A Neural Network based Approach for English to Hindi Machine Translation

Pulkit Kasera
Student
Dept. of Information Technology,
Maharaja Agrasen Institute Of Technology.

ABSTRACT
In this paper we are talking about the working of our framework can decipher English language's straightforward sentences into Hindi. This framework has been executed utilizing feed-forward backpropagation fake neural system. ANN model does structure and Hindi words/tokens, (for example, action word, thing/pronoun and so on.). Neural organize is as the information base and for mapping process bilingual word reference and phonetic guidelines. Bilingual word reference is actualized utilizing neural arrange, stores the importance and semantic highlights connected to the word of English and Hindi. The changeof one common language sentence structure to other regular language is the center of the machine interpretation explicitly when the dialects have diverse linguistic class English and Hindi. Linguistic Structure examination is finished with the assistance of Stanford Tagger and Stanford Parser. The created module can decipher basic sentence of English language. The assessment score accomplished by the framework for around 500 test sentences is: n-gram blue score 0.604;METEOR score accomplished is 0.830 and F-score of 0.816. machine understanding system in figure 1. Structure contains nine modules which are sentence separator and pressure clearing module, parser and tagger module, data extraction module, language structure and sentence structure examination module, ANN and rule based sentence structure mapping module, ANN based word mapping module, sentence age module, rule based phonetic structure development and case stepping module and ANN module. Right when a customer enters a couple content for understanding, content being deciphered encounters the following process.

1. INTRODUCTION

Machine Translation is characterized as interpretation of one common language content to another common language utilizing PC (Hutchins, 1986). As per (Nirenburg and Raskin, 1987), PC must have the option to decipher input content from source language and to create yield message in target language, so that the importance of the objective language content is equivalent to that of the source language content. Machine Translation (MT) is in incredible request now-a-days because of globalization of data, data wanted to be gotten to from everywhere throughout the world. The greater part of this data is accessible in English as it were. In India, every one of the individuals can't get to this data because of language hindrance. Indian government perceives Hindi and English as official dialects of India. Hindi is spoken by around 41% Indian speakers (source: 2011 enumeration). English is spoken by in excess of 100 million Indian speakers (second in Highest in the world and in India likewise) (World Statistics, 2012). Neural systems are very productive in design coordinating and have the capacity of learning by examples.In a work of English to Urdu (Shahnawaz and Mishra, 2011a) and English to Urdu/Hindi machine interpretation (Shahnawaz and Mishra, 2011b), old style rule based approach and neural arrange have been utilized for creating machine interpretation framework.. In other work of English-Sanskrit MT, (Mishra and Mishra, 2010a) and (Mishra and Mishra, 2010b) additionally utilized utilizes neural arrange, case based thinking and from interpretation rules based approach for robotized interpretation. We have isolated this paper into six areas. Next segment presents the framework engineering and examines work process of all the modules in the framework. At that point we have examined counterfeit neural arrange model and preparing process. At that point execution of the framework is talked about. Segment five presents the outcomes got dependent on the framework yield. We have finished up this paper with our continuous work and future work plans.

2. SYSTEM ARCHITECTURE AND PROCESSING

We have displayed the square diagram of our English to Hindi machine understanding system in figure 1. Structure contains nine modules which are sentence separator and pressure clearing module, parser and tagger module, data extraction module, language structure and sentence structure examination module, ANN and rule based sentence structure mapping module, ANN based word mapping module, sentence stepping module and ANN module. Right when a customer enters a couple content for understanding, content being deciphered encounters the following process.

