KABARAK UNIVERSITY
6TH ANNUAL INTERNATIONAL RESEARCH CONFERENCE

A Parallel Corpus Based Translation Using Sentence Similarity

NAME OF PRESENTER : *RUORO SIMON, LAWRENCE SIELE,*

20/07/2016

Introduction / Background

- Text translation is critical for the acquisition, dissemination, exchange and understanding of knowledge in the global information society
- this form the basis of much multilingual research in natural language processing, ranging from developing multilingual lexicons
- Translating large quantities of parallel corpora texts manually, make it difficult to produce consistent translations of text, such as paragraphs, sentences and phrases.
- The parallel Corpus-based translation systems make use of existing parallel texts to guide the translation process

Statement of the problem

- Translating large quantities of parallel corpora texts manually, make it difficult to produce consistent translations of text, such as paragraphs, sentences and phrases, making it impossible to reuse previous translations stored as translation memories and thereby minimizing the chances of producing alternative translations of the same source sentence that provide users with better understanding on word usage in sentences.
- Unlike this approach, traditional translation and dictionaries are limited and users often cannot find explanations concerning words usages

Study objectives

1. To investigate, to what extent sentences can be extracted from parallel corpus on multiple languages.
2. To developed an experimental English-Swahili example based machine translation (EBMT) system, which exploits a bilingual corpus to find examples sentences that match fragments of the input source language
3. To provide an array of sentences, and allow the user to select the best equivalent sentence for the source sentence, and see in what circumstances a word would typically be used in practice.
4. To create a library of multilingual sentences to facilitate translation for English-Swahili languages.

Brief literature review

- According to research there are a lot algorithms available for text similarity after from our analysis we chose to use the edit distance in order to compare the input sentence with different examples in the translation memory for EBMT

Problems with this method.

- It measures differences between strings and not words,
- when the translation memory is built from a parallel corpus, the constituents are quite big sentences

he study provided insight into areas where the recall of translation memory systems can be improved and edit distance

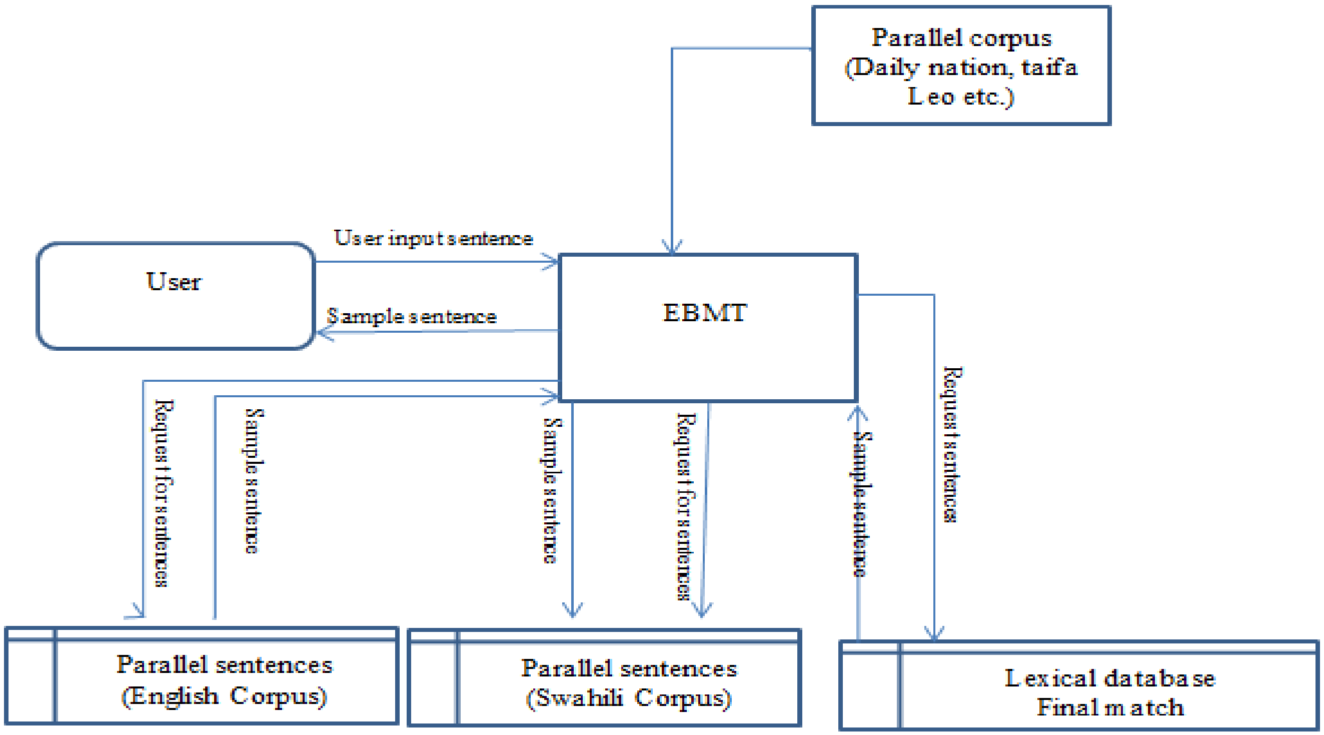
Methodology

- Our aim was to build an easy to use translator of where user will contribute spontaneously in building lexical sentence in the languages they know.
- We expect users to send monolingual search requests in language supported by our system to get multilingual answers.
- Through the use of our search engine user will extract their requests and will be able to add the new searches to the dictionary spontaneously.
- We chose to use Iterative design as it is based on a cyclic process of prototyping, testing, analyzing, and refining a product or process

