MULTILINGUAL TEXT SUMMARIZATION TECHNIQUES

Sherry[1], Anjali Saini[2]

[1]*Assistant Professor, CSE, CGC, Landran*
[2]*Assistant Professor, CSE, CGC, Landran*

Abstract

A Summary is a short document that represents the essential information from the given document. Text Summarization represents essential text information by leaving the irrelevant detail, reduces the details and contents them in short way that meet with requirements of user. Text Summarization may define on the basis of some important features like summaries may be short. They should conserve the required information. Text Summaries may generate from one or more documents. So with the use of multilingual text summaries we can provide information to people which are not able read or understand particular information.

Keywords: - *UNL, Summary, Multilingual.*

----------***----------

1. INTRODUCTION

In today's era electronic documents are becoming a postulate of business and academic knowledge. Many electronic documents are propagated and made accessible from the internet. In order to utilize these on-line documents efficiently, it is difficult to extract the basic idea behind these documents. Having a text summarization system would thus be extremely useful. For example, news information, specific field related information, biographical information, stock market information which is required usually. Now a day's people go to see movies and all on the basis of reviews they had seen. But it is not possible every time to read the whole content. Information should be in reduced form so the user can save the time [1]. Secondly sometimes information is impenetrable to some users. For example, children or people who are learning English as a foreign language. So Multilingual Text Summarization provides summary to user in a language user want and in simplified manner.

2. BUTTONHOLES OF TEXT SUMMARIZATION

Although many summarizing algorithms are available for simple text summarization and few are available for multilingual text summarization. Hence it becomes a challenge to generate algorithm which will produce meaningful and timely summaries.

Another big challenge is how to evaluate the summaries. After applying algorithm one must be confident that summaries should be relevant and considering all the important factors of the main document,*i.e.*, if the resultant summary is textual it should preserve the main foci. Irrelevant information should not be the part of document after filtering. The union of conciseness, readability and completeness would always give good summary [2].

3. PURPOSE OF TEXT SUMMARIZATION

Text Summarization used is in medical field, in multimedia news summarization, in producing intelligent reports, in text for hand held devices, in text-to-Speech for blind people, in education and in summarizing meetings [3]. Many other scenarios use text summarization. For example, an information retrieval system uses automatic summarization to produce the list of retrievals. Now a day's summary of the email messages and news articles is sent to mobile devices as Short Message Service (SMS). Search engines also use summary mechanisms. The summary of the web pages is shown on the screen as a result of particular search.

4. APPROACHES OF TEXT SUMMARIZATION

There are different approaches used for text summarization. They are machine learning approach, deep natural language analysis approach, multi document summary approach and multilingual approach. The methods like naïve Bayes, decision trees, Markov models, Log linear models and neural network are under the machine learning approach. Multi document summary includes graph based method, centroid based method and topic driven methods. Multilingual summarization approach uses mostly Universal Networking Language (UNL) for summary generation and language translation [4].

5. UNL AS AN IMPORTANT TOOL FOR TEXT SUMMARIZATION

Universal Networking language (UNL) is used in multilingual text summarization. It is intermediate language which is used for language translation, *i.e.*, with the help of UNL it is possible to produce a summary of a source natural language document to condensed destination natural language document. The process is first of all source document is converted into UNL document by UNLization process. Summary of this UNL document is produced by using algorithmic approach. Then resultant summary is converted into destination natural language using NLization

process.UNL-based summary provides the following important features which are as follows:

- Text which has to be summarized written in any natural language.
- UNL based approach generates both plain text summary and corresponding summary in any natural language, provided that UNL encoading has been given and all the decoding related to destination natural language have been provided.
- UNL approach make system more independent from any natural language and less complex due to the limited semantics given by UNL.
- UNL-based summarization algorithm can be applied to any web demands, provided text should already encoded into UNL.
- Using UNL in 2n translation n languages can be converted to n-1 languages which describe the efficiency of UNL.

So it provides the features like unambiguity, multilinguality which are beneficial for multilingual summary point of view.

