Chapter 8

Creating Paraphrase Identification Corpus for Indian Languages: Opensource Data Set for Paraphrase Creation

Anand Kumar M.
https://orcid.org/0000-0003-0310-4510
Department of Information Technology, National Institute of Technology Karnataka, Surathkal, India

Shivkaran Singh
Arnek Solutions, Pune, India

Praveena Ramanan
Quantiphy Inc, Bangalore, India

Vaithihe Sinthiya
Karunya Institute of Technology and Sciences, Coimbatore, India

Soman K. P.
Amrita School of Engineering, Amrita Vishwa Vidyapeetham University, India

ABSTRACT

In recent times, paraphrase identification task has got the attention of the research community. The paraphrase is a phrase or sentence that conveys the same information but using different words or syntactic structure. The Microsoft Research Paraphrase Corpus (MSRP) is a well-known openly available paraphrase corpus of the English language. There is no such publicly available paraphrase corpus for any Indian language (as of now). This chapter explains the creation of paraphrase corpus for Hindi, Tamil, Malayalam, and Punjabi languages. This is the first publicly available corpus for any Indian language. It was used in the shared task on detecting paraphrases for Indian languages (DPIL) held in conjunction with Forum for Information Retrieval & Evaluation (FIRE) 2016. The annotation process was performed by a postgraduate student followed by a two-step proofreading by a linguist and a language expert.

DOI: 10.4018/978-1-5225-9643-1.ch008
INTRODUCTION

Paraphrases are pair of sentences that express the equivalent meaning using different wording. A paraphrase is a linguistic phenomenon of a language. It has numerous applications in the area of computational linguistics and teaching languages. In language teaching, paraphrasing reveals the amazing inherent power of human language. All the natural languages show their specific nature through their lexicon and syntactic constructions. The use of different forms of words, word-clusters or multi-words, phrases, and sentences reveal a different kind of dictum. If there is an ambiguity in the expression, it can be removed by paraphrasing (i.e. by using the unambiguous expression). Sometimes paraphrasing can reveal the attitude of the speaker or writer. Automatically detecting the paraphrase in Indian languages using the recent machine learning-based methods requires the paraphrase annotated corpora. The research on paraphrase in Indian Languages is less because of the unavailability of corpora.

Paraphrase detection, on the other hand, is the process of detecting a sentence or paragraph which is represented using different words while preserving the same meaning (Fernando, S., & Stevenson, M, 2008). The process of constructing the semantic representation in natural language is called computational semantics. Computational semantics is very important in many fields such as search engines, summarization, machine translation, question answering, etc. Paraphrase detection requires deep semantic understanding. The quality of paraphrases depends on one’s understanding of deep semantics. The primary importance of paraphrase is to preserve the meaning while rephrasing it (Bhagat, R., & Hovy, 2013). Paraphrase detection is much useful in text summarization, rewriting tools, evaluating machine translation systems, etc. The automatic plagiarism detection models also employ paraphrase detection techniques. To obtain high accuracy in an automatic paraphrase detection task, deep semantic and syntactic analysis of sentences is needed.

This paper aims to develop the paraphrase corpus for Indian languages as paraphrase corpus for any Indian language is not available as an open-source. This is because of the scarcity of research in paraphrase identification for Indian languages. The objective of this work is encouraging the research community working on Indian languages towards the challenging field of paraphrase identification.

TYPOLOGIES OF PARAPHRASES

A paraphrase is a distinct technique to shape different language models (Barreiro, A., 2009). Linguistically, paraphrases are described in terms of meaning or semantics. According to Meaning-Text theory (Mel’čuk, I. A., & Polguere, A, 1987), in a language, if one or more syntactic construction (sentence formation) preserves semantic equality, those are considered as paraphrases. The agreement of semantic likeness between the source and paraphrased text expresses the range of semantic similarity between them. Paraphrasing is typically associated with synonyms. Various other linguistic units such as semi-synonyms, metaphors, linguistic entailment, and figurative meaning are considered as the components for paraphrasing. It is not only seen at the lexical level. It also found in other levels such as phrasal and sentential level (Zhao, S., Liu, T., Yuan, X., Li, S., & Zhang, Y, 2007). Various levels of paraphrasing reveal the diversified classes of paraphrases and the relationship to its source document. Some of the most common paraphrase types are described below (Barrón-Cedeño, A., Vila, M., Martí, M. A., & Rosso, P, 2013).