Acronym Detection in Hindi Texts

- Anubhav Bimbisariye	11131
- Kanishk Varshney	11350

Introduction

- Acronyms are widely used today to make the reading of text easier and faster by shortening the commonly used or famous phrases, organizations, terms etc.
- 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.
- Using acronyms and abbreviations also has a benefit during chatting etc.
- Examples: WHO, radar, P2P, I.I.T., @, e.g., km etc.

Need for recognition

- The goal of text processing in NLP is to understanding text and find its meaning. This requires reading of text properly and understanding the words(tokens) before sentences can be understood.
- These words cannot be recognized from daily text without understanding the meaning of any acronym that might be present.
- Also, a good acronym detector can help users understand some acronym by fetching it's meaning when they need it, or giving alternate meanings of the same acronym when present.
- Uses include improvement in Optical Character Recognition, automated analysis of documents etc.

Related/ Past work

- A lot of work for biomedical corpora has been done.
- Methods include heuristics, scoring techniques, strict rule definitions, data mining, machine-learning schemes based on linguistic or statistical data.
- For problematic cases, Stop-Word libraries are used.
- Three-Letter Acronym system by Yeates. Recall 93%, Precision:88%.
- Various other methods by Larkey, Pustejovsky, Park and Byrd etc. Involving regular expressions, acronym lists, identification rules based on linguistic hints etc.

Motivation

- There has been work to detect acronyms in other languages, while to the best of our knowledge, an acronym detector hasn’t been made for Hindi language. With the same reasoning about need of acronym detection, we want to solve this problem for our national language.

Approach

- We first want to know how acronyms are written and used in hindi.
- Based on this study, we choose a method.

Features of acronyms(English)

Component Usage Example - Letter	
Standard (LF reference, clipping)	Radar, motel, PPP
Word substitution—phonetics-based	RnB, GnR
Syllable/Phoneme substitution	XGA, XML (X for ex)
Roman numeral	GIV
Numeric multiplier	Y2K (e.g. 2K for 2000)
Word substitution—list-based	GIK (K for potassium)
Pluralization	UAVs
Nested acronym	WSR (weather surveillance radar)

Features of acronyms(English)

-Number	
• Standard	MVV12
• Repetition (single or multiple characters can be repeated, before or after the number)	3M, P2PT, AREX2C
• Word substitution(2-to, 4-for)	B2B, P2P, L4D, 2.5G
-Symbol	
• Letter or word replacement	W@H, R&D, AT2i±, C#
• Separator that may be present in Long Form	A/V, I/O, Wi-Fi

Features of acronyms(English)

-Punctuation	
• Separator	C.I.A.
-Case	
• All uppercase characters	PGP, SSL
• Mixed-case characters	eDoc, iFrame
• All lowercase characters	radar, motel

Features of acronyms(हिंदी)

- English Letters: IIT-JEE –आईआईटी-जेईई.
- English words: internet- इंटरनेट , radar- रडार.
- Shortform words- बसपा - बहुजन समाज पार्टी, डीएमके, एआईएडीएमे, भारतीय कम्युनिस्ट पार्टी(भाकपा)
- Words in hindi ‘आप’ for आम आदमी पार्टी instead of आआपा
- Shortened words with periods कि.मी.

Types of Presentation of Acronyms

- Explicit- (..Fx{Fy}..),

  (..Fx{..P..Fy}..),

  (.. Fx .. P .. Fy ..)

- Implicit- (.. Fx .. Fy ..),

  (.. Fx { .. Fy .. }..),

  (.. Fx .. (1 to N sentences)..Fy ..)

