Clause Identification and Sentiment Tagged Parallel Corpus Preparation

Project submitted to

FACULTY OF ENGINEERING AND TECHNOLOGY

JADAVPUR UNIVERSITY

In partial fulfillment of the requirements for the degree of

MASTER OF COMPUTER APPLICATIONS, 2018

BY

Amrita Chandra

Examination Roll: MCA186012

Registration No: 133674 of 2015-2016

Under the guidance of

Prof. (Dr.) Dipankar Das

Assistant Professor, Department of Computer Science & Engineering

Jadavpur University

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

FACULTY OF ENGINEERING AND TECHNOLOGY

JADAVPUR UNIVERSITY

TO WHOM IT MAY CONCERN

I hereby recommend that the project entitled "Clause Identification and Sentiment Tagged Parallel Corpus Preparation" prepared under my supervision and guidance at Jadavpur University, Kolkata by AMRITA CHANDRA *(Reg. No. 133674 of 2015 – 16, Class Roll No. 001510503012 of 2015-16), may be accepted in partial fulfillment for the degree of Master of Computer Applications in the Faculty of Engineering and Technology, Jadavpur University, during the academic year 2017 – 2018. I wish her every success in life.*

…………………………………

Prof. (Dr.) Ujjwal Maulik
Head of the Department
Department of Computer Science and Engineering
Jadavpur University, Kolkata – 700032.

……………………………………

Prof. (Dr.) Dipankar Das
Project Supervisor,
Department of Computer Science and Engineering
Jadavpur University, Kolkata – 700032.

……………………………………

Prof. (Dr.) Chiranjib Bhattacharjee
Dean, Faculty council of Engg. & Tech.
Jadavpur University, Kolkata – 700032.

DECLARATION OF ORIGINALITY AND COMPLIANCE OF ACADEMIC PROJECT

I hereby declare that this project contains literature survey and original research work by the undersigned candidate, as part of her MASTER OF COMPUTER APPLICATIONS studies. All information in this document have been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material results that are not original to this work.

NAME: AMRITA CHANDRA

ROLL NUMBER: 001510503012

PROJECT TITLE:

CLAUSE IDENTIFICATION AND SENTIMENT TAGGED PARALLEL CORPUS PREPARATION

SIGNATURE WITH DATE

JADAVPUR UNIVERSITY
FACULTY OF ENGINEERING AND TECHNOLOGY

CERTIFICATE OF APPROVAL

The forgoing project is hereby accepted as a credible study of an engineering subject carried out and presented in a manner satisfactory to warrant its acceptance as a prerequisite to the degree for which it has been submitted. It is understood that by this approval the undersigned do not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn therein, but approve the project only for the purpose for which it is submitted.

FINAL EXAMINATION FOR
EVALUATION OF PROJECT: 1.________________________

2.________________________

(Signature of Examiners)

ACKNOWLEDGEMENT

I express my honest and sincere thanks and humble gratitude to my respected teacher and guide *Prof. (Dr.) Dipankar Das*, Assistant Professor of the Department of Computer Science & Engineering, Jadavpur University, for his exclusive guidance and entire support in completing and producing this project successfully. I am very much indebted to him for the constant encouragement, and continuous inspiration that he has given to me. The above words are only a token of my deep respect towards him for all he has done to take my project to the present shape.

I would like to thank *Mr. Sainik Kumar Mahata* for valuable support and suggestions to the activities of the project.

Finally, I convey my real sense of gratitude and thankfulness to my family members, specially my elder sister, for being an endless source of optimism and positive thoughts; and last but not the least, my father & mother for their unconditional support, without which I would hardly be capable of producing this huge work.

Amrita Chandra
Examination Roll: MCA186012
Registration No: 133674 of 2015 – 2016

Content

Sl. No.	Topic	Page No.
1	**Introduction**	1-2
	1.1 Background	1
	1.2 Applications	1
	1.3 Challenges	1
	1.4 Problem Statement	2
	1.5 Contribution	2
2	**Literature Survey**	3-6
3	**Methodology**	7-14
	3.1 Overall Architecture	7
	3.2. Collection of English Sentences	7-8
	3.3. Translation using Google Translate API	8
	3.4. Data Preparation	8-9
	3.4.1. POS tagging and Shallow Parsing	8-9
	3.5. Clause Identification	9-11
	3.5.1. Identification of Simple Sentences	9-10
	3.5.2. Identification of Complex and Compound sentences	10-11
	3.6. Phrase Identification	11-12
	3.7. Sentiment Annotation	12-14
	3.7.1. Word Level Sentiment Tagging	13
	3.7.2. Sentiment Annotated Parallel Corpus	13-14
4	**Results and Evaluation**	14-15
	4.1. Observation and Error analysis	14
5	**Conclusions and Future Work**	16
	References	16-18

Abstract

Parallel corpus is a collection of bilingual sentence pair where every sentence is a translation of the other. Such corpus is very essential for a Machine Translation (MT) system to produce good and acceptable output. The purpose of the current work is to build a parallel corpus comprising of English as the source language and Bengali as the target language. We tend to make use of Bengali as it is a low resource language and parallel corpus of a large size is not readily available. We have used Google Translator API for the translation task. Also, we have annotated the prepared corpus with sentiment. We have also tried and identified the noun, verb, adjective, adverb and prepositional phrases. Lastly, we have devised a method by which we will classify the extracted sentences as Simple or Other (Complex/Compound), by identifying the clause structure, and a provision to simplify the Complex/Compound sentences.