5.1 UNL Representation

The information provided by natural languages is expressed in UNL in the form of semantic network. Semantic network comprises various discrete semantic units like Universal Words (UW), Universal Relations and Universal Attributes .Universal Words form a nodes in a graph. Universal Words are the simple English word represents nouns, pronouns, and adjectives *etc*. Universal Attributes are annotations made to nodes in a semantic network. Links in a semantic network is represented by Universal Relations.

5.2 Universal Words

Universal Words represents the content conveyed by natural languages in the form of nouns, verbs, adjectives and adverb *etc*. They are "Universal "in the sense that they comprise the lexicon of the "Universal Language", *i.e*., they convey the ideas that can be expressed in each and every language [5]. Universal Words are expressed by nodes in a semantic network. Universal Words are divided into two categories

5.3 Universal Attributes

Universal Attributes are annotations made to nodes in a semantic network. They actually denote the circumstances under which these nodes are used like tense, mode, aspect *etc* [6]. Attributes may convey three different types of information which is as follows:

1. Information about the role of the node
2. Information conveyed by morphemes and classes such as affixes, determiners, conjunctions, degree adverbs *etc*.
3. Information on the external context i.e. non verbal elements of classification.

"@past", "@present"," @def", "@indef "are some examples of attributes.

Table 1.1: Some UNL Attributes

UNL Attributes
@past,@present,@future
@generic, @def, @indef
@multal, @extra
@pl
@male, @female
@transitive,@symmetric,@identifiable,@disjointed
@habitual, @perfective and @progressive
@ability,@grant,@wish,@will,@obligation,@possible ,@regret,…
@affirmative, @imperative, @interrogative, @request,

5.4 Universal Relations

Universal Relations are known as "links", are labeled arcs which are used to connect one node to another in a semantic network. UNL relations are used to represent the semantic cases or roles such as agent, object and instrument *etc* between Universal Words [7]. Relation is represented by three letter words that specify the kind of semantic relationship between the two Universal Words such as <nameof relation>(<source>;<target>)

5.5 UNL Sentence

UNL sentence is a basic block for expressing the UNL framework. It consists of Universal Words related by Universal Relations and modified by Universal Attributes. When UNL sentence is represented in the form of semantic network consisting of nodes and arcs it is called UNL graph.

5.6 UNL Architecture

Figure 1.3 shows the architecture of UNL. UNLization and NLization processes are carried out with encoders and decoders. Natural language sentences are converted by IAN to UNL during UNLization process. For this process language grammar provided by Transformation Rules (T-Rules), dictionary entries, Disambiguation Rules (D-Rules) are required. Once UNL is produced after that with NLization process of decoder UNL is converted into the other natural language. Decoder uses also language grammar and dictionary rules and Disambiguation Rules.

5.7 UNL Tools

UNL tools are the software programs which are available for transformation,*i.e*., tools which convert the natural language sentence to UNL and then again UNL to target language. Tool which is available for conversion of a natural language to UNL is called EnConverter and tool at other end for conversion of UNL to target language is called DeConverter.UNL contains following two tools :-

5.8 IAN (Interactive Analyzer)

IAN is a tool used to convert natural language to UNL.It contains the grammar for natural language analysis .Word Sense Disambiguation (WSD) is carried out by language expert, but the system can also have a set of Disambiguation rules (D-rules) [8]. IAN deals with natural language rules

(N-rules), transformation rules (T-rules) and disambiguation rules (D-rules).

5.9 EUGENE

It is known as deep to surface natural language generator because it converts UNL language to a desired target natural language. It is language independent and it uses dictionary and grammar rules for its conversions [9].

6. SUMMARY TECHNIQUES

6.1 Semantic and Syntactic (Rule-Based)

There are many semantic analysis techniques which are applied on text summarization to find the relation between different sentences. Following are the three semantic and syntactic summary techniques.