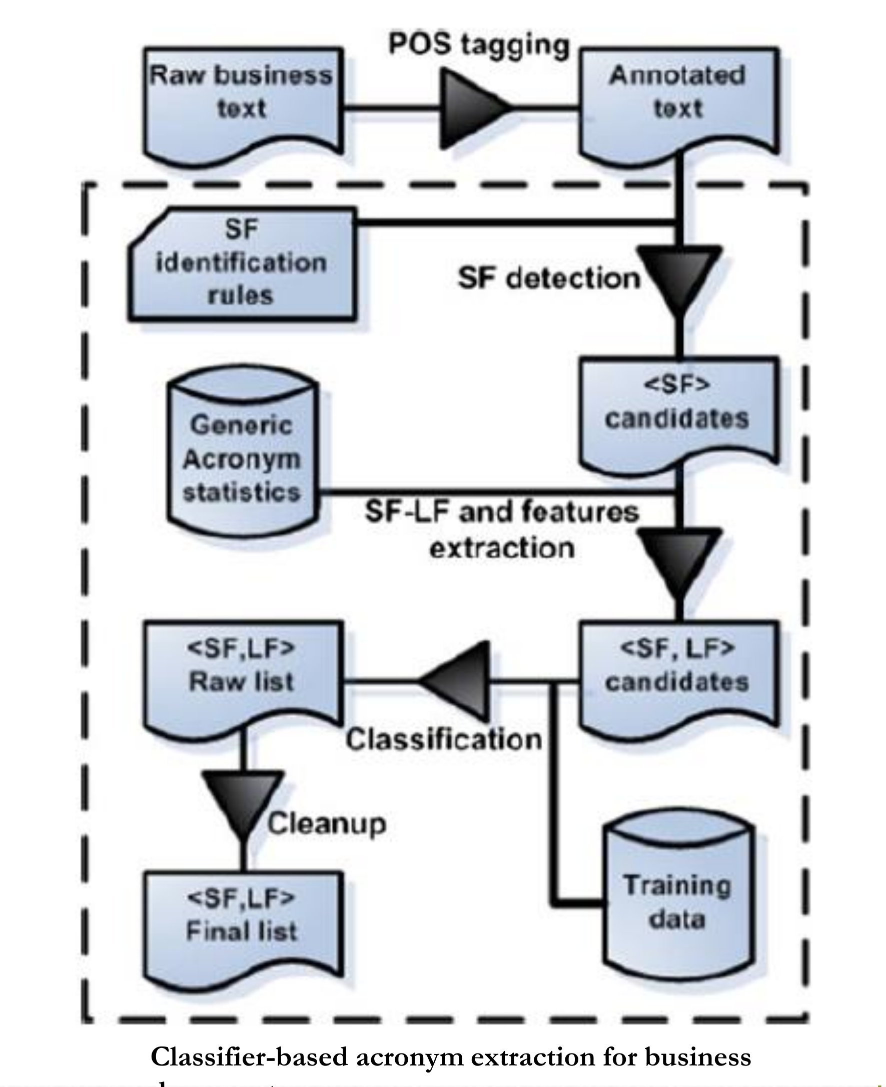
Types of separators: (), { }, [ ],< >, “ ”, << >>, - -, commas, etc.

Challenges in extraction.

- Implicit declaration is difficult to deal with. In English works, ratio of implicit declarations have been found to be 10% in the corpora.
- In a long document, a particular SF may occur a large number of times, so without an explicit presentation, we have to consider implicit presentation around each of these in the document in the search space.
- Other methods of presentations like footnotes, glossaries etc. also need to be handled.
- In Hindi, detection using conversion from English letters can gives false positives. For e.g. detection of “A” in AAP can also lead to detection of “A” in Mughal-e-Azam, because they are written using the same symbol ए.

Method

- First, identify potential acronym candidates.
- Then, potential LF is search, if not identified through an explicit instance by using a context restricted search in the search space(usually 2-3 sentences). Similarity features are used to detect if the potential match is a valid SF-LF pair or not.



Classifier-based acronym extraction for business documents
Pierre André Ménard · Sylvie Ratté

Supervised Learning Approach

- We want to extract the {A,D} pair made of an acronym A and a definition D.

From a sequence of tokens, T = <T1, T2,…, Tn>, there are *n* possible choices for acronym A = Ti and combination of one or more consecutive tokens from <T1, T2,…. Ti-1> or <Ti+1, … Tn>

The candidate pairs must be represented as feature vectors, in order to apply standard supervised learning algorithms .

Supervised Learning Approach(contd…)

- After a candidate acronym-definition pair {A1,D1} is proposed:

  a. In manual annotation of corpus, there exists a correct pair {A2,D2} s.t.

  1. A1 = A2 and D1 = D2, marked as positive.
  2. A1 = A2 but D1 ≠ D2, ignore in training and mark as negative in test.

  b. No official correct pair {A2, D2} s.t. A1 = A2, marked as negative.

Corpus

- Using Wiki pages, because they are more organized in terms of information, so we are likely to encounter more explicit presentations. Wiki has over 100000 Hindi articles.
- Hindi news papers. They include many day to day acronyms and many a times, provide an explicit presentation on first use.
- We also plan on using Hindi Style Guide for searching the stop word acronyms.

References

- Pierre André Ménard & Sylvie Ratté (2010) “Classifier-based acronym extraction for business documents” © Springer-Verlag London Limited 2010.
- David Nadeau and Peter Turney. 2005. A Supervised Learning Approach to Acronym identification. Information Technology National Research Council, Ottawa, Ontario, Canada.
- Dipak Narayan, Debasri Chakrabarty, Prabhakar Pande, and Pushpak Bhattacharyya. An experience in building the indo-wordnet - a wordnet for hindi. In GWC'02: Proceedings of the First International Conference on Global WordNet, Mysore,India, 2002.