1. Introduction

1.1. Background

A parallel corpus is a collection of bilingual translated text. In simple words, if two languages are involved: the source monolingual text is an exact translation of the target monolingual text. Unfortunately, these resources are often scarce, limited in size, and have limited language coverage. Parallel texts are an important resource in many Natural Language Processing (NLP) applications such as MT.

MT is phenomena by which, in semantic level, machine translates one language to another. Due to this approach, the translation quality takes a hit as state-of-art approaches don't dwell into pragmatic level when translating. We thought of indulging MT to pragmatic features such that the quality of translation improves.

1.2. Applications

By solving the problem in hand, the following application areas can be approached.

- Preparation corpus for training Statistical/Neural Machine Translation systems.
- Additionally, to check whether a machine can translate sentiment or not.
- Production of multilingual lexical or semantic resources such as dictionaries or Ontologies.
- Training and testing of multilingual information extraction software.
- Checking the consistency of Automatic translation.
- Training of multilingual subject domain classifiers.
- Testing and bench-marking of sentence alignment software.
- Terminology extraction.

1.3. Challenges

We know that parallel corpus has a major impact on SMT and NMT. The performance of MT should not be checked from pure linguistic perspective rather to increase the performance quality sentiments can be added. Sentiments express the attitude and emotional condition of the speaker. So sentiments play a major role in MT. Our goal is to check whether sentiments are really propagated through the MT processing stages and make the translations more fluent as well as adequate or not. It is obvious that during translation, if sentiment of the source text is propagated to the translated text, it will lead to better translation. That is why to enhance the performance quality of MT; here we have prepared a sentiment tagged parallel corpus.

1.4. Problem Statement

We have tried and created a sentiment tagged parallel corpus, with English and Bengali as Source and Target language, that can be of help when dealing with improving the quality MT.

Also, to complement our problem statement, we have identified the simple sentences from complex and compound sentences and also devised a way to simplify a Complex/Compound sentence. Additionally, we have identified the noun, verb, adverb, adjective and prepositional phrases to create a supplementary resource.

1.5. Contribution

- We have created a English sentence corpus of 115037 sentences consists of simple, complex and compound sentences.
- We translated the English sentences into Bengali using Google Translate API. Then the Parts of Speech (POS) tagging is done using Shallow Parsing.
- We have identified the simple sentences from the phrase structure and analyzing the clause boundaries we have extracted the complex and compound sentences. After that the phrase identification is performed.
- We have built three types of parallel corpuses viz. general corpus, simple sentence corpus and others(complex, compound) sentence corpus after extracting the simple and others(complex, compound) sentences.
- The sentiment annotation is done on these three corpuses. We define some rule to build the sentiment annotated parallel corpus and we have done some analysis on these sentiment annotated parallel corpuses.

2. Literature Survey

Various works has already been done on Parts of Speech (POS) tagging, Shallow Parsing, clause boundary identification, text simplification, sentiment analysis and generation of parallel corpus.

Pattabhi R K et. al. [1] designed a generic hybrid POS tagger for Indian languages. They noted that Indian languages are characterized as free word ordered morphologically productive and agglutinative languages. In this hybrid implementation they have used combination of statistical approach (HMM) and rule based approach. The tag set used was developed by IIIT, Hyderabad and it consisted of 26 tags. They also devised a transformational - based learning (TBL) approach for text chunking. In this technique of chunking, a single base rule (or a few base rules) is provided to the system, and the other rules are learned by system itself during the training phase for reorganization of the chunks. They worked for three Indian languages namely Hindi, Bengali and Telugu. The corpus used for training was provided by SPSAL workshop. The results obtained vary for each language, but are encouraging.

G.M. Ravi Sastry et. al. [2] worked on building an HMM based POS tagger and statistical chunker for 3 Indian languages, viz., Bengali, Hindi and Telugu. They employed the TnT[1] POS tagger for tagging their corpus. The POS tagging accuracies for Bengali, Hindi and Telugu are 74.58, 78.35 and 75.37 respectively. For chunking, they used the training data to extract chunk pattern templates defined as a sequence of POS tags. These templates, in conjunction with the POS tag of the word following the chunk, are used to decide chunk boundaries in the unannotated text. A dynamic programming algorithm is used for finding the best possible chunk sequence. The chunk accuracies obtained are 67.52, 69.98 and 68.32 for Bengali, Hindi and Telugu respectively. The techniques used were generic and are expected to produce comparable accuracies for different languages.

Delip Rao et. al. [3] worked on shallow parsing of several Indian languages utilizing Conditional Random Field models. They showed how performance can be substantially improved by enhancing several features and modeling techniques, including expanding the chunk tag inventory and separating punctuation from linguistic phrases. They also reported results for POS tagging of Hindi, Bengali and Telugu using generative methods.

Avinesh.PVS et. al. [4] worked on POS tagging and Chunking using Conditional Random Fields (CRFs) and Transformation Based Learning (TBL) for Telugu, Hindi and Bengali. They showed that training CRFs can help to achieve good performance over any other Machine Learning (ML) techniques. Improved training methods based on the morphological information, contextual and the lexical rules (developed using TBL) were critical in achieving good results. The CRF and TBL based POS tagger has an accuracy of about 77.37%, 78.66%, and 76.08% for Telugu, Hindi and Bengali, and the chunker performs at 79.15%, 80.97% and 82.74% for Telugu, Hindi and Bengali respectively.

