Survey on Paraphrase Extraction Techniques for Kannada

ASHWINI GADAG[1], Dr.B.M.SAGAR[2], Mr. RAJASHEKAR MURTHY S[3]

Department of ISE, R V College of Engineering, Karnataka, India[1,2,3]

Abstract: Paraphrasing is a technique used to reframe the every word or phrase in a another way. The importance of a paraphrase is to clarify the content by re seeing and re-creating each word or phrase in every line. This paper presents paraphrasing sentences using synonym substitution, statistical model, semantic feature method. These methods are restricted to certain conditions**.** Paraphrase a word/sentence/passage when readers are not sure about its meaning. Restating a complex writing into simple terms may help better understand the content and purpose of the passage. Paraphrase recognition systems depend on lexical, syntactic and semantic features extracted from the candidate texts to identify equivalence. In the first approach, using data mining technology the target word and candidate phrases are retrieved from large corpus. It is simply extracting phrases from synonym dictionary.

Keywords: corpus, paraphrase, semantic, statistical, synonym.

I. INTRODUCTION

Kannada has been estimated to be over 2,500 years old, ranking as the 3rd oldest language after Sanskrit and Tamil. However, the Kannada alphabet evolved around 1,900 years ago. The initial development of Kannada language has followed that of other Dravidian languages, with the development of a vocal identity preceding the written system. During later centuries, Kannada has been highly influenced by Sanskrit vocabulary and literary styles. Kannada is one of the major languages of India, ranking third position in terms of number of speakers in the country, spoken mainly in the southern state of Kannada. Telugu belongs to the Dravidian family of languages, characterized by a rich system of morphology resulting in long and complex word forms.

Developing a Kannada paraphrase is useful in analysis of complex and simple Kannada sentences. Paraphrase conveys the same information but yet is written in different forms. A successful paraphrase is our own explanation or interpretation of another person's ideas. Paraphrasing effective way to restate, or clarify another author's ideas while also providing credibility to our own argument or analysis. The successful paraphrasing is necessary for strong academic writing; unsuccessful paraphrasing can result in unintentional plagiarism. [4] Explains paraphrase recognition system based on support vector machine technique.

The script itself, though derived from the Brahmi script like most other Dravidian languages, is fairly complicated owing to the occurrence of various combinations of "half-letters" or symbols that attach to various letters in a manner similar to diacritical marks. The number of written symbols, however, exceeds the 52 characters in the alphabet because of the fact that different characters can be combined to form compound characters, called "ottaksharas". Each written symbol in the Kannada script corresponds with one syllable, as opposed to one phoneme in languages like English. The script of Kannada is also used in other languages like Tulu, Kodava and Konkani.

The very first step of paraphrase exploration is construction of a paraphrase corpus with high quality and large scale. Paraphrasing is another expression that does not change the meaning of the original statement. The importance of paraphrasing lies in retrieving correct paraphrases. Word-level paraphrasing is sensitive to the context, and its critical indicator is interchangeability. Synonym substitution method explained in [5] using synsets dictionary. Paraphrasing technology has been used for many fields of natural language processing such as machine translation, automatic Question Answering (QA) system, text reuse detection, information extraction, automatic text generation, and automatic text summarization.

Paraphrase a word/sentence/passage when author want our readers to understand all of its points. Once the author’s ideas are conveyed, reader can then elaborate on them or present our opinions of the subject. The words are more worthy when they have comparatively high frequency means high word frequency are used in a lot of context.

II. SYNONYM SUBSTITUTION

The first step in synonym substitution is building a huge synonym dictionary for Kannada. The dictionary is developed such that for every word in the dictionary a set of synonym words would be given. The synonym dictionary is that the words included in it are unique (appear only once) but pay attention to usage since all words that are synonyms have the same meaning or used in the same way. The original words are replaced with words that mean the same. Develop dictionary with a word that comes as close to the meaning of the original as possible. Reread the original passage with the new word(s) in place. See if it makes sense. If it changes in meaning, come up with a new synonym.

The synonym substitution technique for building paraphrasing was used in Spanish language for steganography work. They develop the tool called JANO tool [5], it was built to help the hiding the information by means of synonym substitution.

In general, in Kannada, as in other languages, there are not absolute synonyms; this indicates that words cannot be replaced in any context by of its synonyms without altering the meaning of resulting sentence. Thus, when building the synonym dictionary entries the words that can have different meanings depending on the context can result the effect to category under the same entry that are not really synonyms among themselves, and in fact have very different meanings.

The system [5] detect (according to the rules established) the plural/singular and masculine/feminine of the names, verbs, adverbs, adjectives present in the synonym dictionary. This implies if in the text to modify a word in plural or feminine is found in the synonym dictionary this same word is in singular or masculine, the tool has to generate all its available synonyms in plural or feminine, indicating that it can choose among them according to the rules established.

The words can have more than one synonym; the decision is made in such situation where different procedures are applied to statistically decide how good the substitution of a word is by one of its synonym or by another.

E.g. The word kannu, synonyms are nayana, neetra. The candidate word is extracted depending on frequency of word.

The frequency with which word appears in that specific context should be more precise as the search is in corpus. A good paraphrase employs every single word or phrase in the original without leaving out any ideas and should not repeat any parts of the original using same words. A paraphrase helps us to understand confusing statement.

In general, in Kannada, many words are not used frequently, hence to show those words/phrases in such situation synonym substitution is useful.
E.g. ravi prakaashisuttiddane/
Bhaanu prakashisuttiddane

In the above example, the word Bhaanu may be not known to the common people. The frequently used words are substituted by non-frequent used words.

