Acronym detection (and possible full-forms) for Hindi

Kanishk Varshney (varskann@iitk.ac.in)
Anubhav Bimbisariye (anubhab@iitk.ac.in)
Advisor: Dr. Amitabha Mukherjee
Department of Computer Science and Engineering,
IIT Kanpur, India

February 28, 2014

Introduction

Acronyms are a subset of abbreviations and are generally formed with capital letters from the original word or phrase, however many acronyms are realized in different surface forms i.e. use of Arabic-numbers, mixed alpha-numeric forms, lowercase acronyms etc.[5] For instance, deoxyribonucleic acid (DNA) and ribonucleic acid (RNA), and यूरोप, मिडिल-ईस्ट एन्ड एशिया(ई.एम.ई.ए.).

Motivation

Over the past few decades there has been an explosion in the number on online technical documents, news stories and other formal documents. Along with these documents comes the creation of many author defined acronyms, as well as the use of many predefined acronyms.

The task of automatically extracting acronym-definition pairs from biomedical literature has been studied, almost exclusively for English, over the past few decades using technologies from Natural Language Processing (NLP), but no work has been done exclusively for Hindi language. Our main motivation for this topic is that often the acronyms used in Hindi do not appear in acronym-definition pairs, but as abbreviations.

Related Work

One of the earliest acronym identification systems *(Taghva and Gilbreth, 1999) is AFP (Acronym Finding Program)* [1]. The AFP system first identifies candidate acronyms, which the authors define as uppercase words of three to ten letters. It then tries to find a definition for each acronym by scanning a $2n$-word window, where n is the number of letters in the acronym.

Yeates (1999) [2] proposes the automatic extraction of acronyms-definitions pairs in a program called TLA (Three Letter Acronyms). Although the name

suggests that acronyms must have three letters, the system can find *n*-letter acronyms as well.

Larkey et al. [3] extracted Web pages to feed an online acronym list with four heuristic based methods: simple canonical, canonical, contextual, and canonical/contextual.

Pustejovsky *et al.* [4], present an approach with weak constraints, designed to capture the wide range of acronyms that are abundant in medical literature. For example, "PMA" stands for "phorbol ester 12-myristade-13-acetate" and "E2" stands for "estradiol-17 beta". Pustejovsky *et al.*'s acronym resolution technique searches for definitions of acronyms within noun phrases.

Dana Dannells [5], presented an approach to recognize and extract acronym-definition pairs in Swedish language text. A rule-based method was used to solve the acronym recognition task and compares and evaluates the results of different machine learning algorithms on the same task.

Pierre André Ménard & Sylvie Ratté (2010) [6], developed a classifier based approach for acronym extraction from Business text.

Proposed Approach

Keeping in mind the lack of reliable linguistic resources for Hindi, we propose a method which is not limited to parallel corpora only. We plan to utilize Wikipedia articles to generate comparable corpora. This can generate a large amount of data, as the number of Wikipedia articles in Hindi is over 100,000.[10]We also plan on using Hindi Style Guide[9] for searching the stop word acronyms(कि.मी., ई., etc.)

References

[1] Taghva, K. and Gilbreth, J. (1999), Recognizing acronyms and their definitions, *International journal on Document Analysis and Recognition*, pages 191-198.

[2] Yeates, S. (1999), Automatic extraction of acronyms from text. *In Third New Zealand Computer Science Research Students' Conference*, pages 117-124.

[3] Larkey LS et al (2000) Acrophile: an automated acronym extractor and server. In: ACM fifth international conference on digital libraries, DL '00. ACM Press, Dallas

[4] Pustejovsky, J., Castao, J., Cochran, B., Kotecki, M., Morrell, M. and Rumshisky, A. (2001) "Extraction and Disambiguation of Acronym-Meaning Pairs in Medline", *unpublished manuscript.*

[5] Dana Dannells "Automatic Acronym Recognition for Swedish language", Department of Swedish Language Goteborg University, Goteborg, Sweden

[6] Pierre André Ménard & Sylvie Ratté (2010) "Classifier-based acronym extraction for business documents" © Springer-Verlag London Limited 2010.

[7] David Nadeau and Peter Turney. 2005. *A Supervised Learning Approach to Acronym identification*. Information Technology National Research Council, Ottawa, Ontario, Canada.

[8] 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.

[9] www.microsoft.com/Language/StyleGuides.aspx

[10] Rahul Arora, Prabhat Pandey(2012), Cross-Lingual Word Sense Disambiguation, Indian Institute of Technology, Kanpur