[1] http://www.coli.uni-saarland.de/~thorsten/tnt/

Asif Ekbal et. al. [5] reported a POS tagger based on the Hidden Markov Model (HMM) and a rule-based chunker on three languages-Bengali, Hindi and Telegu.

Erik F. Tjong Kim Sang et. al. [6] worked on dividing texts into syntactically related non-overlapping groups of words, a so-called text chunking. They gave background information on the data sets, presented a general overview of the systems and discussed their performance.

Sarah E. Petersen et. al. [7] worked on text simplification for language learners. Simplified texts are commonly used by teachers and students in bilingual education and other language-learning contexts. Their goal was the development of tools to aid teachers by automatically proposing ways to simplify texts. Their paper presents a detailed analysis of a corpus of news articles and abridged versions written by a literacy organization in order to learn what kinds of changes people make when simplifying texts for language learners.

Claire Cardie et. al. [8] found out that finding simple, non-recursive, base noun phrases are an important subtask in many natural language processing applications. They presented a corpus-based approach for finding base NPs by matching part-of- speech tag sequences. The training phase of the algorithm was based on two successful techniques: first the base NP grammar is read from a Treebank[2] corpus; then the grammar is improved by selecting rules with high benefit scores. Using this simple algorithm with a naive heuristic for matching rules, they achieved surprising accuracy in an evaluation on the Penn Treebank Wall Street Journal.

R. Vijay Sundar Ram et. al. [9] worked on the detection of clause boundaries using a hybrid approach. The Conditional Random fields (CRFs), which have linguistic rules as features, identified the boundaries initially. The boundaries marked were checked for false boundary marking using Error Pattern Analyzer. The false boundary markings were re-analyzed using linguistic rules. The experiments done with their approach showed encouraging results and are comparable with the other approaches.

Erik F. Tjong Kim Sang et. al. [10] used seven machine learning algorithms for one task: identifying base noun phrases. The results were processed by different system combination methods and all of these outperformed the best individual result. They have applied the seven learners with the best combinatory which is a majority vote of the top five systems to a standard data set and managed to improve the best published result for this data set.

Kerstin Denecke [11] introduced a methodology for determining polarity of text within a multilingual framework. The method leveraged on lexical resources for sentiment analysis available in English SentiWordNet[3]. First, a document in a different language than English was translated into English using standard translation software. Then, the translated document was classified according to its sentiment into one of the classes “positive” and “negative”. For sentiment classification, a document is searched for sentiment bearing words like adjectives. By means of SentiWordNet, scores for positivity and negativity were determined for these words.

[2] https://catalog.ldc.upenn.edu/ldc99t42

[3] http://sentiwordnet.isti.cnr.it/

An interpretation of the scores then led to the document polarity. The method was tested for German movie reviews selected from Amazon and is compared to a statistical polarity classifier based on n-grams. The results showed that working with standard technology and existing sentiment analysis approaches was a viable approach to sentiment analysis within a multilingual framework.

Aurangzeb khan et. al. [12] proposed the rule based domain independent sentiment analysis method. The proposed method classified subjective and objective sentences from reviews and blog comments. The semantic score of subjective sentences was extracted from SentiWordNet to calculate their polarity as positive, negative or neutral based on the contextual sentence structure. The results showed the effectiveness of the proposed method and it outperformed the machine learning methods. The proposed method was achieved an accuracy of 87% at the feedback level and 83% at the sentence level for comments and 97% at feedback and 86 % at sentences for customer reviews.

Federico Zanettin [13] worked on how small bilingual corpora of either general or specialized language can be used to devise a variety of structured and self-centered classroom activities whose aim was to enhance the understanding of the source language text and the ability to produce fluent target language texts.

Colin Bannard et. al. [14] worked on Using alignment techniques from phrase based statistical machine translation, they showed how paraphrases in one language can be identified using a phrase in another language as a pivot. They define a paraphrase probability that allows paraphrases extracted from a bilingual parallel corpus to be ranked using translation probabilities, and show how it can be refined to take contextual information into account. They have evaluated their paraphrase extraction and ranking methods using a set of manual word alignments, and contrast the quality with paraphrases extracted from automatic alignments.

Daniel Varga et. al. [15] worked on e a general methodology for rapidly collecting, building, and aligning parallel corpora for medium density languages, illustrating their main points on the case of Hungarian, Romanian, and Slovenian. They have also described and evaluated the hybrid sentence alignment method which they are using.

Philip Resnik et. al. [16] worked on using the STRAND [21, 22] system for mining parallel text on the WorldWideWeb (WWW). They first reviewed the original algorithm and results and then they presented a set of significant enhancements. These enhancements include the use of supervised learning based on structural features of documents to improve classification performance, a new content based measure of translational equivalence, and adaptation of the system to take advantage of the Internet Archive for mining parallel text from the Web on a large scale. Finally, the value of these techniques is demonstrated in the construction of a significant parallel corpus for a low-density language pair.

Constantin Oraˇsan [17] proposed a hybrid method for clause splitting in unrestricted English texts which required less human work than existing approaches. The results of a machine learning algorithm, trained on an annotated corpus, were processed by a shallow rule-based module in order to improve the accuracy of the method. The evaluation of the results showed that the machine learning algorithm is useful for identification of clause's boundaries and the rule-based module improved the results. Using some very simple rules they reported precision of around 88%.

