Rule based Sentence Simplification for English to Tamil Machine Translation System

Poornima C, Dhanalakshmi V
Computational Engineering and Networking
Amrita Vishwa Vidyapeetham

Anand Kumar M, Soman K P
Computational Engineering and Networking
Amrita Vishwa Vidyapeetham

ABSTRACT
Machine translation is the process by which computer software is used to translate a text from one natural language to another but handling complex sentences by any machine translation system is generally considered to be difficult. In order to boost the translation quality of the machine translation system, simplifying an input sentence becomes mandatory. Many approaches are available for simplifying the complex sentences. In this paper, Rule based technique is proposed to simplify the complex sentences based on connectives like relative pronouns, coordinating and subordinating conjunction. Sentence simplification is expressed as the list of sub-sentences that are portions of the original sentence. The meaning of the simplified sentence remains unaltered. Characters such as ('.','?') are used as delimiters. One of the important pre-requisite is the presence of delimiter in the given sentence. Initial splitting is based on delimiters and then the simplification is based on connectives. This method is useful as a preprocessing tool for machine translation.

Keywords
Sentence simplification, Sentence segmentation, POS tag, Machine translation

1. INTRODUCTION
Language is important for human communication. There are many countries where people speak more than one language. In India there are more than 18 constitutional languages such as Hindi, Bengali, Guajarati, Oriya, Punjabi, Telugu, Kannada, Tamil, Malayalam, etc. In India Hindi is the national language and English is the common language for all states. Even though Hindi is our national language, Hindi is spoken only in northern states but in the southern region especially in Tamil Nadu most of the people speak only in their regional language (i.e. Tamil). So, for better communication English to Indian language machine translation is necessary.

Machine translation is the process of translating a text from source language into a target language with the help of computers. The translation process converts a text in one human language to another which preserves not only the meaning, but also the form, effect and style. Nowadays most of the online information is available in English. In a multi-lingual society different languages are spoken in different regions. So, for this purpose machine translation is required. Research in machine translation is going on for 50 years but not yet succeeded. Machine translation will work for simple sentences but, machine translation system faces difficulty while translating long sentences and as a result the performance of the system degrades. Problems when dealing with long sentences often include embedded clauses such as relative clause. To overcome this problem, the long sentence is simplified into smaller sentences. Sentence simplification can be used as a pre-processing tool in several applications such as Natural Language Processing, Query Processing and Speech Processing. In Text Summarization sentence simplification is used to shorten the original Text without losing the meaning of the content. Recently, many splitting technique has been developed for machine translation systems from English to other languages.

Sentence simplification and segmentation can be performed by two approaches, either rule based or corpus based. In our paper rule based technique is used for simplification. In some research work such as Berger et al. and Takezawa recommended a word-sequence characteristic based technique. A few rule-based systems have been developed for text simplification. These systems contain a set of manually created simplification rules that are applied to each sentence. Satoshi Kamatani, Tetsuro Chino and Kazuo Sumita Proposed Hybrid Spoken Language Translation Using Sentence Splitting Based on Syntax Structure [5]. In our paper rule based technique is used to simplify the relative pronouns, coordinating and subordinating conjunctions to obtain simple sentences for machine translation.

2. PAGE SIZE
Sentence simplification is an important task in machine translation. Various approaches are available for sentence simplification. Zhemin Zhu, Delphine Bernhard, Iryna Gurevych [10] proposed A Monolingual Tree-based Translation Model for Sentence Simplification. In Input sentence splitting and translation, Takao Doi, Eiichiro Sumita uses splitting technique as a pre-processing method based on N-gram of POS subcategories for machine translation [1]. This method is derived from that of Lavie (1996) and modified especially for Japanese. Katsuhito Sudoh et al. proposed Divide and Translate: Improving Long Distance Reordering in Statistical machine translation [2]. Nakajima H. et al. proposed The Statistical Language model for Utterance Splitting in Speech. Chandrasekar R, Srinivas B. proposed Automatic induction of rules for text simplification [6]. David Vickrey and Daphne Koller use Sentence Simplification technique for Semantic Role Labeling. Jia Xu, et al. proposed Sentence Segmentation using IBM Word Alignment Model 1. Entity-Focused Sentence Simplification for Relation Extraction is proposed by Makoto Miwa, Rune Saetre, Yusuke Miyao, and Jun'ichiTsujii. Deepa Gupta used splitting technique in Contributions to English to Hindi machine translation using Example-Based Approach [3].

