RESULTS

Exact Match Ratio: the ratio of papers where predicted and gold categories are completely agreed.
Micro-F: micro average of F-measure for all categories.
Macro-F: macro average of F-measure for all categories.

REFERENCES