AUTOMATIC ACRONYM DETECTION IN HINDI

COURSE: CS365A

INSTRUCTOR INCHARGE: DR. AMITABHA MUKERJEE

BY- ANUBHAV BIMBISARIYE 11131
KANISHK VARSHNEY 11350

INTRODUCTION

PROBLEM STATEMENT

Automatic detection of acronyms occurring in daily Hindi texts like wiki articles, newspapers.

Motivation

- While reading a Hindi text, in order to fully understand it, it is necessary to realize an acronym when it is present and to know its meaning.
- Acronyms' usability depends on the reader's familiarity with them. If a user is not familiar with the acronym, they find it more difficult to read.
- Without the acronyms' understanding, it is harder to process the text and make something useful out of it. So, for an NLP agent, realizing the meaning of acronyms which might occur in a text is crucial. So the first step comes with acronym detection.
- Acronym detection might prove to be useful for various NLP problems.

RELATED/ PAST WORK

- No work done in past for Hindi language.
- Three letter acronym approach by Yeates.
- Classifier based approach for French by Ménard and Ratte.
- Various other methods by Larkey, Pustejovsky, Park and Byrd etc. Involving regular expressions, acronym lists, identification rules based on linguistic hints etc.

PROBLEMS FACED

- Source codes for past approaches not available due to commercialization and patents.
- Parallel English Hindi corpora based approach fails due to lack of a good corpus. Pure Hindi corpus existing corpus (>60000 lines) also has extremely few acronyms.
- Needed to collect new corpus from good sources with significant amount of acronyms.
- Hindi acronyms differ a lot from that of other languages in term of frequency of type of acronym, acronym governing rules, way of declaration.
- General Hindi text contains a lot of mixtures from English language, Roman numerals, English numerals, Hindi numerals etc.
- Difficulties dealing with tools used to handle Unicode.

OUR WORK

- COLLECTED CORPUS FROM A FEW WIKI ARTICLES AND A LOT OF NEWS, AND A LITTLE FROM BOOKS AND ALTERNATE SOURCES.
- DIFFERENT CODES TO RECOGNIZE DIFFERENT KINDS OF ACRONYMS STARTING FROM MOST ABUNDANT ONES.
- MERGING OF RESULTS OF ALL TECHNIQUES.
- CASES HANDLED:
  - ENGLISH LETTER ACRONYMS, USING REGULAR EXPRESSIONS. WE DEAL WITH 2-10 LETTER ACRONYMS, BUT IT CAN BE VERY EASILY EXTENDED TO MORE.
  - STOP WORD ACRONYMS.
  - NON DICTIONARY WORDS, AND ACRONYMS ACKNOWLEDGED BY WORDNET.

RESULTS

- Code recognizes 3 out of 4 types of acronyms.
- Total number of acronyms detected, including duplicate instances in the test corpus.
- Correct acronyms detected: 175.
  Total acronyms detected:195.
  Correct acronyms present: 193.
- Recall: 90% (will need another implementation to improve)
- Precision: 89.74% (can potentially be tweaked more in our existing code)
- Menars and Ratte's classifier based approach for French corpora yields similar results with precision of 89.1 % and recall of 90.9%.
- Explicit declaration out of all acronyms present in the hand tagged test corpus: 2.
- Future scope: word dictionary elimination based on word frequency of a large corpus. Will highly benefit from POS tagging. Dictionary to be trained from a very very large corpus having less number of acronyms.
- Future Scope: Implicit pattern searches to match the 4th kind of acronyms.
- Elimination from existing dictionary had a bad affect on the performance(nouns are not present, but acronyms are present in the dictionary)(Precision, recall ~ 60%).