The different techniques are used for paraphrase recognition, of which machine learning technique, support vector machines. Machine learning techniques usually employ lexical, semantic and syntactic features extracted from text. The resources such as WordNet are used as basis for extracting semantic similarity features. The various techniques have been used for retrieving paraphrases from synonym dictionary. Such as probability statistical model, compares target-candidate semantic feature similarities in various sense.

When replacing the original word with a definition it is not possible to use synonym. Be aware while selecting the words that are not familiar with. It is not possible to use synonyms for the specialist terms such as microeconomics, forces, aluminum. However, in many field-specific terms – i.e., nomenclature, jargon, specific terminology, Technical terms, and such – often have no appropriate synonyms and cannot be changed.

The one of the difficulty in using synonym substitution is words can have different synonyms among which to choose.

III. STATISTICAL METHOD

In the statistical method only spellings are consider and ignore the meaning of the words. [1] Describes Statistical similarity between words. The sentences similarity is computed based on word order, word set, word vector, edit distance and word distance. Each statistical measures are calculated on set of sentences. The word set of each sentence is constructed by jaccard similarity or dice similarity. The word order vector of two sentences is constructed. The position of the word in sentence, the orders (before and after) between word pairs could be established. In Word distance the sentence between word pairs is considered.

The sentences/phrases with higher similarity have higher probability to become paraphrases. In the candidate extraction similarity between two sentences are considered as main features. Each sentence compared with other sentences to measure similarity.

Named entities has important role in paraphrase extraction. Named entities are person names, location names, organization names, numbers, times, dates and technical terms. The similarity between named entities is measured. Two stage word-level Chinese Paraphrase extraction methods described in [3]. In the first stage, look-up method was used to identify the phrases, candidate phrases were retrieved from the large corpus it also named as data mining technique. In the second stage, statistical model was established with seven similarity feature values. The candidate phrases with high similarity values are taken out. These seven similarity feature values are used to train large corpus of data. These seven features are then divided into two category, the targeting sentence is similar to the candidate sentence; targeting word is similar to the candidate word.

IV. SEMANTIC TECHNIQUE

In semantic technique semantic meaning of word is considered. Semantic technique is based on the lexical matching; each word in one sentence is compared with most similar word in other sentence. Word-to-word similarity measures are described in [1]. Semantic similarity is estimated based on large corpus of data.

Paraphrase recognition in [6] uses six patterns. All the six patterns use different strategies to identify the paraphrase. Type 1 uses sentence similarity. Type 2 and 3 deals with syntactic analysis, these three types uses set of rules that sentences should follow when changing voice-form or word-form. For other types instead of rule based BING-QUAN LIU, SHUAI XU, BAO-XUN WANG used supervised learning method. The supervised learning method described in [6], uses trained corpus to recognize the paraphrase pairs. Then use learned knowledge to decide two sentences are paraphrases. For simple sentence similarity that is type 1 most of the sentences are same as ones in the other sentences, only some words are substituted by synonyms.

The subject of the one sentence is change to complement of the other sentence during paraphrasing, mean while its object is change to subject of the other sentence.

E.g. kappu mooDagaLu suRyanannu aavarisive.
- suRyanu kappu moDagaLimdaa aavarisalpaTTiddane.

The above two sentences are paraphrases. The phrase “kappu moDagalu” is the subject in first sentence while its adverbial modifier in sentence two. This is similar to the change the voice of sentence from active to passive.

During paraphrasing the object of one sentence can change to subject of other sentence, at the same time its subject changes to the attribute of another sentence’s object and its predicate changes to another sentence’s object.

E.g. Graham Bell telephonannu avishkarisida.
Telephone omdu Graham Bellnaa aavishkaara.

In the first sentence ‘Graham Bell’ is the subject, but it is attribute of ‘invention’ of second sentence. Object in first sentence is Telephone while the subject in the second sentence.The adverbial modifier of one sentence can change to attribute of another sentence, while its predicate changes to another sentence’s object.

E.g. telephonnu Graham Bell imdaa aavishkarisalpaTTitu.
Telephonnu Graham Bell aavishkarisidaa.

V. CONCLISION

In this paper, different paraphrase recognition methods are discussed. Supervised and rule method is not dependent on single corpus, so it can be used in many areas. Each paraphrase technique has merits so better to use in combination instead of single technique. This is just the beginning of research on paraphrases, we pay our attention mainly on sentences with simple structure. The sentence similarity is calculated by dice similarity computation.

REFERENCES

[1] Savitha Sam Abraham, Sumam Mary Idicula. Comparison of Statistical and Semantic Similarity Techniques for Paraphrase Identification. International Conference on Data Science & Engineering 2012

[2] Hu Hongsi, Zhang Wenbo, Yao Tianfang. Paraphrase Extraction from Interactive Q&A Communities,

[3] HE Xian-jiang, YU Zhong-hua. A Research on Multi-feature Word-level Paraphrase Extracting System Based on Context. Fourth International Conference on Multimedia Information Networking and Security 2012.

[4] A.Chitra, Anupriya Rajkumar. Evolutionary Approach for building Efficient Paraphrase Recognizers. World Congress on Information and Communication Technologies 2011.

[5] Alfonso Muñoz, Justo Carracedo, Irina Argüelles Álvarez. Measuring the security of linguistic steganography in Spanish based on synonymous paraphrasing with WSD. 10th IEEE International Conference on Computer and Information Technology 2010.

[6] BING-QUAN LIU, SHUAI XU, BAO-XUN WANG. A Combination of rule and supervised learning approach to recognize paraphrases. Proceedings of the Eighth International Conference on Machine Learning and Cybernetics, Baoding, 12-15 July 2009