Kim and Ehara proposed a rule based method for splitting Japanese long sentences for Japanese to English translation. Sarah E. Petersen and Mari Ostendorf developed Text Simplification for language learners. Caroline Gasperin et al.

developed learning when to simplify sentences for natural text simplification. O. Furuse et al. proposed splitting ill formed input for robust Multi-lingual speech translation [7]. KatrinTomanek et al. developed Sentence and Token Splitting Based on Conditional Random Fields [4]. Orasan, C. proposed a hybrid method for clause splitting in unrestricted English texts.

3. RULE BASED SENTENCE SIMPLIFICATION

Sentence simplification is the process of simplifying the complex sentences into simpler sentences. The method proposed in this paper is simpler one. Based on rules, the sentences are simplified in order to get exact translation. When a clause stands on its own and is independent, it is called main clause. Subordinate clauses are those clauses which cannot stand alone but depend on main clause for their meaning. Most of the sentences contain conjunctions and sentences are split based on conjunctions. Independent clauses can be joined by a coordinating conjunction to form complex or compound sentences. Dependent clauses often begin with a subordinating conjunction or relative pronoun.

Our system handles coordinating conjunctions, subordinating conjunctions and relative pronouns. Coordinating conjunction includes for, and, not, but, or, yet and so. Subordinating conjunction includes after, although, because, before, if, since, that, though, unless, where, wherever, when, whenever, whereas, while, why. Relative pronoun includes who, which, whose, whom.

Framework

The proposed approach follows in following steps:

1. Split the sentences from the paragraph based on delimiters such as "." and "?"
2. Delimiters such as (comma, {,}, [,],) are ignored from the sentences.
3. Individual sentences are split based on coordinating and subordinating conjunction.

The text can be of any form i.e. paragraphing format, individual sentences, etc. Presence of delimiter such as (? and .) is an important pre-requisite as the initial splitting is done based on delimiters. The obtained individual sentences are parsed using Stanford parser. Stanford parser gives POS tag as well as dependency information; based on the information the rules are generated.

Our system deal with the following techniques,

1. Splitting
2. Simplification

Splitting is used to break the sentences which contain coordinating and subordinating conjunction whereas sentence simplification is to simplify the sentences which contain relative pronoun.

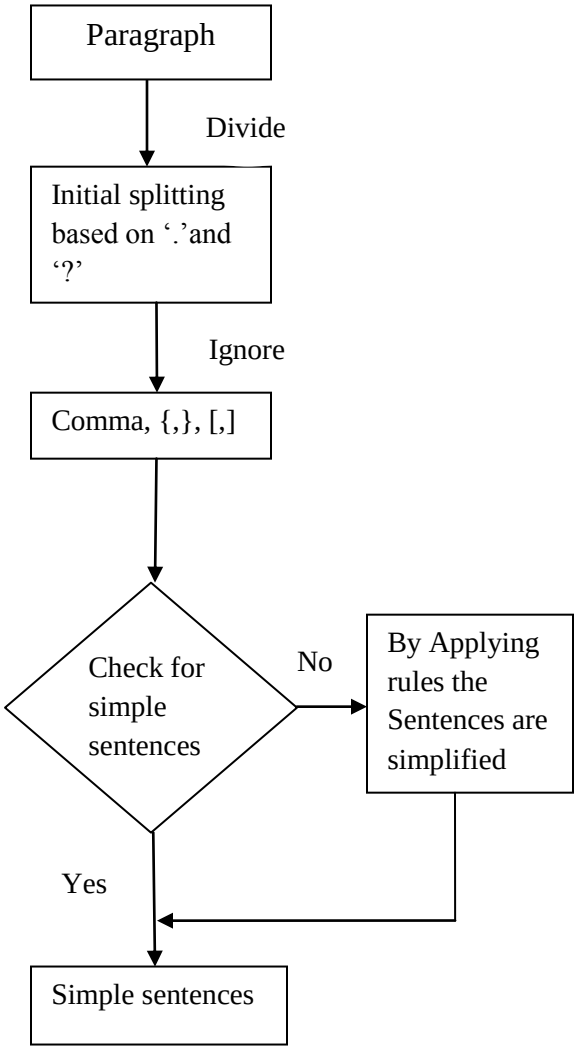

Fig 1: Frame work of our method

3.1 Algorithm for Sentence Simplification

There are several "wh" connectives available out of which "who, whom, which, whose" are dealt. In this case, the relative clause can occur either in between the main clause, or after the main clause. In both the cases, the connective words contain two possible dependency tags i.e. either "subject" or "object". The algorithm for all these possibilities is shown in Table 1.