Stefano Baccianella et. al. [18] worked on presenting SENTIWORDNET 3.0, a lexical resource explicitly devised for supporting sentiment classification and opinion mining applications. SENTIWORDNET 3.0 was an improved version of SENTIWORDNET 1.0 which is a lexical resource publicly available for research purposes currently licensed to more than 300 research groups and used in a variety of research projects worldwide.

Himanshu Agrawal[19] proposed his approach for a part of speech tagger and chunker for South Asian Languages. They have used a Conditional Random Fields based approach to train the system on the corpus made available by the SPSAL workshop at ICJAI 2007. They have worked on improving the machine's learning without using any language specific tools like dictionaries, morphological analyzers etc. Apart from the annotated training data they have also used a large raw unannotated text. The average performance figures over all the three languages were 79.13% for POS tagging and 92.36% for chunking over the 3 languages. The highest was being 84.90 % for Hindi.

Santanu Pal et. al. [20] worked on how sentiment analysis can improve the translation quality by incorporating the roles of sentiment holders, sentiment expressions and their corresponding objects and relations. We also demonstrated how a simple baseline phrase based statistical MT (PB-SMT) system based on the sentiment components can achieve 33.88% relative improvement in BLEU for the under-resourced language pair English-Bengali.

3. Methodology

3.1. Overall Architecture

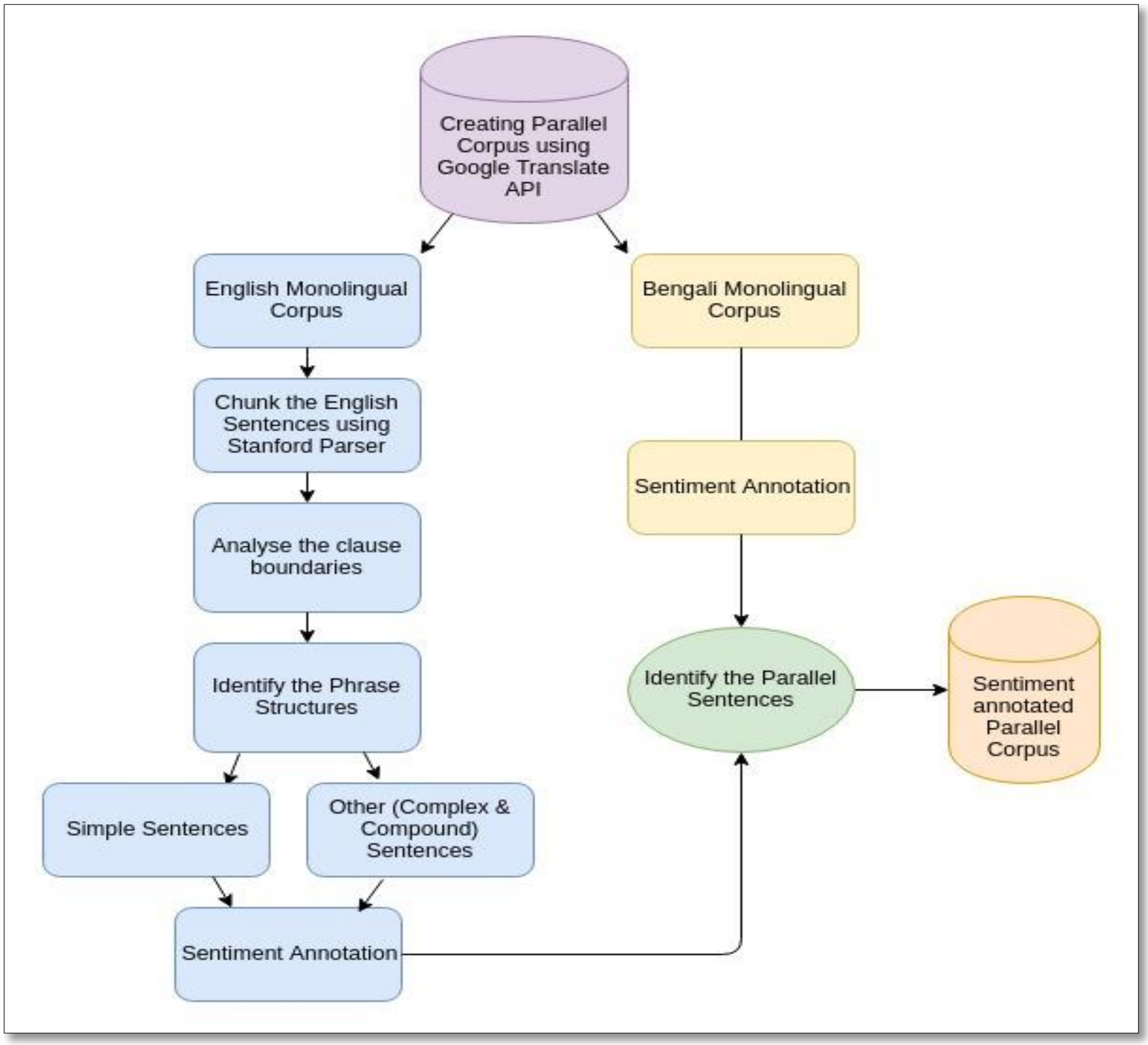
Creating Parallel Corpus using Google Translate API
English Monolingual Corpus
Bengali Monolingual Corpus
Chunk the English Sentences using Stanford Parser
Sentiment Annotation
Analyse the clause boundaries
Identify the Phrase Structures
Identify the Parallel Sentences
Sentiment annotated Parallel Corpus
Simple Sentences
Other (Complex & Compound) Sentences
Sentiment Annotation

Fig 2: Architecture followed for the methodology.

