International Journal of Advance and Applied Research (IJAAR)
Peer Reviewed Bi-Monthly

ISSN – 2347-7075
Impact Factor – 0.899
Vol.6 No.3 Jan – Feb 2019

DEVELOPMENT OF NMT MODEL FOR URDU – HINDI MACHINE TRANSLATION

Mohd. Khalid Mubashir Uz Zafar[2] Tuba Sabahat[3]

[1]Professor, Department of Translation, Maulana Azad National Urdu University, Hyderabad

[2]Research Scholar, Department of Translation, Maulana Azad National Urdu University, Hyderabad

Abstract

Since Urdu is one of India's official scheduled languages, the majority of machine translation work has been done by the government. Work on machine translation (MT) is crucial if Urdu is to continue to exist in the future. However, the field faces numerous obstacles at the moment, including internal ambiguity, linguistic complexity, and diversity between the source and target languages. The principles that provide linguistic information are typically the foundation of machine translation. Examples of corpus-based machine translation methods like Example Based MT (EBMT) and Statistical MT (SMT) are currently in use. In the current data-driven paradigm, these two corpus-based techniques have distinct frameworks, among other things. Potential is used by SMT systems to generate output, while EBMT systems combine a lot of training data with examples to translate input text. Urdu MT is still in its infancy, and the necessary data and computational resources are scarce.

Keywords: bilingual corpora, multilingual, alignment, parallel corpus

Introduction:

This paper examines the creation and analysis of a parallel Hindi-Urdu corpus. The alignment of sentences and words across two corpuses is one of the tasks. Therefore, the ultimate objective of alignment work is to discover word correspondence and sentences. The university has provided a bilingual Hindi-Urdu corpus for this task. The algorithm generates XML-formatted output from a text file that is used as its input. The number of words in the text is what we use because the alignment algorithm uses the idea that a text's sentence length is highly correlated by the length of its translation. The straightforward approach has been successful. Every word was correctly aligned, and the evaluation was based on parallel corpora from various fields. We close by

examining words with different interpretations. Depending on the corpus examined, the accuracy varies; In any case, the strategy can likewise be helpful for various language matches, which are dialects with phonetic similitudes.

According to Hock (1991), Hindi and Urdu are sister languages with a common ancestor. They use similar postpositions, verb morphology, and complex predicate verb structure, and their structures are very similar. Can we develop English-Urdu translation using this similarity from the machine translation (MT) system for English and Hindi? This paper tries to find a solution to this problem. "Why not consider Hindi to Urdu machine translation as a separate task instead of limiting it to the translation that is obtained from the English-Hindi MT system?" is the obvious question that comes to mind. Grammatical analysis of Hindi, including POS tagging, chunking, parsing, and transformation, if it is chosen as the source language—must be carried out. Brown et al.'s statistical machine translation system For Hindi-Urdu, a large, representative parallel corpus is required (Cohen, 1990). Such equal corpora are not accessible. An interlingua approach is yet another approach (Trujillo, 1999; Goodman, 1989; 1992 (Hutchins and Somers). The knowledge representation schema, on the other hand, is crucial to the success of this strategy. The system's capacity to extract sufficient information from the source language to produce text in the target language that is actually a translation of the source language is what determines the translation's quality and accuracy. According to Sinha (2004), a compromise is to employ a group of structurally similar languages as a pseudo-interlingua. A text generator must be developed for each target language in order to accomplish this. A simpler approach for obtaining a translation to Urdu from the Hindi English-Hindi MT system without requiring a comprehensive grammatical analysis of the source language or the development of a complete text generator from the interlingua representation was presented. For this case study, the English-Hindi MT system has been used (Sinha, 2004). Primarily, lexical mapping of Hindi words to Urdu had been used in the event that the lexicons differ in gender, and the output sentence was changed to reflect gender agreement. Megerdoomian and Parvaz (2008) use a similar English-Persian strategy to acquire Tajiki MT. In Tajiki's Cyrillic script, compounds that