Consider an example,

POS tag of Input sentence: The/ DT, people/ NNS, **who**/ WP, live/ VBP, in/ IN, Scotland/NNP, are/ VBP, called/ VBN, Scots/ NNS.

Our work is sentence simplification. The simplifying sentences will work for any translation system with English as source

language. From the translation point of view, here English to Tamil language is considered.

ஸ்காட்லேண்டில் வசிக்கும் மக்களை ஸ்காட்ஸ் என்று அழைப்பர்.

Here, 'who' is the subordinating conjunction. The sentence should be simplified based on 'who'. In the above example, two words are present before "who". Make ensure that any of these words contain "verb" tag. If so, then the sentence is not embedded within the main clause. But in the above sentence, the parsed information of the first two word "The" and "people" is {DT, NNS}. So this indicates that the relative clause is embedded within the main clause.

For separating the main clause and the relative clause first find the second occurrence of "VERB". The first verb occurs at the 4th position and the second verb occurs at the 7th position of the given sentence. So the first two words are concatenated with 7th to 9th words form the main clause. The delimiter of the sentence will be the delimiter to the main clause.
"Who live in Scotland" form the relative clause. Here "who" contain "subject" tag so, who is simply replaced by the word "The people" based on dependency information. Delimiter for the RC is always ".".

Input sentence: The people **who** live in Scotland are called Scots.

The dependency information for the input sentence is given below,

Typed Dependency: [det (people-2, The-1), nsubjpass (called-8, people-2), nsubj (live-4, who-3), rcmod (people-2, live-4), prep (live-4, in-5), pobj (in-5, Scotland-6), auxpass (called-8, are-7), dobj (called-8, Scots-9)]

The people *live in Scotland. (RC)*

The people are called Scots. (MC)

ஸ்காட்லேண்டில் மக்கள் வசிக்கிறார்கள்.

அந்த மக்களை ஸ்காட்ஸ் என்று அழைப்பர்.

After translating the two sentences "அந்த" should be added before the main clause. Suppose if the connective word is an "object" then the sentence is simplified as,

Input Sentence: The shoes ***which*** I bought yesterday are very comfortable.

Splitting the above sentence will result as shown below.

which I bought yesterday.

The shoes are very comfortable.

This splitting gives meaningless sentence. In our method, simplification is done in such a way that the splitted sentences are meaningful. In the above sentence, "which" indicate "The shoes". The dependency tag for "which" is "rel" (i.e. relative pronoun). So, find the words that contain {nsubj, rcmod} after "which" and store the position in **K** and place "The shoes" after **K** by ignoring "which". Then the sentence become

I bought ***the shoes*** *yesterday.*

The shoes are very comfortable.

Table 1. Sentence Simplification Algorithm

Let the input sentence be **"S"** and its length be **"n".**
For i= 0 to n
Find position of "WP, WDT, WP$" and stored in "a".
From (0 to a)
Check the presence of "verb" or "auxiliary verb"
If so flag= 1
Else
Flag= 0
End if
If (flag==1)
For i=1 to n
Check whether "who, which or whom" is "subject or object" and store in **"b"**.
If b = subject (or) relative clause and (b+1)! =subject
Then
$(S_0\ to < S_a)$ → Main Clause
Find "determinant or object tag" in main clause and store in **"w"**
$(S_{a+1}\ to < S_n)$ → **Z**
Relative Clause→ **w + Z**
End if
If b! = subject and (b+1) = to subject
Then
$(S_0\ to < S_a)$ → Main Clause
$(S_{a+1}\ to < S_n)$ →**R**
Find "verb" tag in **R** and store in **"d"**
Find "determinant or object tag" in main clause and store in **"w"**
$(S_{a+1}\ to < S_d)$ → **Y**
$(S_{d+1}\ to < S_n)$ →**Z**
Relative Clause→ **w** + Y + Z
End if
End for
End **if**
If (flag==0)
$(S_{a+1}\ to < S_n)$ → **K**
Find the second occurrence of "verb or auxiliary verb or relative clause modifier" in **K** and the position is stored in **"v"**
$(S_0\ to < S_a)$ + $(S_{a+1}\ to < S_v)$ → Relative clause
For i=1 to n
Find "who, which, whose or whom" and position is store in **"b"**.
(b-1 to 0) find the words with any of the following tags {subject, determinant, nn, subjpass, amod, poss} and the position is store in **"z"**
z + $(S_v\ to < S_n)$ → Main clause
End ***for***
End **if**
End **for**