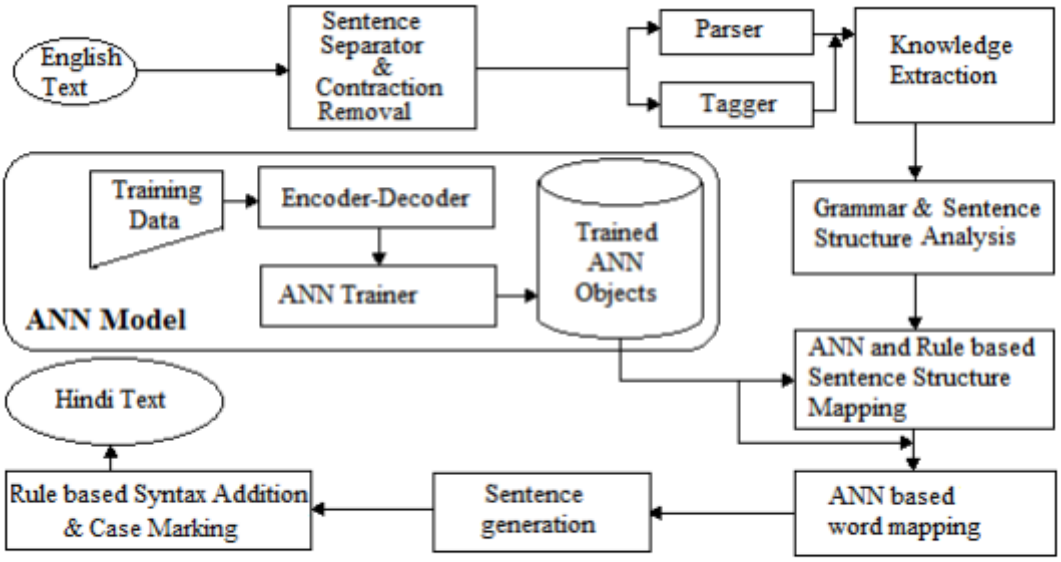

Fig 1: System Architecture

2.1 Sentence Separator and Contractions Removal Module

This module clears the text and prepares it for further processing. System scans the text being translated and finds out individual sentences from the text based upon the English language punctuations. This module also examine for the contractions used in the text. If contractions have been found in the text, this module transforms all the contraction to corresponding normal form and generate new sentence without contraction. Algorithm for this module is follows:

Read user text
IF text contains more than one sentence
    Split text into sentences
FOREACH sentence in the text
    IF there is contraction in the sentence
        Remove contraction and replace the original sentence with this new sentence.

2.2 Parser and Tagger Module

This module utilizes Stanford composed reliance parser (Stanford Parser, 2012) and most extreme entropy tagger for parsing and labeling the English language content. Stanford parser is an usage of the probabilistic characteristic language parsers, exceptionally enhanced probabilistic setting free language structure (PCFG) what's more, lexicalized reliance parsers, and a lexicalized probabilistic setting free language structure (PCFG) parser. Stanford POS tagger (Stanford Tagger, 2012) utilizes the Penn Treebank label set and is executed utilizing most extreme entropy labeling calculation. Grammatical form Tagger (POS Tagger) appoints parts of discourse tag to each word and tokens, for example, numbers, things, action words, modifiers and so on. Parsing and labeling process take place in the accompanying advances:

Load parser into memory

Load tagger into memory

FOREACH *sentence in the text*

Apply typed dependency parser

Apply POS-tagging

2.3 Knowledge Extraction Module

This module procedure parsed and labeled sentences from Parser what's more, Tagger module. Sentences are handled to separate data from the sentence and this data is at that point joined with each word. Presently each word thinks about itself like what is its situation in the sentence, which words are straightforwardly subject to it, regardless of whether it is thing, pronoun or action word and so forth. These words are changed over to items to convey this data, we call them proficient items. So every sentence in the content currently is an assortment of learned objects. Steps for changing sentences into learned objects are as per the following:

FOREACH tagged word in the tagged sentence
    Create an object and save

        Its pos-tag,

        Position in the sentence

        Word itself

Typed dependencies governing and dependent

Words which governs this word and which are dependent on it

2.4 Grammar and Sentence Structure Analysis Module

This module distinguishes lumps in the sentence for example subject, object, backhanded article, primary action word, assistant action word, prepositional articles, "ing" words and infinitive and so forth. This module additionally discovers about the dynamic or uninvolved voice of the sentence. After ordering every one of these pieces and voice of the sentence, this module moves to recognize tense of the sentence by the assistance of helper and primary action word. We likewise make sense of different properties like whether the sentence is confident or inquisitive and so on. Linguistic structure of the sentence is created by utilizing the data present in the learned articles, sentence pieces and qualities (tense, voice, type and so on.). A scrap of the calculation for sentence structure examination and sentence structure age is as per the following:

