A Broad Coverage of Corpus for Understanding Translation Divergences

Simran Kaur Jolly, Rashmi Agrawal

***Abstract*: The objective of natural language understanding is to exploit the rich resources like text corpora for semantic categorization of texts. In natural language understanding corpus based statistical approaches are being used for language modeling and translation modeling. In this paper we applied the sentence pre-processing using factored base translation models on Europarl dataset and results show that pre-processing reduces the number of out of the vocabulary words accurately. This paper also defines methodology for preprocessing the parallel dataset using factored based model from Europarl dataset which can be used in machine translation ahead.**

***Index Terms*: BOW , Corpus, (Bags of Words), , Out of the Vocabulary Words (OOV), Parts of Speech Tagger (POS), Segmentation.**

.

I. INTRODUCTION

In past few Years Corpus based Statistical Machine Translation Systems have gained popularity in language modelling. Corpus based approach is a paradigm[2] that describes relation between source and target text which is a theoretical description. Corpus based approach [5] is the current state of art approach which helps in adding richer information to the machine translation models. If the source language is morphologically rich than target language, then pre-processing techniques are applied before translation to avoid data scarcity. Hence pre-processing the data is the main challenge discussed ahead with translation process.BOW (bag of words) models are used in corpus based approaches which extracts features from the corpus for pre processing of the sentences before machine translation .This technique handles the size of vocabulary by creating a group of words called as bag of words and handles the data sparseness by giving vectors '1' or '0' to words in the corpus where '1' represents presence of a word and '0' represents absence of a word or OOV (out of the vocabulary words) [7].Data driven machine translation (corpus based approaches) system require large amount of data to function accurately. The task of obtaining parallel corpus for different languages is time consuming, expensive and complex. The challenges that were seen earlier while translating from source language to target language using Corpus based approaches are:

a. Mitigating effect of rare OOV (out of the vocabulary) words.

b. Handling data scarcity.

This paper comprises of following four sections. Section 2 presents overview of existing Sentence pre-processing phases in Corpus based translation. Section 3 presents the related work of the existing Corpus based approaches in machine translation. Section 4 presents the methodology to create vocabulary for machine translation on the existing Europarl Corpus and the pre-processing of the corpus for handling the challenges mentioned above.

II. SENTENCE PREPROCESSING

Sentence pre-processing in Corpus based translation means to convert the sentence into its constituent words before they are converted to its target language. Various Machine Learning methods (word2vec, stemming algorithms and tokenization) have been used to segment sentences into its constituent tokens or words. The segmentation and tokenization helps in decreasing the size of vocabulary and solving the problem of data scarcity. Sentence segmentation further helps in word alignment or merging the sentences. The major phases of sentence segmentation are:

a) Identifying the Factors b) Sentence Alignment

The major sentence pre processing phases used in machine translation are:

A. Identifying the Factors

Identifying the factors of Machine translation is the first phase of sentence pre-processing. Factored based translation models, Bag of Words, Word2vec are the techniques used in identifying the factors during machine translation. Factored based translation Model is discussed in this section for machine translation. Factored based Models are the statistical models used for sentence pre processing before it undergoes the translation process. This model basically brings the output lemma to its base form (stemming) by first converting input to output lemma and then annotating them with parts of speech tagger and bringing them to their base form. Factored based translation models [10] are the models which combine translation model and language models of translation. In order to find the probability of translation

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Source language to target language the following factors are considered in machine translation:

$$P(e / f) = 1/ Z (\sum_{i=1} \lambda i hi(e, f)$$

Where

∑= summation of the translation component and weight of words component

e = English sentence/Source Language

f = foreign sentence/Target Language

Z= normalization constant that is ignored

$h_{i=}$ translation component: the translation component computes the probability of each feature or word in the corpus.

λ_i = weight of words: the feature component basically calculates conditional probability of input and output factors.

The factored translation model has a source language which has surface word, root word, and set of factors S which affect morphology and on target language we have surface word, root word and suffix. Suppose we have two languages A as Source language and B as Target language.

A. Surface|root|{S}(set of factors)(Source language)

B. Surface|root|suffix(target language)

The factored based translation models treat words as vectors. The above mentioned factors of factored translation model describe how the set of factors (S) affect the target language.