are composed distinctly in the Perso-Arabic script are written together, which is one of the issues discussed there. Such compositions are uncommon for lexical mapping between Hindi and Urdu. Nevertheless, the replacements might be made for a group of nouns. The fact that the scripts used to write Hindi and Urdu do not match is an important point to make here. Hindi is written in the order Devanagari, from left to right. The Urdu script, which adds six additional characters primarily to map the sounds of English and Hindi, is based on the Perso-Arabic alphabet. Urdu is shorthand for "zaban-i-urdu," which translates to "camp language." Urdu became a language when Moghul soldiers in Indian camps interacted with the locals and began mixing words from Arabic, Farsi, and Turkish with those from the local language. Consequently, the nature and origin of the words used in Hindi and Urdu differ primarily. Urdu is spoken in a number of different ways in different parts of the world. Arabic-Farsi words are typically used more frequently by Urdu language purists. In contrast, Sanskrit is the source of Hindi. Gilchrist (1796) claims that Hindustani is frequently referred to as a language that uses words of various origins more frequently than either Hindi or Urdu. In this work, we have concentrated on the Hindustani language written in Urdu.

Development of Corpus:

A corpus is a collection of natural language spoken or written utterances, typically accessible electronically, in computational linguistics. According to their properties, corpora can be categorized in a number of different ways into various types and categories. One different ways is to recognize corpora that incorporate just a single language (monolingual corpora) and corpora that incorporate a few dialects (multilingual corpora). Parallel and non-parallel corpora are two types of multilingual corpora. Equal corpora are alluded to as regular language expressions and their interpretation with arrangement between comparing fragments in various dialects. A common source document and one or more translations of this source (target documents) are typically included in parallel corpora. Bitexts and bitext segments are terms used to describe corresponding parts of bilingual parallel corpora. In human studies, parallel corpora have been utilized. For translation research and multilingual natural language processing

(NLP) tasks, numerous applications make use of parallel corpora. For some time, bilingual concordances have been used to support human translation. For data-driven NLP tasks, parallel corpora have become a more accessible resource in recent years.

Domain Specific Corpus:

The essential justification for leading exploration in this space is the absence of earlier examination on messages written in Hindi and Urdu. The purpose of this paper is to align parallel bilingual Hindi and Urdu corpora. An xml file would contain the alignment, which would shift from the alignment of the sentences to the alignment of the words that correspond. A Hindi sentence and a translated Urdu sentence will form the pair as the sentences align. The lack of prior research on texts between Hindi and Urdu is the primary motivation for conducting research in this area. There are many things that Hindi and Urdu have in common. It is intuitive to assume that length-based or cognate-based alignment methods for Hindi-Urdu texts will yield favorable results due to this similarity. Because these languages have similar grammatical structures, we can align at the word level; However, actual research is required to either confirm or refute such hunches. Our corpus is largely built on the basis of word alignment, which allows for 1:1, 1:2, and 2:1 sentence word alignments. In another language, some words are split into two words and represented as two words rather than being the same phonetically. Words should be arranged 1:2 or 2:1 in these situations. Our primary goal is to assess how well the suggested approaches work for the languages in Hindi texts and modify them so that they work better for the parallel corpus of Hindi and Urdu. Parallel corpus analysis, in which multiple translation words are identified and an automatic bilingual dictionary of aligned words is produced, is the second objective.

Preparation of Data:

The Hindi to Urdu mapping is derived from the English to Hindi lexical database utilised in the previous MT system. Each Hindi-meaning item in the lexicon is given syntactic, semantic, and morphological details in the English-Hindi lexical database. It also contains details on the limitations on meaning selection. The Hindi-Urdu mapping table only needs to hold Urdu meanings and information