Find whether the sentence is in active or passive voice

FOREACH knowledgeable object in the sentence, process it according to voice

IF typed dependency of the word in knowledgeable object shows it as syntactic subject

Find out all its directly dependent words to form a noun phrase

IF sentence does not have dummy subject like it/there consider this noun phrase as subject

Construct subject chunk and update all knowledgeable objects which belong to the chunk
ELSE-IF subject is dummy like it/there

Construct subject chunk and update knowledgeable object

IF sentence does not have copula verb

Governor word of the typed dependency which shows syntactic subject is main verb ELSE
Copula is main verb

2.5 ANN and Rule based sentence structure mapping module

Language structure and sentence structure examination module produces the syntactic structure for each sentence in the content. This module makes rules for the produced syntactic structure of the sentence and sentence traits. Encoder encodes this rule in the arrangement which is reasonable to be utilized as contribution for prepared neural system object. Neural arrange model returns the equal objective language syntactic structure in encoded structure. Decoder interprets the acquired objective language syntactic structure of the sentence for further preparing.
This procedure pursues following advances:

Gather all the attributes and grammatical structure of the sentence

Create rule with this information for this sentence

Encode rule by Encoder
Select ANN object to be used

Send this rule to ANN model

Retrieve encoded structure from ANN model

Decode this structure.

2.6 ANN based word mapping module

Every one of the words or tokens in each sentence lump are encoded by the encoder and given as contribution to ANN bilingual word reference individually to get the objective language code comparable to this word or token. This code additionally has some heuristic data connected with it separated from importance of the word. This data might be number, individual, sex, powerless action word and action word structure and so on. Implications of the words are kept as near base which means as could be allowed with the goal that affectation should be possible concurring to the use of the words in the sentence or content. Decoder disentangles this code to acquire meaning and other data.
Word mapping process calculation is practically like the one in ANN and Rule based sentence structure mapping module.The calculation is as per the following:

Select ANN bilingual dictionary object to be used

FOREACH word in each chunk of the sentence

Encode word Send this code to ANN model

Retrieve code for encoded meaning with attached heuristic information from ANN model

Decode this code

Substitute the source language word from chunk by the decoded word and attached heuristic information

2.7 Sentence Generation module

Presently every one of the pieces of the sentence are orchestrated by the linguistic structure got in ANN and Rule based sentence structure mapping module. Source language words are as of now being supplanted by the decoded word and connected heuristic data acquire in ANN Based word mapping module. So the yield of this module is the sentence produced by the objective language sentence structure in which target language words additionally have some joined heuristic data.

FOREACH chunk of the sentence

Place at appropriate place according to the grammatical structure obtained in ANN and Rule based sentence structure mapping module.

2.8 Rule Based Syntax Addition and Case marking module

Sentence age module changes the source language sentence as indicated by the objective language punctuation acquired from ANN and Rule based sentence structure mapping module in which target language words likewise have a few appended heuristic data. This module utilizes sentence traits got from Grammar and Sentence Structure Examination module and coupled heuristic data of words gotten from bilingual ANN based word mapping module to include the linguistic structure with the objective language words and for case stamping to create significant interpretation (Shahnawaz and Mishra, 2011a). Calculation for this procedure is as per the following:

FOREACH sentence

FOREACH words in the chunk

Add rule based syntaxes which are compatible with heuristic information

Apply case marking and remove heuristic information from all the words.