During translation from source language to target language following factors must be handled by factor based translation models:

a) Data scarcity: Data scarcity in machine translation means not having enough morphological forms of the source language while translating to a target language. In order to handle this factor, morphological forms are injected on target language side.

b) Correct choice of Inflection: In Inflection Morphology one word can achieve many forms so the system should make correct choice of word.

a. Sentence Alignment

Sentence Alignment is the second phase of sentence pre- processing in machine translation. Hierarchal approach for sentence alignment, Integrated Phrase alignment, Bleualign are the techniques associated with sentence alignment. Hierarchal approach to Sentence alignment in machine translation means to segment sentences to phrases and find their equivalent alignment in the opposite target language sentence. The alignment of sentences in machine translation is useful in many ways in order to meet the challenge of data scarcity and word inflection. The hierarchal approach to sentence alignment uses a one to one mapping approach by exploiting the internal structure of the sentence. The sentence alignment described in this section uses lexical information (hierarchal approach) from the available dictionaries for both the source and target language so that there is no word sense disambiguation. After extracting sentences from the corpus they are further pre- processed and converted into chunks in order to align the sentences. The rules followed for hierarchal approach to sentence alignment are:

a. There should be one to one mapping i.e. one sentence should map with one sentence in target language.

b. There should be no cross mapping as well i.e. if one sentence is aligned to another sentence then the other sentence pair cannot exist.

B. Issues in Sentence Alignment

The Issue with Sentence Alignment are extraction of phrases, generating pair of phrases. The translation probability of source language to target language in machine translation is affected by:

a. There should be an equal word matching from translation from source language to target language, both syntactically and semantically.

b. Segmentation of sentences into phrases or words must be done accurately.

c. The challenge of handling OOV (out of the vocabulary words) [11] while translating from source language to target language is there because these words are not translated to their equivalent word forms.

III. RELATED WORK

Various methodologies have been deployed for Corpus Based Translation. Three types of Corpora are defined for Statistical Research Problems by Baker et.al (1995)[12]: Comparable, Parallel and Multilingual Corpora which is the Bakers classification of Corpora. G.S Josan et.al (2007) [13] evaluated direct statistical machine translation systems having some similarities. The major inaccuracy in the system was due to ambiguous and postpositional errors in Hindi language. V Goyal et.al (2011)[8] investigated the need of root word and projected many ways of transliteration based on root lexicon. A word can be shown as Indian or foreign based on n-gram probability distribution. These results were shown on english-hindi, english-telugu datasets. The only limitation was out of the vocabulary words were not properly handled like proper names.

The corpus based approach given by Baker et al. (2004) consists of three components: monolingual, parallel, annotated corpora and the EMILLE monolingual corpora consisting of total 92,799,000 words. The facility of supporting Unicode format was the main contribution in the monolingual corpora.

Many techniques have been added to the already existing statistical models focusing on the lemmas and pre processing and post processing. Ney et.al, (2001) [15] used a more Complex framework for Machine

Translation called as phrase base translation on three datasets and observed that monotone search reduces the problem of word re-ordering. Yang et.al, (2006)[14] reviewed that languages that are morphologically rich must be processed when going through a translation process as inflectional morphology is the main challenge which must be dealt before translation modeling. Some Structural problems Colins et.al, (2005)[16] have also been seen during translation such as word reordering in the sentences.

Habash et.al, (2008)[10] proposed a tool called REMOOV for removing the out of the vocabulary words by using phrase based techniques. This pre-processing method used a novel dictionary based approach to handle the unknown words in the corpus.

Bird et.al, (2006)[4] used the natural language toolkit for preprocessing of Corpus and writing linguistic rules for the sentence conversion from one language to another. This tool basically defines the tokenization and stemming of sentences to be used further in natural language understanding.

Ondřej Bojar et.al, (2008) [6] presented HindiEnCorp that is a parallel corpus for hindi and English. The corpora are obtained from different web sources and are further pre-processed for machine translation. The parallel aligned corpus consists of smaller datasets as collected by Bojar et al. (2010) [6] including the sub corpus of EMILLE. They presented that smaller datasets give better translation quality.

Koehn et.al, (2005) [17] described a method for adding syntactic information to existing statistical machine translation systems. The first step is parsing of the source language and then applying reordering techniques for proper alignment of the sentences. The second step is to generate the target language word order that is close to target sentence.