that affects the composition of Urdu texts because Hindi and Urdu employ postpositions in the same sequence. While planning with predicate action words, cautious word choice is required. A light verb follows a noun or an adjective/adverb in a predicate verb. There will be entries in the mapping table that correspond to each of these constituent words in this instance. If a Hindi and an Urdu light verb are the same, there is no need to add a new entry. Nevertheless, if a Hindi predicate verb maps onto a non-predicate Urdu verb or a predicate verb maps onto a different light verb (or vice versa), all morphological derivations must be entered. For instance, the Hindi translation of the English verb "achieve" that can be found in the lexical database is a predicate verb called "#." $%&" (!"# $ %&'#&) where !"#(%&'#&) is a noun, and the verb "$%&" is a light verb. "()* (()!*)" is the Urdu translation of the noun. and it already exists there. The ambiguous $%&"(!"# $) does not change. Therefore, there is no need to add a new entry to the mapping table. Take into consideration the Hindi form of the English verb "get" ("!&). "()* $%&"(!"#" is the equivalent translation in Urdu. $ ()!*) which has a predicate and is a verb. The mapping table must now contain all Hindi verb forms. Some of these entries are listed in Table 1. In a similar vein, the predicate verb,-./0 $%&" (!"#, in Hindi is the Hindi word for "stall" and "stop" in English $ %+,-.') and the non-predicate verb in Urdu that is equivalent to it, %1$&"(!/$01). Additionally, all predicate verb forms must be entered here. A Hindi word's number and gender are crucial to its composition. In addition, a Hindi word may have multiple parts of speech (POS). The addition of some of the postpositions may have an effect on the structure of the target text. In Urdu, this most often occurs when the gender, number, or inflection of the associated word changes. For all inflected words, the user can enter both inflected and uninflected words as options for human post-editing. In a similar vein, the postposition (!)2 encompasses all adverbs is added as an option. After being altered, humans are given one of these options. Automatic selection is currently not supported by our system. The paradigm numbers assigned to each noun, verb, and adjective in the lexical database are what are used to automatically generate the various morphological forms. The study depicts the procedure for creating the Hindi-Urdu mapping table. These entries are taken out of the

mapping table because they include a lot of common Hindi and Urdu words that are spoken in India. The production of the planning table is completely computerized; only the appropriate Urdu entry can be entered manually. The English to Hindi translation immediately benefits from the mapping table's building of the English to Urdu translation.

Set Up the Data Training Model:

Step 0: Install OpenNMT-py

pip install --upgrade pip

pip install OpenNMT-py

Step 1: Prepare the data

To get started, we propose to download a toy Hindi-Urdu dataset for machine translation containing 10k tokenized sentences:

wget <URL>toy-hiur.tar.gz

tar xf toy-hiur.tar.gz

cd toy-hiur

The data comprises both source (src) and target (tgt) data with a token at the beginning of each line and a sentence at the end of each line:

src-train.txt

tgt-train.txt

src-val.txt

tgt-val.txt

toy_hi_ur.yaml

Where the samples will be written

save_data: toy-hiur/run/example

Where the vocab(s) will be written

src_vocab: toy-hiur/run/example.vocab.src

tgt_vocab: toy-hiur/run/example.vocab.tgt

Prevent overwriting existing files in the folder

overwrite: False

Corpus opts:

data:

corpus_1:

path_src: toy-hiur/src-train.txt

path_tgt: toy-hiur/tgt-train.txt

valid:

path_src: toy-hiur/src-val.txt

path_tgt: toy-hiur/tgt-val.txt

onmt_build_vocab -config toy_hi_ur.yaml -n_sample 10000

Train the model:

the vocabulary path(s) that will be used: can be that generated by

onmt_build_vocab;

training specific parameters.

toy_en_de.yaml

Vocabulary files that were just created

src_vocab: toy-hiur/run/example.vocab.src

tgt_vocab: toy-hiur/run/example.vocab.tgt

Train on a single GPU

world_size: 1

gpu_ranks: [0]

Where to save the checkpoints

save_model: toy-hiur/run/model

save_checkpoint_steps: 500

train_steps: 1000

valid_steps: 500

onmt_train -config toy_hi_ur.yaml

On both the encoder and decoder, this configuration will execute the standard model, which comprises a 2-layer LSTM with 500 hidden units. It will use a single GPU to execute (world_size 1 & gpu_ranks [0]).

Translate

onmt_translate -model toy-hiur/run/model_step_1000.pt -src toy-hiur/src-test.txt -output toy-hiur/pred_1000.txt -verbose

validation:

1,50,000 lines with 31,50,725 words (3.1Millions)

This data is from news domain collected from internet and other sources.

Data breakup

Training: 80%

Validation: 10%

Test: 10%

Training steps: 1,00,000

We train model with 100000 training steps and test the model. below is the

Accuracy.