Send the translated text to the output

3. ARTIFICICAL NEURAL NETWORK AND TRAINING PROCESS

ANN coach trains Feed-Forward Back-Propagation Neural System for the preparation information and makes prepared neural organize objects for bilingual lexicon and syntax rules. The main term, "feed-forward" portrays how the neural arrange forms information and reviews designs. In the feed-forward system, neurons re associated foreword as it were. Each layer of neural system contains associations with next layer (for example from the info layer to the concealed layer), however there are no back associations. A feed-forward fake neural arrange (ANN) comprises of layers of preparing units, each layer sustaining contribution to the following layer in a feed-forward way through a lot of association qualities or loads. The least complex such arrange is a two layer organize. As if there should arise an occurrence of Example Based Translation a lot of information yield design sets is given relating to a discretionary capacity changing a point in the M-dimenstional input design space to a point in the N-dimenstional yield design space, the issue of catching the inferred useful relationship is known as a mapping issue. A multilayer feed-forward neural coordinate with at least two moderate layers can play out an Example mapping task. The expression "back-proliferation" portrays the preparation procedure of this kind of neural systems. Back-proliferation is a type of managed preparing. The preparation designs are applied in some arbitrary request individually, and the loads are balanced utilizing the back-proliferation law. Every utilization of the preparation set examples is known as a cycle. The examples may must be applied for a few preparing cycles to acquire the yield blunder to an adequate low worth. Once the neural arrange is prepared, it tends to be utilized to review the fitting design for another information design. In regulated preparing technique of neural systems, the neural system must be given example inputs and the foreseen yields. Foreseen yields are analyzed against genuine yields for the given information. The backpropagation preparing calculation by utilizing the envisioned yields takes the determined blunder and changes loads of different layers in reverse from the yield layer to include layer (Yegnanarayana, 1999). The preparation designs in backpropagation calculation are applied in an arbitrary request one by one by modifying the loads utilizing the back-spread law. Every use of the preparation set designs is known as a cycle. The examples may must be applied for various preparing cycles to get the yield mistake to a worthy low worth (Yegnanarayana, 1999). The outcomes displayed by (Hagan and Menhaj, 1994) show that Levenberg-Marquardt calculation is very effective for preparing the systems having up to a couple hundred loads. Neural arrange has demonstrated very valuable in different common language preparing errands (Koncar and Guthrie, 1999). PARSEC (Jain, 1991), JANUS (Waibel and Jain, 1991) and English-Sanskrit MT framework (Mishra and Mishra, 2010a) utilize neural system approach for characteristic language preparing task and computerized machine interpretation. Following calculation works for preparing ANN linguistic structure and ANN bilingual word reference objects. Calculation for Counterfeit Neural Network Model and Training process is as pursues:

Read source and target language data from comma separated text file.

Encode data

Make the data of uniform size by adding false values (e.g. 0)

Classify data for ANN input and output

Create ANN object and train with input and output data

Save the trained ANN object

3.1 ANN Based Mapping Process

In the ANN based model, we use feed forward backpropagation Artificial Neural Network for the determination of proportionate syntactic structure and action word, thing and so forth. This procedure happens in the accompanying advances:

1) Encoding of English words/tokens or punctuation structure to numeric code.

2) Mapping of English numeric code: Data sets are encouraged to Neural Network from which ANN chooses what might be compared to the English words/tokens or language structure accommodated Translation.

3) Decoding the code of the got Hindi words/tokens or language structure. When we got the proportionate words/tokens or sentence structure, Hindi significance and data is extricated and prepared.

3.2 Encoder-Decoder

We made an informational collection of info yield sets of English-Hindi words with related information and another informational index of input-yield sets of punctuation rules. Encoder-Decoder changes over this preparation information into numeric coded structure which is *Create files to store encoded input or output data*

Read the language token to be trained as input or output

Encode each character according to code presented in the English Alphabet Encoding table

Save this code in the corresponding file.