3.2. Collection of English Sentences

A wide array of different types of corpora has been constructed for use in the field of machine translation. They reflect the criteria according to which they are designed and the purpose for which they are created. Such a corpus used in translation is a bilingual corpus. Language pairs are put together either on the basis of "parallelism" or/and "comparability." Parallel bilingual corpora consist of texts in language *"A"* and their translation into language *"B"*, or vice versa. The relationship between texts is directional, i.e. it goes from one text; the source language (SL) text to the target language (TL) text. To prepare such a parallel corpus; for English as SL and Bengali as TL; we collected 7053 English sentences from various websites. In addition, we obtained 57985 English sentences from the resource of Machine Translation in Indian Languages (MTIL) shared task[4], organized by Amrita University and 49999 English sentences from Technology Development for Indian Languages Programme (TDIL)[5]. This is shown in Table 1.

Source	**Data Size**
Various websites	7053
Amrita University	57985
TDIL	49999
Total	**115037**

Table 1: Data Information Table

3.3. Translation using Google Translate API

We translated the English sentences into Bengali using Google Translate API[6]. We have done these translations for three type of corpora viz. general, simple sentence and others (complex, compound) sentences. Then the English sentences and their Bengali translations are merged in parallel order in a file.

3.4. Data Preparation

3.4.1. POS tagging and Shallow Parsing

In corpus linguistics, POS tagging, also called grammatical tagging or word-category disambiguation is the process of marking a word in the text as its corresponding part of speech, based on both its definition and its context. A simplified form of this is commonly taught to school-age children, for identification of words as nouns, verbs, adjectives, adverbs, etc.

Shallow Parsing is an analysis of a sentence in which constituent parts of sentences (nouns, verbs, adjectives, etc.) are identified and then higher order units that have discrete grammatical meanings (noun groups or phrases, verb groups, etc.) are linked. While the most elementary parsing algorithms simply link constituent parts on the basis of elementary search patterns (e.g. as specified by Regular Expressions), approaches that use machine learning techniques (classifiers, topic modeling, etc.) can take contextual information into account and thus compose parses in such a way that they better reflect the semantic relations between the basic constituents. We have used Natural Language Toolkit (NLTK)[7] and Stanford Dependency Parser[8] for performing the POS tagging and shallow parsing. Example of POS tagging and shallow parsing is given in Table 2.

[4] http://nlp.amrita.edu/mtil_cen/

[5] http://tdil.meity.gov.in/

[6] https://translate.google.com/

[7] http://www.nltk.org/

[8] https://nlp.stanford.edu/software/stanford-dependencies.shtml

Sentence before parsing	After shallow parsing
Initially, it ran on 6 routes which joined most of Delhi's parts.	S (ADVP (RB initially)) (, ,) (NP (PRP it)) (VP (VBD ran) (PP (IN on) (NP (NP (CD 6) (NNS routes)) (SBAR (WHNP (WDT which)) (S (VP (VBD joined) (NP (NP (JJS most)) (PP (IN of) (NP (NP (NNP Delhi) (POS 's)) (NNS parts))))))))))) (. .))
This place is also the part of UNESCO's world heritage.	S (NP (DT This) (NN place)) (VP (VBZ is) (ADVP (RB also)) (NP (NP (DT the) (NN part)) (PP (IN of) (NP (NP (NNP UNESCO) (POS 's)) (NN world) (NN heritage))))) (. .))

Table 2: Example of POS tagging by parsing

3.5. Clause Identification

3.5.1. Identification of Simple Sentences

A simple sentence in this context is defined as a sentence which contains only one independent clause and has no dependent clauses. Generally, whenever two or more clauses are joined by conjunctions (coordinating and subordinating), it becomes a complex or a compound sentence accordingly, to get a hold on handling the conjunctions, we used the dependency parser to chunk the English sentences into phrases. (viz. NP (Noun Phrase), VP (Verb Phrase), PP (Preposition Phrase), ADJP (Adjective Phrase) and ADVP (Adverb Phrase)).

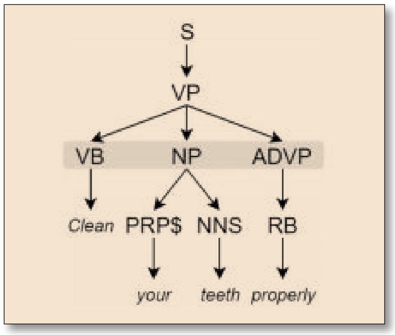

Figure 2: Extraction of phrase chunks.

We noticed that, simple sentences have a unique phrase structure that consists of combinations of NP, VP and PP. In conjunction to this theory, we applied a rule based approach to extract simple sentences from the English corpus.

We subjected 3046 simple sentences to chunking, using Stanford dependency parser, and extracted the unique phrase structures for the rules by which we further mined for simple sentences from the English corpus. We extracted 205 unique rules, the surface forms of which are shown in Table 3.

Extracted Rules
PP NP* PP VP NP*
PP NP* VP PP NP*
ADVP NP* VP* ADVP NP*
NP VP PP NP PP NP
NP ADVP VP* NP*
NP* VP NP*
NP* PP NP VP* NP
NP VP PP NP*
NP VP* NP* PP* ADJP* ADVP*

Table 3: Surface forms of the extracted rules where "*" means one or more occurrence of an item.

