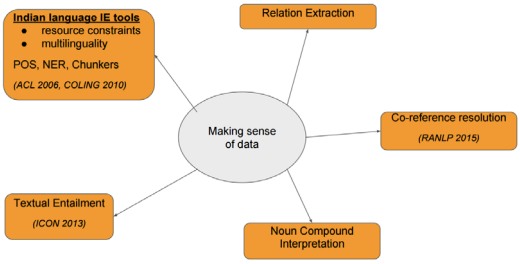
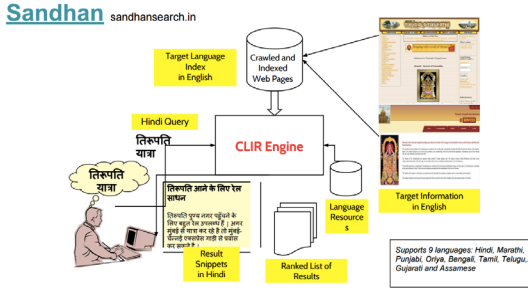
Information extraction and retrieval





Information extraction (IE) is the task of automatically extracting structured information from unstructured and/or semi-structured machine-readable documents, while information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers), using queries.

The Centre for Indian Language Technology (CFILT) at IIT Bombay has a long-standing research in both information extraction and retrieval. Under the umbrella of Information extraction, the research focuses on multiple problems:

- **Named entity recognition (NER)** is the task of identifying named entities (person names, location names, etc.) in a text. The main emphasis at CFILT is on multilingual NER. This is crucial for resource-scarce languages, as is the case with many Indian languages.
- **Relation extraction** is the problem of identifying entity mentions (noun / pronoun mentions of persons, locations, etc.) and determining meaningful relationships between these mentions. The focus is on jointly extracting entity mentions and relations, using sentence level joint inference in Markov logic networks (MLNs).
- **Noun compound interpretation** identifies semantic connection between the components of a noun compound (a sequence of two or more nouns). For example, 'student protest' means 'protest BY students', but 'student housing' means 'housing FOR students'. The work at CFILT in this area lies in two directions: (i) Data driven approach where Web is used as a corpora, and (ii) using a lexical resource such as FrameNet to identify a missing predicate.
- **Co-reference resolution** resolves co-referential mentions with referent entities. This work is towards development of knowledge base with meta-information. Out of the available paradigms for formulating this problem, the mention-pair model was followed where the mention pairs are classified as co-referent or not, followed by clustering in order to form co-referent chains.
- **Multiword expression extraction** is the task of identifying a sequence of words (such as 'zebra crossing') whose semantic, syntactic, or lexical properties cannot be fully predicted from their constituent words. The research at CFILT has explored semantic and ontological features to detect multiword expressions. The approaches have been tested on multiple Indian languages, viz., Assamese, Bengali, Hindi, Marathi, and Punjabi, but can be easily adapted to other Indian languages.

In case of IR, the focus at CFILT lab is cross-lingual information access (CLIA). This is a challenging problem because the query is in different language than the document language. The CFILT lab is a part of Ministry of Communication & Information Technology sponsored project started in 2009 which lead to the development of **Sandhan**: a CLIA search engine for Indian Languages in which query can be in one of the nine Indian languages and retrieved documents will be in English, Hindi and input language.

Prof. Pushpak Bhattacharyya, Department of Computer Science and Engineering, pb@cse.iitb.ac.in