The Decoder process is done in the following steps:

Read the retrieved code form ANN object

Decode each character according to code presented in the English Alphabet Encoding table

Return the decoded String

4. Implementation

We have utilized java as the fundamental programming language for actualizing therules and every one of the modules separated from the neural arrange model which have been executed in Matlab. Stanford parser and tagger library is likewise accessible in java.Wehave prepared, tried and effectively executed neural arrange model in Matlab. The info information for neural arrange preparing is encoded into numeric structure from literary structure by the Encoder which is likewise executed in Java. Neural arrange fills in as the information base for semantic rules and bilingual lexicon. Bilingual word reference doesn't just store the significance of English word in Hindi yet in addition stores etymological information (for example action word, thing, pronoun, number, individual and sexual orientation and so on.) appended to the Hindi words. Levenberg-Marquardt back-engendering calculation is utilized for preparing the two-layer feed-forward neural system. We have made diverse neural arrange for syntactic rules and bilingual word reference. In bilingual lexicon target language words are likewise encoded with heuristic data like action word, thing, number, individual, pronoun, and sex. The information layer of linguistic structure arrange contains 42 hubs, shrouded layer contains 100 hubs and yield layer contains 30 hubs. Mean squared mistake objective was set to preparing mistake of 10-8 which was accomplished after 29 ages. We have prepared neural system for punctuation structure rules with an information set of around 465 input-yield pair of sentence structure rules. The neural organize for educated bilingual word reference has been prepared with an informational index of around 9000 info yield pair of English-Hindi words with related information. The info layer of bilingual word reference arrange contains 10 hubs, concealed layer contains 100 hubs and yield layer contains 32 hubs (for significance and other data). Mean squared blunder objective was set to preparing mistake of 10-8 which was accomplished after 333 ages. A java class does coding and deciphering of the tokens and semantic rules and provides for the neural systems as contribution for mapping them to their identical objective language tokens and etymological rules. To robotize the procedure we have made a java class for encoding preparing information in numeric form. Encoder java class changes over preparing information into numeric structure from a book document where information is available in comprehensible form.Numeric structure is hard to peruse by a human yet simple for a program. Neural arrange then map these numeric qualities and produces equal bring about numeric structure which are of course passed to the java class which disentangles numeric yield recovered from neural arrange back to intelligible structure with the assistance ofdecoder.This information is additionally handled and target language meaning and joined data is removed. Postfix in the action word and

marker with the subject are appended on the premise of information acquired from the neural system and data got in the Grammar Analysis and Sentence Structure Recognition module. These parts are then organized as indicated by the language structure got from syntactic structure organize and the yield is introduced in Romanized structure. We have displayed here the yield delivered by our MT framework for the example English content.

Sample English Text:Sunil Kumar Singh is a student. He lives in Shimla. Shimla offers you refreshing environment. He enjoys playing hockey. He likes singing. He went to the fare with his father. He saw an old man in the fare. The old man was buying a ring for his wife from the shop. He bought a book for his sister. He met his friends. They wanted to go to watch the magician show. He decided to watch the show.

Translated Hindi Text: SUNIL KUMAR SINGH ekchātrahai | wahSHIMLA me rahtāhai |SHIMLAtumkotāzāvātāvaranpradānkaratāhai | wahhaukikhelnaānaṅdletāhai | wahgānāpasandkartāhai | wahapnepitākesāthmelākogayāthā | wahmelā me ekboodhāādamīdekhāthā | boodhāādamīmelā se apnīpatnīkeliyeekaṅgūthīkharīdrahāthā | wahapnībahankeliyeekpustakkharīdāthā | wahapnedostoṅmilāthā | vejādūgartamāshādekhnejānachāhate the | wahtamāshādekhnafaislākiyāthā |

5. RESULTS AND DISCUSSION

Different strategies have been utilized for assessing the nature of machine interpretation yield. A few highlights can be assessed naturally for instance familiarity can be checked by n-gram examination of reference interpretations are accessible and some can't as significance feeling of interpretation. We have utilized BLEU (Papineni et al., 2002) to ascertain the score of framework yield. BLEU (Bilingual Evaluation Understudy) is an IBM-created metric and utilizations altered n-gram exactness to think about the competitor interpretation against reference interpretations. It takes the geometric mean of altered exactness scores of the test corpus and afterward increases the outcome by exponential quickness punishment factor to give the BLEU score.