- Graph Representation
- Lexical Chains
- Natural language processing

In the graph representation lexical graphs, Graph matching, weighted graphs and unweighted graphs are used for summarization. In lexical chains word net, co-reference chains and lexical semantics *etc* are used for summarization. Natural language processing used information extraction, part of speech tagger for the summarization. Summarization techniques under NLP are divided into two categories as follows

- Plain Text Summarization
- Multilingual Summarization

In Plain text summaries resultant summary is in the same natural language but in multilingual text summarization resultant summary is in different natural language. Initial work in plain text summarization was started in 1950's. Most initial work on the text summarization was targeted on technical documents. information, biographical information. According to the Luhn summary has different categories, some of the summaries are difficult to generate than other. Different categories are Extractive, Abstractive, Indicative Multilingual text summarization is come into existence in 2005. This technique is still in early stage but this different framework has many advantages in the newswire field in which information is combined from different foreign news agencies. Evans (2005) described the scenario in which there is always a preferred language in which summary is required, different multiple source documents are in demand and in different languages are available. They preferred English as a source language and documents are from the news articles in English language and Arabic. The logic was to generate the summary of English articles without discarding the details contains in Arabic. IBM's machine is used to do a transformation of Arabic language to English. The system checks the transformed document in Arabic corresponding to a document of English for each sentence. If match is found then sentence is found relevant for summary. Hence more grammatical summary is found this way, since machine translation is still not perfect of that. To find out the similarities between sentences Simfinder tool was used. This is a clustering text based tool based upon similarity over different semantic and lexical features which is using long linear regression model. Universal Networking Language is mostly used in multilingual summarization.

sMartins and Rino proposed algorithm for the text summarization using UNL. They presented UNLSumm model to prune the UNL text by means of heuristics that totally focus upon unnecessary binary relations. The system used decoder to produce corresponding summary in Brazilian Portuguese. Their pruning heuristics are based upon the relations of UNL. Although each relation is not candidate for pruning because some relations like "agt" or "obj" convey important information [10]. Only some of the relations are candidates for pruning. According to this algorithm initially there was 84 heuristics were divided into two groups A and B shown Group A considers 39 heuristics. It also called as single pruning and removes the independent binary relations one by one. Group B heuristics are complex than the Group A heuristics. Group B heuristics are called chained pruning *i.e.* once the binary relation is excluded the interconnected binary relation is also excluded.

There are some more limitations of this approach.

- Sometimes it covers non-relevant information.
- There is an upper bound to the number of heuristics applied for each entry.
- Application order is relevant and providing satisfactory results or not.

Managaikarasi, Gunasundari (2012) proposed an idea of text Summarization. The most important work they have done is improved methodology, which scans the document and transform into UNL graph. The system introduced UNL as a language for knowledge representation and information representation that can be describe in natural language conversation. They proposed method to find the summary of UNL document. The documents are collected from websites based upon the education domain. These documents contain images and unwanted information also. In the First Step stop words are removed. The sentence splitter is used to split document into sentences. The delimiter used is blank space here. In the next step sentences are again splitted into words. Then Morphological analyzer analyzes these words to find out the root word. These root words provide to UNL dictionary. Tenses and heuristic relations of the root words are indentified. The graph is constructed from given information. During graph construction counter field is also updated. Counter field is provided to find the important concepts, based upon threshold. Highest concepts sentences are finally picked for the resultant summary of the document. The system is tested on education domain document for summary. There was manually test on the summary with senior people. During summary preparation the data is collected from the news service providers. Each document includes the irrelevant information like images, tables *etc*. So there is a need of creation of ideal summary for evaluation of results. For the ideal results the documents are distributed to three judges and rank is given to the sentences according to their importance in text document.

The future work for project is develop a well managed tool for evaluations and updating of UNL dictionaries with the help of root words provided by Morphological analyzer. It also identifies more and more UNL relations with the help of heuristic rules [11].