Suppose if the relative clause is present after the main clause. First check whether the connective word is a subject or object. If the connective word is a "subject" then the sentence is simplified as,

Input Sentence: I know the policemen ***who*** chased the thief.

I know the policemen.

The policemen *chased the thief.*

Suppose if the connective word is an "object" then the sentence is simplified as,

Input Sentence: I met Sita ***who*** I called yesterday.

Splitting the above sentence with relative pronoun will result as,

I met Sita.

who I called yesterday.

Here, the second sentence "*who I called yesterday*" is meaningless. Here, "who" indicate "Sita". The dependency tag for "who" is "dobj" (i.e. direct object). So, find the second verb in the above sentence and place "Sita" after the second verb by ignoring who. Now the sentence become,

I met Sita.

I called ***Sita*** *yesterday.*

Table 2. Examples for "wh" connectives

1. Thames is a river ***which*** runs through London.
   Thames is a river.
   Thames runs through London.

2. The book ***which*** is on the table belongs to Brandon.
   The book belongs to Brandon.
   The book is on the table.

3. The box ***which*** Samuel dragged into his backyard turned out to be full of toys.
   Samuel dragged The box into his backyard.
   The box turned out to be full of toys.

4. There is nobody other than politicians aiding and abetting the students ***who*** are carrying out this strike.
   There is nobody other than politicians aiding and abetting the students.
   The students are carrying out this strike.

5. Do you know the girl ***who*** is talking to ram?
   Do you know the girl?
   The girl is talking to ram.

3.2 Sentence Segmentation

In this case, the sentences are split based on the conjunctions. Coordinating conjunction includes (for, and, not, but, or, yet, so) and POS tag for coordinating conjunction is **"CC"** and the dependency tag is "cc". Subordinating conjunction includes (when, whenever, where, wherever, if, because, unless, though, etc.). Here, the relative clause can occur before the main clause, or after the main clause.
Consider an example,

Input Sentence: Ravi/NNP, waited/VBD, for/IN, the/DT, train/NN, ***but/***CC, the/DT, train/NN, was/VBD, late/JJ.

ரவி ரயிலுக்காக காத்திருந்தான் ஆனால் ரயில் தாமதமாக வந்தது.

In the above example relative clause is present after the main clause. Here, *'but'* is the coordinating conjunction and it is the splitter word. Here the sentences will be split into two simple sentences based on the splitter word. The connective word is always present in the relative clause.

Ravi waited for the train

but the train was late.

ரவி ரயிலுக்காக காத்திருந்தான்

ஆனால் ரயில் தாமதமாக வந்தது.

Table 3. Examples for conjunction

1. The students are studying ***because*** they have a test tomorrow.
   The students are studying
   because they have a test tomorrow.

2. I want to own my own company ***and*** I want to pay all my workers a lot of money.
   I want to own my own company
   and I want to pay all my workers a lot of money.

3. Do you know a shop ***where*** I can buy used laptops?
   Do you know a shop?
   where I can buy used laptops.

4. I want to own my own company ***and*** I want to pay all my workers a lot of money.
   I want to own my own company
   and I want to pay all my workers a lot of money.

5. ***Unless*** the coffee is hot I will not drink it.
   Unless the coffee is hot.
   I will not drink it.

6. ***Though*** he enjoyed the movie he will not buy the DVD ***because*** he only watches films once.
   Though he enjoyed the movie.
   he will not buy the DVD.
   because he only watches films once.

7. Kannan is young ***so*** it would be good ***if*** someone would accompany him to Chennai.
   Kannan is young
   so it would be good
   if someone would accompany him to Chennai.