We tested our system on 2876 sentences (1438 simple sentences and 1438 complex/compound sentences) and got an accuracy of 89.22%. Table 4 shows the various validation metrics. We used this system to extract 16654 simple sentences from the generated English corpus.

	Other	Simple	Kappa
Other	1275	90	0.78
Simple	220	1291	
Prec.	93.41%		
Recall	85.28%		
Acc.	89.22%		
F1	89.16%		

Table 4: Confusion matrix for the rule based approach.

3.5.2. Identification of Complex and Compound sentences

The POS tag assigned to every token by the POS tagger is used to discover these positions. The chunk boundaries are identified by some handcrafted linguistic rules that check whether two neighboring POS tags belong to the same chunk or not. If they do not, then a chunk boundary is assigned in between the words.

The phrase structures of the sentences are extracted from the shallow parsing into a file. Rules for extracting complex sentences are as follows.

- If a line in a shallow parsed file contains 'SBAR' tag in between two sentences then we can say that the sentence is complex sentence.
- If a line in a chunked file contains 'SBAR' in starting of the sentence and ',' in middle of the sentence then the sentence is considered as complex sentence.

Similarly, there are rules for extracting compound sentences. If the shallow parsed sentence has a 'CC (coordinating conjunction) ' tag followed by 'S (starting of sentence)' tag, then the sentence is a compound sentence. Symbolically we can define the rule as follows:-

$$CC_{POS} \rightarrow S \text{ (Starting of another sentence)}$$

There are 7 coordinating conjunctions. Below is the list of coordinating conjunctions. We also devised a way by which the extracted complex and compound sentences can be split into simple sentences.

- *For, Nor, Or, So, And, But, Yet*

Type	**Number of Sentences**
Simple	16654
Complex	39068
Compound	59315

Table 5: No. of sentences extracted using rule based approach

3.6. Phrase Identification

As mentioned above POS tagging is the task of assigning grammatical classes to words in a natural language sentence. Similarly chunking consists of dividing a text in syntactically correlated parts of words. Identifying the POS tags and chunk tags for the words in a given text is an important aspect in any language processing task. Both are important intermediate steps for full parsing. Table 6 shows the list of POS tags (for English) of Penn Treebank Project, used by NLTK.

POS tag	**Description**	**POS tag**	**Description**
CC	coordinating conjunction	PDT	predeterminer 'all the kids'
CD	cardinal digit	POS	possessive ending parent's
DT	Determiner	PRP	personal pronoun I, he, she
EX	existential there	PRP$	possessive pronoun my, his, hers
FW	foreign word	RB	adverb very, silently,
IN	preposition/subordinating conjunction	RBR	adverb, comparative better
JJ	adjective 'big'	RBS	adverb, superlative best
JJR	adjective, comparative 'bigger'	RP	particle give up
JJS	adjective, superlative 'biggest'	TO	to go 'to' the store.
LS	list marker 1)	UH	interjection
MD	modal could, will	VB	verb, base form take
NN	noun, singular 'desk'	VBD	verb, past tense took
NNS	noun plural 'desks'	VBG	verb, gerund/present participle taking
NNP	proper noun, singular 'Harrison'	VBN	verb, past participle taken
NNPS	proper noun, plural 'Americans'	VBP	verb, sing. present, non-3d take
VBZ	verb, 3rd person sing. present takes	WP	wh-pronoun who, what
WDT	wh-determiner which	WRB	wh-abverb where, when

Table 6: POS tag list

Using these POS tags and the chunked file noun, verb, adverb, adjective, prepositional phrases are identified. Phrase identification is very important task in various NLP applications. It also helps in evaluating the performance for MT. The following is the example of noun, verb, adverb, adjective, prepositional phrase identification:-

Sentence	Little children amuse easily.
Noun Phrase	Little
Verb Phrase	amuse
Preposition Phrase	[]
Adjective Phrase	[]
Adverb Phrase	easily
Sentence	She allowed us near the house.
Noun Phrase	She, us, the
Verb Phrase	allowed
Preposition Phrase	[]
Adjective Phrase	near
Adverb Phrase	[]

Table 7: Example of phrase identification

After identifying the phrases we have analyzed them i.e. we have counted the average number of noun, verb, preposition, adjective and adverb phrases in the corpus and is shown in Table 8.

Phrase	**Average value**
Noun	2.71
Verb	5.42
Preposition	8.13
Adjective	10.84
Adverb	13.55

Table 8: Average count of phrases

3.7. Sentiment Annotation

Opinion mining (sometimes known as sentiment analysis) refers to the use of natural language processing to systematically identify, extract, quantify, and study affective states and subjective information. Sentiment analysis is widely applied to voice of the customer materials such as reviews and survey responses, online and social media, and healthcare materials for applications that range from marketing to customer service to clinical medicine. Generally speaking, sentiment analysis aims to determine the attitude of a speaker, writer, or other subject with respect to some topic or the overall contextual polarity or emotional reaction to a document, interaction, or event. So sentiment annotation is very important in NLP.

Here we use sentiment annotation in parallel corpus to help increasing the performance of MT. Here English is the source language and Bengali is the target language. Our main aim is to develop a parallel corpus for low resource language such as Bengali since there are not many resources available for these language pairs and also we will compare the sentiment annotated sentences parallel and find out whether it is a word to word translation or generated translation. The following are the steps to achieve the above mentioned:-

3.7.1 Word Level Sentiment Tagging

- SentiWordNet of positive and negative words for English and Bengali[9] were applied to the English and Bengali files separately. We have also used AFINN-96[10], AFINN-111[11], Taboada Grieve 2004-SO[12], vendersentiment[13] of positive and negative words for English. This step was repeated for the general corpus, Simple sentence corpus and the "Other" (Complex, Compound) sentences. A snippet of the result of this step is shown in Table 9.