Koehn at.al (2008)[18] further investigated how source language translation improved by adding more syntactical information to the corpus. hence inflections of target words were injected into the existing translation systems. Chahuneau et. al, (2013)[3] showed how the factored based translation model can be further improved by adding phrases.

IV. CORPUS PREPROCESSING ON EUROPARL DATASET

Parallel Corpora [8] are documents read by computer having two parallel languages which have sentences aligned parallel to each other. In order to deploy corpus based architecture various NLP tools like Stanford POS tagger, Syntactic Parsers, Corpora, bilingual dictionaries, lexicon has been adopted [4]. This paper describes the corpus pre-processing approach for building and analyzing the data using factored based translation models and sentence alignment for machine translation. Further we can use python libraries like NLTK [4] to process data and add rich information to the existing Corpus.

Corpus Pre-processing in machine translation is basically used to clean the corpus by removing noisy data. Noisy data is removed by applying tokenization, stemming and removing out of the vocabulary words. In order to show corpus pre-processing we acquire pair of English-Italian dataset and apply pre-processing steps to the corpus.

Word	Words forms	I bought fruits from market.
Root word	Root word	I buy fruit from market
Tagging	Whether it is a noun, verb or adjective.	Subject, Root, Noun object, prepostion, Proper noun object.

Table 1: Factorization of English Sentences

Corpus pre-processing in Machine translation using factored based models does factorization of sentences into its constituent tokens and assigning tags to the word. Table1 above shows how an English sentence is factored using the factor based translation model on the basis of factors suffix, root word and Inflection.

The phases of Corpus Pre-processing are:

A. Preprocessing Pipeline

The processing pipeline begins with loading the source documents and extracting plain text from the corpora. The initial stages of pre-processing include identifying the language and preprocessing them. The experiments in the paper for preprocessing of the corpus for machine translation are performed on a parallel corpus called Europarl ([14]). This corpus is extracted from the proceedings of the European Parliament from April 1996 to December 2001.it consists of sentences spoken by politicians in parliament. The Corpus consists of 20 million words in 743,880 sentences of each language. The corpus is available in an alignment framework for aligning sentences and preprocessing them.The factored corpus consists of a noun identifier and sentence aligner in order to find the out of the vocabulary words in the above mentioned approach. Hence, for acquiring the corpus we take an instance of Europarl Corpus for pre-processing of corpus before the translation modeling.

The steps followed in Corpus pre-processing are:

1. Language Identification: The first

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

step of corpus pre-processing is identifying the language and converting it into a binary format.

2. Loading the Corpus: The corpus is loaded into the memory and then it is converted to binary file format i.e. the pkl(pickle cleaned files) files and trim the corpus by removing the out of the vocabulary words accurately.

3. Corpus Generation: After the corpus is loaded into the translation memory then there is need to clean the sentences by tokenizing them on the basis of white spaces and normalizing to lowercase letters and converting them to some binary file formats.

4. Vocabulary Generation: After the corpus is pre-processed to its binary file format , the out of the vocabulary words are removed by loading the pickled clean files into the memory

B. Corpus Statistics

The Europarl Corpus is available online easily for non-commercial use. The table below explains the pre-processing of the English-Italian Corpora and how it can be further preprocessed and number of words generated after pre-processing. The table 2 shows the words before and after pre-processing the corpora using factored translation models. The pre-processed corpora can be further used in machine translation for conversion of source language to target language accurately.

Table 2: Vocabulary Generation

English Vocab before preprocessing	New Vocab	Italian Vocab (before)	New Italian vocab
104655(words)	41564(words)	171079(words)	67378(words)

C. Sentence Alignment Pipeline

The hierarchal approach for sentence alignment follows a lexical based approach as well as the sentence length of source language and target language in the parallel corpus. Lexical information is important when two sentences are of equal lengths. Statistical information that means the length of sentence is important when there is 1:1 alignment of the sentences. So in order to have full alignment at sentence level a combination of lexical and statistical based approach is followed. The major steps followed in this are:

1. First we define the alignments between the pair of sentences and fix it to 1:1.
2. Obtain the difference between the length of the sentences in source and target language i.e. the minimum length and maximum length called as distance measure.
3. If the distance measure between source and target language is less, then both the source and target language are aligned to each other correctly by 1:1.