Let us consider another example,

Input Sentence: ***If*** /IN you/PRP don`t/VBP obey/RB the/DT traffic/NN rules/NNS you/PRP will /MD meet/VB with/IN accidents/NNS.

நீ சாலை விதிகளை கடைப்பிடிக்காவிட்டால் நீ ஆபத்துக்களை சந்திப்பாய்.

In this sentence the relative clause is present before the main clause. Here *'if'* is the subordinating conjunction and its dependency tag is *"mark"*. In this case second subject is considered to break the sentences. Therefore the above sentence will be split as,

If you don`t obey the traffic rules
you will meet with accidents.

நீ சாலை விதிகளை கடைப்பிடிக்காவிட்டால்

நீ ஆபத்துக்களை சந்திப்பாய்.

After translation, the above two sentences are combined to get a meaningful sentence. Our splitting technique splits the sentences containing more than one connective. The examples for sentence segmentation are given in Table 3.

4. RESULTS AND DISCUSSION

Our ultimate goal is to provide better accuracy for translation system. 200 sentences are given to the rule based English to Tamil machine translation system out of which 140 sentences are incorrect because of syntax and reordering error. Same 200 sentences were tested with simplification. 115 sentences are translated correctly after simplification.

Table 4. Results of our experiment

Rule based system	**#of Sentences for translation (Testing)**	**#of sentences with Syntactic error**	**#of sentences with Reordering error**	**#of sentences with both Syntactic + Reordering error**
Before simplification	200	48	63	29
After simplification	200	22	46	17

The long sentences are tested with Rule based machine translation system and the accuracy obtained is 30%. After The same long sentences are simplified and given to the translation system. The accuracy improves by 28% after simplification. Here, Syntax and reordering error alone is considered and not morphological error. Reordering error is to calculate the divergence in the source and target sentence such as English and Indian languages. Error in the structure of the sentence is called syntactic error. The simplified sentences are tested with Statistical machine translation and it also gives better accuracy.

5. CONCLUSION

For simple sentences, high-quality machine translation systems generate good translation, whereas for longer sentences it is difficult. In machine translation system 100% accuracy is not possible. Even though if long sentences are simplified into smaller sentences and given to the machine translation system, meaning can be retained but exact translation is not possible. The proposed algorithm deals with the sentences with more than one connective. This paper shows how the complex sentences are simplified into simple sentences and also how the sentences are translated to Tamil without changing the meaning of the sentence. Many of the Indian languages such as Telugu, Kannada, etc. are syntactically similar to Tamil. So, our sentence simplification technique will also help for the development of machine translation systems from English to other languages in India. These algorithms give better result for complex sentences. It has been proved that the splitting technique can lead to remarkable improvements in machine translation system.

6. REFERENCES

[1] Takao Doi and EiichiroSumita. 2003. "Input sentence splitting and translation", Proc. of Workshop on Building and using parallel Texts, HLT-NAACL 2003.

[2] Katsuhito Sudoh et al. 2010. "Divide and Translate: Improving Long Distance Reordering in Statistical Machine translation".

[3] Deepa Gupta. 2005. Contributions to English to Hindi Machine translation using Example-Based Approach.

[4] KatrinTomanek et al. 2003 "Sentence and Token Splitting Based on Conditional Random Fields", Jena University Language & Information Engineering (JULIE) Lab, Forman, G.

[5] Satoshi Kamatani, Tetsuro Chino and Kazuo Sumita, "Hybrid Spoken Language Translation Using Sentence Splitting Based on Syntax Structure", Corporate Research and Development Center, Toshiba Corporation.

[6] R. Chandrasekar R, B. Srinivas. 1997. Automatic induction of rules for text simplification.

[7] Furuse, O., et al. 2001. "Splitting Ill-formed Input for Robust Spoken-Language Translation", Transactions of IPSJ, vol.42 No.45.

[8] Kim, Y. B. et al. 1994. An Automatic Sentence Breakingand Subject Supplement Method for Japanese to English Machine translation.

[9] Orasan, C. 2000. A hybrid method for clause splitting in Unrestricted English texts, Proceedings of ACIDCA 2000, Monastir, Tunisia.

[10] Zhemin Zhu, Delphine Bernhard and Iryna Gurevych 2010. "A Monolingual Tree-based Translation Model for Sentence Simplification", Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010).