3.7.2. Sentiment Annotated Parallel Corpus

- Parallel sentences were found out by considering the following rule.

  R1) If the English sentence is having one or more 'POS' tag and its corresponding Bengali sentence is also having one or more 'POS' tag, they were considered as parallel.

  R2) if the English sentence is having one or more 'NEG' tag and its corresponding Bengali sentence is also having one or more 'NEG' tag, they were considered as parallel.

  R3) If the English sentence is having one or more 'POS' and 'NEG' tag and its corresponding Bengali sentence is also having one or more 'POS' and 'NEG' tag, they were considered as parallel.

  R4) If the English sentence is having one or more 'POS' and 'NEG' tag and its corresponding Bengali sentence is also having one or more 'POS' tag, and vice versa, they were considered as parallel.

  R5) If the English sentence is having one or more 'POS' and 'NEG' tag and its corresponding Bengali sentence is also having one or more 'NEG' tag, and vice versa, they were considered as parallel.

- We also analyzed that whether it is a direct translation (word to word translation) or it is a general translation.

[9] http://amitavadas.com/sentiwordnet.php

[10] http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=5981

[11] https://udel.instructure.com/courses/1310886/files/57548375

[12] https://www.coursehero.com/file/p76pc0gm/Taboada-and-Grieve-2004-improved-performance-of-an-earlier-version-of-the-SO/

[13] https://books.google.co.in

English Sentiment annotated sentence	Bengali Sentiment annotated sentence
The tourists’ admired\POS the paintings.	পর্যটকরা পেইন্টিংসকে প্রশংসিত\POS করেছে।
I admired\POS him for his honesty.	আমি তার সততা\POS জন্য তাকে প্রশংসিত।
The enemy\NEG soldiers submitted to us.	শত্রু\POS \NEG সৈন্য আমাদের জমা\POS \NEG দেওয়া।
Ellen warned\NEG Helen that the party\POS would be tonight.	এলেন হেলেনকে সতর্ক\POS \NEG করে\POS দিয়ে বলেন যে পার্টি আজ রাতে হবে।

Table 9: Example of Sentiment Annotation

4. Results and Evaluation

We have collected 115037 English sentences and then translated it into Bengali using Google translate API. Then the sentences are parsed using Stanford parser and NLTK. After that analyzing the clause boundaries and training by the phrase structure of the simple sentence we have identified the simple and others (complex, compound) sentences. Next step is to identify the noun, verb, adverb, adjective and prepositional phrases. These phrases are identified using the parsed file and with the help of POS tag list. After that the sentiment annotation is performed. Sentiment annotation is performed in three different parallel corpuses i.e. general, simple and others (complex, compound) corpuses. Table 10 quantifies the various observations related to the sentiment tagged parallel corpus generation.

4.1 Observation and Error Analysis

The table below gives us the statistics of the parallel corpuses. We have total of 115037 sentences, out of which 16654 simple sentences and 98383 others (complex, compound) sentences. From these parallel corpuses we have separated the sentiment annotated sentences and built a parallel corpus of these sentiments annotated sentences. After that we have calculated no. of positive and negative words in a sentence and comparing English sentences with their Bengali translations for the positive and negative value=1, 2 and 3.

Here we only deal with the simple, complex and compound sentences. So our experiment cannot detect the mixture of complex and compound sentences.

Parallel corpus Type	**Total no. of sentences**	**No. of parallel sentiment annotated sentences**	**English**	**Bengali**	**No. of sentences**	**English**	**Bengali**	**No. of sentences**
			Positive	**Positive**		**Negative**	**Negative**	
General	115037	70358	1	1	14289	1	1	12861
			2	1	5936	2	1	5554
			3	1	1914	3	1	1896
			1	2	7653	1	2	5361
			2	2	4797	2	2	3256
			3	2	2089	3	2	1521
			1	3	2891	1	3	1594
			2	3	2452	2	3	1352
			3	3	1386	3	3	801
Simple	16654	6700	1	1	2153	1	1	1717
			2	1	472	2	1	430
			3	1	86	3	1	61
			1	2	837	1	2	470
			2	2	334	2	2	153
			3	2	62	3	2	33
			1	3	183	1	3	79
			2	3	104	2	3	39
			3	3	33	3	3	11
Others	98383	63619	1	1	12121	1	1	11140
			2	1	5462	2	1	5124
			3	1	1828	3	1	1835
			1	2	6810	1	2	4889
			2	2	4463	2	2	3103
			3	2	2027	3	2	1488
			1	3	2705	1	3	1513
			2	3	2348	2	3	1310
			3	3	1353	3	3	790

Table 10: Statistics of generated sentiment tagged parallel corpus.

5. Conclusions and Future Work

In this report we have developed a sentiment annotated parallel corpus for MT to improve its performance and to enhance the performance quality of MT i.e. whether it results correctly or not for a sentiment annotated parallel corpus and also we have identified clauses. Here the source language is English and the target language is Bengali. We choose these languages since Bengali are a low resource language and there are not many resources available. Here we have collected the sentences then translate them into Bengali. Then the sentences are POS tagged by performing shallow parsing. We have identified the simple and others (complex, compound) sentences by identifying the clause boundaries. After that the phrase identification is done. Then the sentiment annotation is performed for three parallel corpuses i.e. general, simple and others (complex, compound) corpuses respectively and then we have done some analysis on these sentiment tagged parallel corpuses.