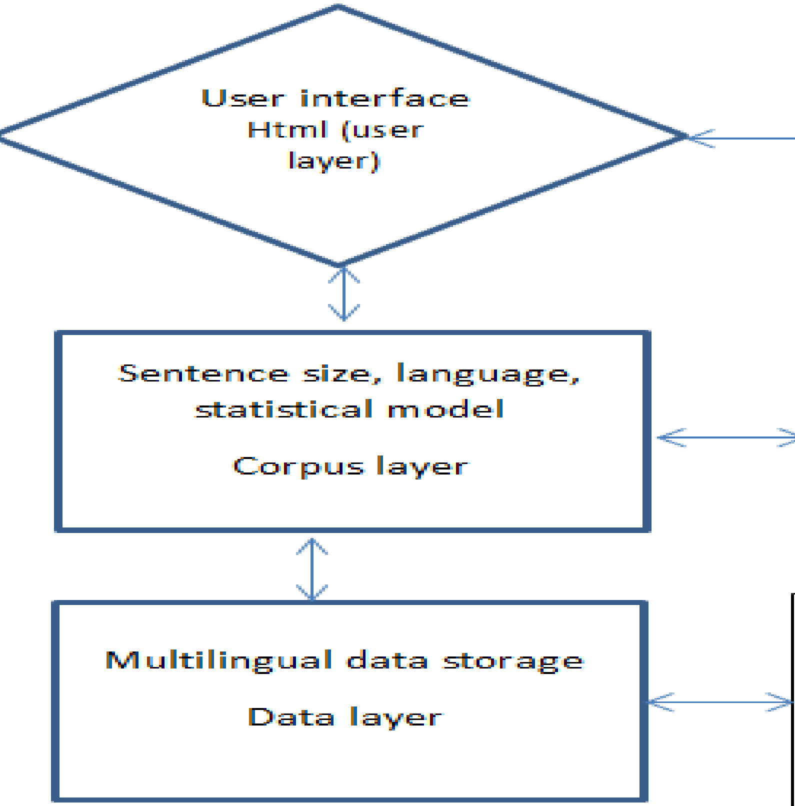
Functionality

- Content management this enables the *user to organize, modify content, and deleting as well maintenance of files and data from* news websites where primary sentence (English) and secondary sentence (Kiswahili) are extracted.
- Machine translation for English and Swahili

Structure of the EBMT



Multilingual Translation Structure



User interface Html (user layer)

Sentence size, language, statistical model

Corpus layer

Multilingual data storage

Data layer

User layer

User layer interact with users by a simple set of HTML pages and web forms that provide sentences retrieved at the corpus layer.

Corpus layer

At this layer corpus are preprocessed in order to get additional linguistic information about the source language, e.g. frequencies of collocations while tokeniser takes a SL sentence as input. One of its tasks is the separation of words and punctuation adding tags to words and sentences. And use statistical models n-gram and levenshtein algorithm to find similar sentences in Swahili to the source sentence in English.

Data layer

At the data layer, multilingual dictionary entries were indexed into a database, which give dictionary lock-up facilities, the structure can be changed dynamically to include any kind of information needed later. Its simplicity make it easy for users to learn, their searches are automatically transfer into an XML document and indexed into the dictionary.

This provides a variety of Web results and sentences to help users get context for a word or phrase.

Findings / Results

. we conducted our experiment two bilingual corpora both containing translations examples of about 3000 sentence the system was tested in all aspect and also the effect of the topic classifier

n the following aspect

. The comprehensiveness of sentence retrieved from multiple resources, conversion to a desired format and integration to the multilingual database.

. The accuracy of extracted sentence considering similarity measure

. use of classifying text into their domain/topic did show some improvement.

Conclusions

- The system developed has demonstrated a promising potential for using sentence similarity in an example-based machine translation
- sentence provided better performance
- we were able to solve the problem of consistency in translation by using these tool based on translation memories
- We also made it possible to reuse old translations stored as translation memories of previous versions of handbooks and thereby reducing the chances of producing variant translations of the same source sentence improving on the quality of the translation memories that are being put to use

Recommendations

1. From our tool we are able find new sentences in a parallel corpus of comparable html documents, which performed pretty well in terms of precision
2. To find different techniques for building a new classifier for extracting sentences equivalents from a corpus of comparable html documents.

Areas for further study

- ext to speech
- quality of translations produced
- Vord ambiguity
- emantic similarity
- tructure similarity

End