Validation Accuracy: 96.22 %

Test Accuracy: 92.41 %

Error Analysis:

Discussion On Results:

The average accuracy of the algorithm is 95% for sentence alignment and 75.55% for word alignment. The corpus's complexity affects how exact it is, and the more complex the corpus, the less precise it is. Complexity refers to the word distribution in the target file. The program will have trouble aligning the words if any of these categories, 1:2 and 2:1, occur simultaneously in a single sentence. Due to the frequent appearance of these categories, the corpus is more complicated. When these kinds of cases are distributed separately across the various sentences in the corpus, the outcome of this program is impressive. Whenever these techniques are used on complicated, noisy corpora, performance often falls dramatically. When texts contain components of formatting (layout), painting, and other visual elements in addition to the text itself, it becomes more challenging to adapt texts at the word level. In bilingual texts with well-arranged sentences, sentence alignment is beneficial and significantly improves alignment accuracy; As a result, texts with well-organized sentences should be written using this program.

As previously mentioned, the Hindi entries for the Hindi-Urdu mapping table must first be created. There are roughly 400000 of these entries. However, this number is roughly 100,000 because Urdu uses the same postpositions. It takes a long time to enter the appropriate Urdu meanings. Part of this step is automated. Various morphological forms from the root are generated using the

Urdu paradigm file and entered in accordance with Hindi word tags. After that, these are checked by hand. Alternative post-positions can be entered in the event of alternatives, and a human post-editor will select the appropriate post-position. The gender change module and the POS resolution module are gradually improved through experimentation. The system is currently being tested and improved. The quality of Urdu output translation is nearly identical to that of English to Hindi translation.

Conclusion:

The alignment of French words and sentences constitutes the majority of the research hansards in German, English, or Chinese for a universally reliable bilingual database. However, Hindi-Urdu texts do not contain such hansards. Consequently, we are utilizing the parallel corpus mentioned earlier. For the word and sentence arrangement, the storm's length-based strategy is utilized in the proposed calculation. The method is built on a simple statistic model as its foundation. The idea for the model came from the observation that shorter text sections typically have shorter translations, whereas longer text sections typically have longer translations. This work will have a significant impact on the development of a bilingual machine translation and dictionary system. Consequently, the project's goal is achieved. The proposed algorithm is also somewhat useful for languages that are closely related, with a few minor adjustments.

This paper offers a straightforward strategy for producing an Urdu translation from an English to Hindi machine translation. The parsing, POS tagging, and chunking that would be required for any other source language have not been applied to Hindi. Instead, the grammatical investigation of English provides all of the information needed to map from Hindi to Urdu. It is essential to keep in mind that this kind of system cannot be used to translate directly from Hindi to Urdu because a number of ambiguous mappings need to be resolved. In general, it will be necessary to solve the issues that arise when resolving translation divergence between Hindi and Urdu.

Future Scope:

There is a lot of room for improvement given that this is a new research project. In the future, we anticipate expanding the method by utilizing linguistic data. At the word level in plain text, the fundamental approaches to word alignment work. Some discriminative approaches are suggested in order to incorporate a variety of syntactic and lexical clues into the alignment models and enhance the quality of the alignment models.

References:

[1] D. Wu. Aligning a Parallel English-Chinese Corpus Statistically with Lexical Criteria. In: Proc. of the 32nd Annual Conference of the ACL: 80-87. Las Cruces, NM, 1994.

[2] D. Melamed, "A Geometric Approach to Mapping Bitext Correspondence," Proceedings of the First Conference on Empirical Methods in Natural Language Processing (EMNLP'96), Philadelphia, PA, 1996a.

[3] Dengjun Ren, Hua Wu, Haifeng Wang, ”Improving Statistical Word alignment with various clues” In proceeding of MT SUMMIT XI, Copenhagen, Denmark pages 391-397, 2007.

[4] Gale, W. A. and K. W. Church. "A program for aligning sentences in bilingual corpora," Computational Linguistics, vol. 19, pp. 75-102, 1993.

[5] Moore, R.C. "Fast and Accurate Sentence Alignment of Bilingual Corpora," AMTA 2002, pp. 135-144, 2002.

[6] Melamed, I. D. "Bitext Maps and Alignment via Pattem Recognition," Computational Linguistics, 25(1), pp.107-130, March, 1999.

[7] Melamed,I. D., "Bitext Maps and Alignment via Pattern Recognition,"Computational Linguistics, 25(1), pp.107-130, March, 1999.

[8] XU Yang1, WANG Hou-feng1, LÜ Xue-qiang2 “Research of English-Chinese Alignment at Word Granularity on Parallel “Peking University, Beijing 100871, CHINA.

[9] http://unicode.org.

[10]http://en.wikipedia.org/wiki/Natural_language_Processing