As a future prospect, we will try and classify sentences as Simple, Complex and Compound using deep learning architectures. Our experiment will help in evaluating and enhancing the performance quality of MT.

References

[1] T Pattabhi R K Rao, Vijay Sundar Ram R, Vijayakrishna R and Sobha L.A Text Chunker and Hybrid POS Tagger for Indian Languages,In the Proceedings of Shallow Parsing for South Asian Languages, pages 9-12,2007

[2] Sastry G.M. Ravi , Sourish Chaudhuri and P. Nagender Reddy. An HMM based Part-Of-Speech tagger and statistical chunker for 3 Indian languages, In the Proceedings of Shallow Parsing for South Asian Languages, pages 13-16, 2007

[3] Rao Delip , David Yarowsky.Part of Speech Tagging and Shallow Parsing of Indian Languages,In the Proceedings of Shallow Parsing for South Asian Languages, pages 17-20, 2007

[4] PVS Avinesh., Karthik G.Part-Of-Speech Tagging and Chunking using Conditional Random Fields and Transformation Based Learning, In the Proceedings of Shallow Parsing for South Asian Languages, pages 21-24,2007

[5] Ekbal Asif , Samiran Mandal , Sivaji Bandyopadhyay. POS Tagging Using HMM and Rule-based Chunking, In The Proceedings of SPSAL,pages-25 to 28, 2007

[6] Sang Erik F. Tjong Kim , Sabine Buchholz.Introduction to the CoNLL-2000 Shared Task: Chunking, In the Proceedings of CoNLL-2000 and LLL-2000, pages 127–132, Lisbon, Portugal, pages 127-132, 2000

[7] Petersen Sarah E., Mari Ostendorf.Text Simplification for Language Learners: A Corpus Analysis, Speech and Language Technology in Education (SLaTE 2007),pages 69-72, 2007

[8] Cardie Claire, David Pierce.Error-Driven Pruning of Treebank Grammars for Base Noun Phrase Identification,In the Proceeding COLING '98 Proceedings of the 17th international conference on Computational linguistics - Volume 1,Proceeding ACL '98 Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1, pages 218-224,1998

[9] Ram R. Vijay Sundar, Sobha Lalitha Devi.Clause Boundary Identification Using Conditional Random Fields, A. Gelbukh (Ed.): CICLing 2008, LNCS 4919, pp. 140–150, 2008. © Springer-Verlag Berlin Heidelberg 2008

[10] Sang Erik F. Tjong Kim, Walter Daelemans, Herv´e D´ejean, Rob Koeling, Yuval Krymolowski, Vasin Punyakanok, Dan Roth.Applying System Combination to Base Noun Phrase Identification, In the Proceedings of COLING 2000, pages 857–863, Saarbru¨cken, Germany, 2000

[11] Denecke Kerstin.Using SentiWordNet for Multilingual Sentiment Analysis, Data Engineering Workshop, 2008. ICDEW 2008. IEEE 24th International Conference on, pages 507-512, 2008

[12] Khan Aurangzeb, Baharum Baharudin.Sentiment Classification by Sentence Level Semantic Orientation using SentiWordNet from Online Reviews and Blogs, International Journal of Computer Science \& Emerging Technologies, pages 539-552, 2011

[13] Zanettin Federico.Bilingual Comparable Corpora And The Training Of Translators, Meta: Journal des traducteurs/Meta: Translators' Journal,pages 616-630,1998

[14] Bannard Colin, Chris Callison-Burch.Paraphrasing with Bilingual Parallel Corpora, In Proceedings of the 43rd Annual Meeting of the ACL, pages 597–604, Ann Arbor, June 2005. C 2005 Association for Computational Linguistics

[15] Varga D´aniel, P´eterHal´acsy , Andr´asKornai, ViktorNagy, L´aszl´oN´emeth, ViktorTr´on.Parallel corpora for medium density languages, Amsterdam Studies In The Theory And History Of Linguistic Science Series 4, pages 247, 2007

[16] Resnik Philip ,Noah A. Smith.The Web as a Parallel Corpus, Computational Linguistics, pages 349-380, 2003

[17] Oraˇsan Constantin.A hybrid method for clause splitting in unrestricted English texts, In the Proceedings of ACIDCA’2000, 2000

[18] Baccianella Stefano, Andrea Esuli, Fabrizio Sebastiani.SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining, LREC,pages 2200-2204, 2010

[19] Agrawal Himanshu.POS tagging and Chunking for Indian Languages, Shallow Parsing for South Asian Languages, pages 37-40, 2007

[20] Pal Santanu, Braja Gopal Patra, Dipankar Das, Sudip Kumar Naskar , Sivaji Bandyopadhyay, Josef van Genabith. How Sentiment Analysis Can Help Machine Translation, In the Proceedings of the 11th International Conference on Natural Language Processing, pages 89-94, 2014

[21] Philip Resnik 'Mining the Web for Bilingual Text' in Proceedings of the 37th annual meeting of the Association for Computational Linguistics (ACL), pages 527–534, University of Maryland, College Park, Maryland, 1999

[22] Philip Resnik, "Parallel Strands: A Preliminary Investigation into Mining the Web for Bilingual Text", in In 3rd Conference of the Association for Machine Translation in the Americas, pages 72–82, Springer, 1998