6. CONCLUSION AND FUTURE WORK

The interpretation results acquired from the framework assessed utilizing machine assessment techniques and physically and it has seen that the framework works proficiently on the prepared semantic rules and bilingual word reference. The MT assessment scores acquired for the framework more than 500 test sentences are: n-gram blue score 0.604; METEOR score accomplished is 0.830 and Fscore of 0.816. So an upgrade to the language structure rules and size of bilingual word reference will prompt the productive and precise machine interpretation framework. Case stamping is one of significant factor for the semantic precision of the deciphered content. In Hindi, sentence significance can change if just case markers. We have likewise seen from the consequence of framework that on the off chance that case checking is improved in the framework, framework will be capable to deliver increasingly proficient outcomes.

7. REFERENCES

[1] A.N. Jain: Parsing Complex Sentences with Structured Connectionist Networks, Neural Computation, 3, pp. 110-120, 1991.

[2] A. Waibel, A.N. Jain, A.E. MCNAIR, H. Saito, A.G. Hauptmann and J. Tebelskis: JANUS: A Speech-toSpeech Translation System using Connectionist and Symbolic Processing Strategies, Proceedings of the 1991 International Conference on caustics, Speech and Signal Processing (ICASSP-91), pp. 793-796, Toronto, Canada, 1991.

[3] B. Yegnanarayana, Artificial Neural Networks, New Delhi, India: Prentice-Hall of India, 1999.

[4] Banerjee S. and Lavie A., METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments, 2005

[5] Martin T. Hagan and Mohammad B. Menhaj, Training Feedforward Networks with the Marquardt Algorithm, IEEE Transactions on neural networks, Vol. 5, No. 6, November 1994.

[6] Mishra V., and Mishra R. B., ANN and Rule Based Model for English to Sanskrit Machine Translation. INFOCOMP Journal of Computer Science, 9, 80-89, (2010a).

[7] Mishra V., and Mishra R. B., Approach of English to Sanskrit machine translation based on case based reasoning, artificial neural networks and translation rules. Int. J. of Knowl.Eng. Soft Data Paradigm, 2, 328-348, (2010b).

[8] Mishra, Vimal and Mishra R. B., Performance Evaluation of English to Sanskrit Machine Translation System, International Journal of Computer Aided Engineering and Technology (IJCAET), InderScience Publication, UK, Vol.4, No.4, pp 340-359, 2012.

[9] NenadKoncar, Dr. Gregory Guthrie: A natural language translation neural network, In International Conference of the International Conference on New Methods in Language Porcessing (NeMLaP), pages 71 -77, Manchester, UK, 1994.

[10] Papineni K., Roukos S., Ward T., and Zhu W.-Jing, BLEU: a Method for Automatic Evaluation of Machine Translation, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, p. 311-318, July 2002.

[11] Sergei Nirenburg , Victor Raskin: The Subworld Concept Lexicon And The Lexicon Management System, Computational Linguistics, Vol (13), pages 276-289, 1987

[12] Shahnawaz and Mishra R. B., Translation Rules and ANN based model for English to Urdu Machine Translation. INFOCOMP Journal of Computer Science, 10, 36-47, 2011.

[13] Shahnawaz, Mishra R. B. ANN and Rule Based Model for English to Urdu-Hindi Machine Translation System. Proceedings of National Conference on Artificial Intelligence and agents:Theory& Application (AIAIATA 2011), 2011, pp 115-121

[14] Stanford Parser, http://nlp.stanford.edu/software/lexparser.shtml, online access 2012

[15] Stanford Tagger, http://nlp.stanford.edu/software/tagger.shtml, online access 2012

[16] Turian J. and Shen L. and Melamed I. D., Evaluation of Machine Translation and its Evaluation, In Proceedings of MT Summit IX, 2003.

[17] W. J. Hutchins: Machine translation: past, present, future. (Ellis Horwood Series in Computers and their Applications.) Chichester, Ellis Horwood, 1986. 382p. ISBN: 0-85312-788-3.

[18] World Statistics,http://www.nationmaster.com, online access 2012