Pandian and Kalpana (2013) also proposed text Summarization mechanism using UNL. They focused on the tourism domain document which is UNL based. The Bengal UNL system is developed by them. UNL representation is for simple sentences not for complicated sentences. The main focus was mostly on deconversion part which converts the universal networking language to Tamil language. The source document is scanned and it is converted into intermediate language. It further undergoes generation process for final output. For the summary process the source document undergoes a process of enconversion which includes the steps like Parts of speech tag, Parts of speech parsing, Identification of entity, Identification of relation, Creation of dictionary and Generation rules. Source document is converted into UNL document of UNL expressions. In the first phase parts of speech tagging and parts of speech parsing is carried out with the help of Stanford Parser. The outcome of the parser is used to find the entities and relationship between entities. Further rules are constructed and knowledge base is obtained for the generation of UNL expressions. UNL document containing UNL expressions are passed to the deconvertor for the generation of the final summary and final output for three levels of users (level1 user, level2 user and level 3 users). The Deconversion module is constructed in such a way that it will perform the function of both summarizer and deconvertor. To obtained the summary deconversion module scan the word dictionary and find the relation between the different universal words, attributes of the universal words are collected and relation between universal words are taken. Further the unnecessary information like determiners, prepositions are reduced to obtained the final summary document. Final summary document is produced for the different levels of user's base upon the classification of ages. The distribution level of the summary document based upon the IQ level. Deconversion module produce summary in three steps produce summary in three steps[12].

- Analysis and preparations of the dictionary information.
- Preparing Deconversion rules
- Deconversion to produce the output summary document

Sornlertlamvanich *et al.* proposed an approach for Summarization using Universal Networking Language. While producing summary this approach considers surface and semantic information of the UNL. The multilinguality can also be realized using deconvertors from the summarized UNL document to the resultant target natural language document under the framework of UNL. Algorithm consists of four steps. In the first step the score of each UNL sentence is calculated. Score of the sentence is calculated by using weight of each universal word. Weight of each universal word is calculated by using the factor of frequency and inverse document frequency. After the score calculation some top most sentences based upon score are chosen for the future summary. By using the semantic information of the UNL the redundant words are removed from the summary in third step. This is mathematically calculated by using contribution function. The values obtained through contribution function are compared with the threshold value 1.5. To make summary more natural and real different sentences are merged based upon the head of the sentence and no of words in the sentence in fourth step. This algorithm is applicable for multiple document summarizations. Their experiment proved that use of the UNL improves the summary quality as compare to the plain text summarization. The semantic information of the UNL can also be applied to improve the naturalness in sentence level of summary [13].

7. CONCLUSION

The multilingual summaries are very important. They are used for language translation as well for summary calculation. There are still areas in world where language translation is required. All the above techniques have some pros and cons. But in today's era this field has remarkable demand which is growing day by day.

REFERENCES

[1] Lal, Partha. "Text Summarization." (2002).
[2] Mangairkarasi, S., and S. Gunasundari. "Semantic based text summarization using universal networking language." *Int. J. Appl. Inf. Syst* 3.8 (2012): 18-23.
[3] Lal, Partha. "Text Summarization." (2002).
[4] Das, Dipanjan, and André FT Martins. "A survey on automatic text summarization." *Literature Survey for the Language and Statistics II course at CMU* 4 (2007): 192-195
[5] Unlweb.net,. 'Universal Words - UNL Wiki'. N.p., 2015. Web. 20 May 2015.
[6] Unlweb.net,. 'Universal Attributes - UNL Wiki'. N.p., 2015. Web. 20 May 2015.
[7] Unlweb.net,. 'Universal Relations - UNL Wiki'. N.p., 2015. Web. 20 May 2015.
[8] Unlweb.net,. 'Introduction to UNL - UNL Wiki'. N.p., 2015. Web. 21 May 2015.
[9] Martins, Camilla Brandel, and Lucia Helena Machado Rino. "Revisiting UNLSumm: Improvement through a case study." *the Proceedings of the Workshop on Multilingual Information Access and Natural Language Processing*. Vol. 1. 2002
[10] Mangairkarasi, S., and S. Gunasundari. "Semantic based text summarization using universal networking language." *Int. J. Appl. Inf. Syst* 3.8 (2012): 18-23.
[11] Kalpana, S. "UNL based Document Summarization based on Level of Users."*International Journal of Computer Applications* 66.24 (2013).
[12] Sornlertlamvanich, Virach, Tanapong Potipiti, and Thatsanee Charoenporn. "UNL Document Summarization." *Proceedings of the First International Workshop on Multimedia Annotation*. 2001.