For instance a sentence in English and its equivalent translated sentence in French are shown in table below. The word pairing is done for sentence alignment and translation of source language (English) to target language (French). The table below shows the sentence pair aligned and not aligned to each other. The sentence alignment is shown by two ways:

1. Pair1- 1:1 alignment: This means that there are equal number of words in both source and target language.

^{1}Resumption ^{2}of ^{3}the ^{4}session

Lexical translation

Reprise de la session

Statistical translation

^{1}Reprise ^{2}de ^{3}la ^{4}session

Aligned in 1:1

2. Pair2-Not aligned: This means that there are unequal number of words in source and target language resulting in improper alignment.

I[1] declare[2] resumed[3] the[4] session[5] of[6] the[7] European[8] Parliament[9] adjourned[10] on[11] Friday[12] 17[13] December[14] 1999[15], and[16] I[17] would1[8] like1[9] once[20] again[21] to[22] wish[23] you[24] a[25] happy[26] new[27] year[28] in[29] the[30] hope[31] that[32] you[33] enjoyed[34] a[35] pleasant [36] festive[37] period[38].

Lexical Translation

Je déclare reprise la session du Parlement européen qui avait été interrompue le vendredi 17 décembre dernier et je vous renouvelle tous mes vux en espérant que vous avez passé de bonnes vacances.

Statistical Translation (alignment)

Je[1] déclare[2] reprise[3] la[4] session[5] du[6] Parlement[7] européen[8] qui[9] avait[10] été[11] interrompue[12] le[13] vendredi[14] 17[15] décembre[16] dernier[17] et[18] je[19] vous[20] renouvelle [21] tous[22] mes[23] vux2[4] en[25] espérant[26] que[27] vous[28] avez[29] passé[30] de[31] bonnes[32] vacance

Table 3: Sentence Alignment

Sentence one(English)	Sentence two(French)	Matching words	Alignment

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Resumption of the session	Reprise	Alignment([(0, 0), (1, 1), (2, 2), (3, 3)])	1:1
I resumed the session of the Europ ean Parliament adjourned Friday December 1999, and I would like once again to wish you a happy new year in the hope that you enjoyed a pleasant festive period.	Je décembre dernier et je vous renouvelle mes vux en espérant que vous avez passé de bonnes vacances .	Alignment([(0, 0), (1, 1), (2, 2), (3, 3), (4, 4), (5, 5), (6, 5), (7, 7), (8, 6), (9, 8), (9, 9), (9, 10), (9, 11), (10, 12), (11, 13), (12, 14), (13, 15), (14, 16), (16, 17), (17, 18), (18, 19), (19, 20), (20, 20), (21, 20), (29, 24), (31, 25), (32, 26), (33, 27), (33, 28), (34, 29), (35, 30), (36, 20), (36, 31), (37, 20), (37, 21), (37, 22), (37, 23), (37, 32), (38, 32), (39, 33)])	Not aligned

V. DISCUSSION

The idea of sentence pre-processing using factored based translation model and sentence alignment has been shown on European language pair i.e. English-Italian and English-French. The results are shown in Table 2 and Table 3. Table 2 shows Corpus pre-processing reduces number of out of the vocabulary words after applying factor based models on the existing corpus. Table 3 shows that when the sentence is pre-processed in corpus it has to be aligned to the target language sentence word by word. The above preprocessing steps are applied on the available datasets to align the pair of preprocessed sentences from corpus. After the sentences are aligned they can be translated accurately from source language to target language accurately.

VI. CONCLUSION

The field of Language Modeling and Corpus modeling requires a dedicated corpus which consists of source and target language text from many languages for creation of a Vocabulary. In this paper, the acquisition of a personalized corpus is done which primarily addressees study on properties of the translation divergence. The corpus that we use is the Europarl Corpus that has 11 different languages and sentence aligned parallel to English language. In future, there is no option rather than adding more texts from other languages and Asian languages so it is a big resource for language modeling .Tagging or Annotating the data is further improvement in the Corpus based approach for supporting more translation features.

REFERENCES

1. Al-Rfou, R., Perozzi, B., and Skiena, S. (2013). Polyglot: Distributed word representations for multilingual NLP. CoNLL-2013, page 183.
2. Sara Laviosa. 1998. The corpus-based approach: A new paradigm in translation studies. Journal des traducteurs / Meta: Translators' Journal, 43(4):474–479.
3. Chahuneau, Victor, Eva Schlinger, Noah A. Smith, and Chris Dyer. 2013. Translating into Morphologically Rich Languages with Synthetic Phrases. Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics.
4. Bird, S. (2006). NLTK: the natural language toolkit. In Proceedings of the COLING/ACL on Interactive presentation sessions, pages 69–72.Association for Computational Linguistics.
5. HindEnCorp - Hindi-English and Hindi-only Corpus for Machine Translation. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14).
6. Bojar, Ondřej, Straňák, Pavel, and Zeman, Daniel. (2008). English-Hindi Translation in 21 Days. In Proceedings of the 6th International Conference On Natural Language Processing (ICON-2008) NLP Tools Contest, Pune, India, December. NLP Association of India.
7. Daumé III, H. and Jagarlamudi, J. (2011). Domain adaptation for machine translation by mining unseen words. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2, pages 407–412.Association for Computational Linguistics.
8. Vishal Gupta and Gurpreet Singh Lehal, "Automatic Keywords Extraction for Punjabi Language", International Journal of Computer Science Issues, Vol. 8, No. 5, pp. 327-331 (2011).
9. V. Goyal and G. S. Lehal, "N-Grams Based Word Sense Disambiguation: A Case Study of Hindi to Punjabi Machine Translation System", International Journal of Translation, Vol. 23, No. 1, pp. 99-113 (2011).
10. Habash, N. (2008). Four techniques for online handling of out-of-vocabulary words in Arabic-English statistical machine translation. In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers, pages 57–60. Association for Computational Linguistics.
11. Habash, N. and Metsky, H. (2008). Automatic learning of morphological variations for handling out-of-vocabulary terms in Urdu-English machine translation. Proceedings of the Association for Machine Translation in the Americas (AMTA-08), Waikiki, HI.
12. Mona Baker. 1995. Corpora in translation studies. An overview and suggestion for future research. Target, 7(2):223–243.
13. G S Joshan and G S Lehal, "Evaluation of Direct Machine Translation System from Punjabi to Hindi", International Journal of Systemics, Cybernetics and Informatics, pp. 76-83 (Jan 2007).
14. Yang, M. and Kirchhoff, K. (2006). Phrase-based backoff models for machine translation of highly inflected languages. In Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL).
15. Nießen, S. and Ney, H. (2001). Toward hierarchical models for statistical machine translation of inflected languages. In Workshop on Data-Driven Machine Translation at 39th Annual Meeting of the Association of Computational Linguistics (ACL), pages 47–54.
16. Collins, M., Koehn, P., and Kucerova, I. (2005). Clause restructuring for statistical machine translation. In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05), pages 531–540, Ann Arbor, Michigan. Association for Computational Linguistics
17. Koehn, P. (2005). Shared task: Statistical machine translation for european languages. In ACL Workshop on Parallel Texts

AUTHORS PROFILE

Simran Kaur Jolly: Pursuing Ph.d in Computer Applications, doing Research on Natural Language Processing. Worked on research such as designing new learning algorithms, challenges in machine perception, data mining, machine learning, and natural language understanding, Taught courses such as database management systems, fundamentals of C,advanced database management systems (ADBMS), data mining and data warehousing, machine learning & artificial intelligence, contributed to research communities, including publishing papers in various international machine learning journals (e.g.: IGI Global, IJMLNCE, JETIR, ICICC, IJETTCS, IJCS)

Dr. Rashmi Agrawal: is working as Professor in Department of Computer Applications in MRIIRS, Faridabad. Dr. Agrawal has a rich teaching experience of more than 17 years. She is UGC-NET(CS) qualified. She has completed PhD, M.Phil, MTech , MSc and MBA(IT). She has completed her PhD in the area of Machine Learning. Her area of expertise includes Artificial Intelligence, Machine Learning, Data Mining and Operating System. She has published more than 30 research papers in various national and International conferences and Journal and authored many books and chapters in edited books. She has organized various Faculty Development Programmes and participated in workshops and Faculty development Programmes. She is actively involved in research activities. She is a life time member of Computer Society of India. She has been a member of the Technical Programme Committee in various conferences of repute.