Enhancing Aspect Extraction in Hindi

Arghya Bhattacharya, Alok Debnath and **Manish Shrivastava**
Language Technologies Research Center (LTRC),
Kohli Center for Information Systems (KCIS),
International Institute of Information Technology, Hyderabad (IIIT-H)
{arghya.b, alok.debnath}@research.iiit.ac.in
m.shrivastava@iiit.ac.in

Abstract

Aspect extraction is not a well-explored topic in Hindi, with only one corpus having been developed for the task. In this paper, we discuss the merits of the existing corpus in terms of quality, size, sparsity, and performance in aspect extraction tasks using established models. To provide a better baseline corpus for aspect extraction, we translate the SemEval 2014 aspect-based sentiment analysis dataset and annotate the aspects in that data. We provide rigorous guidelines and a replicable methodology for this task. We quantitatively evaluate the translations and annotations using inter-annotator agreement scores. We also evaluate our dataset using state-of-the-art neural aspect extraction models in both monolingual and multilingual settings and show that the models perform far better on our corpus than on the existing Hindi dataset. With this, we establish our corpus as the gold-standard aspect extraction dataset in Hindi.

1 Introduction

Recent literature has seen an increase in the amount of work being done in fine-graining downstream NLP tasks. One common method of fine-grained analysis is the use of aspect information. An aspect term is an entity of interest which identifies a unique aspect of a predefined topic or domain (Pontiki et al., 2014). For example, in the *restaurant* domain, *service* and *seasoning* are aspects. While aspect extraction (AE) has been often seen as a subtask of fine grained aspect-based sentiment analysis (ABSA), recent advances in literature have established it as an independent task which can be used in other downstream tasks as well, such as summarization (Fremann and Klementiev, 2019) and topic-specific information retrieval such as opinion mining (Asghar et al., 2019).

Aspect extraction (as a subtask of aspect-based sentiment analysis) datasets and models have been developed for multiple languages. ABSA has been a shared task in SemEval 2014 (Pontiki et al., 2014), 2015 (Nakov et al., 2015), 2016 (Pontiki et al., 2016), and as a part of the overall task of sentiment analysis on Twitter in SemEval 2017 (Rosenthal et al., 2017). These tasks have garnered a lot of attention in various languages including Arabic, Chinese, Dutch, French, Russian, Spanish and Turkish. Each monolingual dataset consisted of one or two domains with each language having anywhere between 4,000 to 9,000 sentences overall (including the train and test split). For Indian languages, there has been some work in developing a dataset for aspect extraction in Hindi (Akhtar et al., 2016) and Telugu (Regatte et al., 2020).

Limited work has been done on improving the state of AE and ABSA in Hindi beyond the development of a singular dataset, namely Akhtar et al. (2016). Existing evaluations show that existing sequence tagging models (both general and specific to AE) have performed very poorly on this dataset when their performance is compared to English AE as well as in similar sequence tagging tasks in Hindi such as named entity recognition (NER) and event detection.

In this paper, we thoroughly analyze the existing dataset for AE in Hindi and explain the reason for the poor model performance. We then propose the creation of a parallel corpus, by manually translating the SemEval-2014 ABSA corpus (Pontiki et al., 2014). We provide detailed guidelines and challenges faced during the creation of this resource. We show that our dataset performs much better than the existing dataset for Hindi using baseline as well as state-of-the-art neural models for AE. Finally, we leverage the SemEval-2014 corpus to perform zero-shot and fine-tuned aspect extraction in Hindi using multilingual BERT with baseline and SoTA neural models in the dataset we have created.

Therefore, the main contributions of this paper

Proceedings of the 4th Workshop on e-Commerce and NLP (ECNLP 4), pages 140–149
August 5, 2021. ©2021 Association for Computational Linguistics

are:

- providing an in-depth qualitative and quantitative analysis of the existing Hindi AE dataset,
- creating a new resource for aspect extraction in Hindi by translating the SemEval 2014 corpus into Hindi[1],
- providing detailed guidelines and challenges associated with the creation of this corpus, as well as explaining the quality of the translations and annotations, and
- evaluating the new dataset using state-of-the-art neural sequence labeling models for aspect extraction in Hindi in monolingual and multilingual settings using transfer learning.

We establish that our corpus is a more robust and representative corpus for aspect extraction in Hindi, and its parallel nature can be exploited for a large number of downstream tasks including review translation, cross-lingual opinion mining, and aspect-based sentiment analysis.

2 Dataset Development

As discussed in Section 1, Akhtar et al. (2016) is the only corpus for aspect term extraction and aspect-based sentiment analysis in Hindi. In this section, we discuss the inadequacy of this corpus by analyzing the data qualitatively, statistically and through experiments using established aspect extraction models. We also detail the process of creating a parallel aspect extraction datset from the English gold standard dataset (Pontiki et al., 2014). This resource can be treated as an individual Hindi aspect extraction dataset or can be considered a parallel resource for the aspect extraction task.

The annotation format used for the existing dataset and the dataset being created is the Begin-Inside-Outside or `BIO` sequence labeling format (Ramshaw and Marcus, 1999). This format annotates each word with a corresponding label where a word labeled `B` denotes the first word of an aspect, `I` denotes any word within the aspect span and `O` denotes the words outside the aspect span.

2.1 Analyzing Existing Datasets

In this section, we aim to prove based on a qualitative and statistical analysis of the Hindi ABSA dataset for AE, and compare it to the SemEval-2014 English ABSA dataset. Some of the metrics for comparison include the number of sentences, number of aspects, ratios of `B`s, `I`s, and `O`s and the number of marked sentences (sentences with one or more aspects). We explain how these comparisons explain the quality of the dataset for this task as well. We also show a quantitative performance analysis of these datasets on baseline model as well as the state-of-the-art models in sequence tagging and aspect term extraction in Section 3.1.

The Akhtar et al. (2016) dataset consists of $5,147$ sentences, and a total of $4,509$ aspect terms. The combined SemEval-2016 dataset shows a similar trend, with $6,092$ sentences and $6,072$ aspect terms. However, on a closer analysis of the dataset, detailed in Table 1, we see three prominent distinguishing factors:

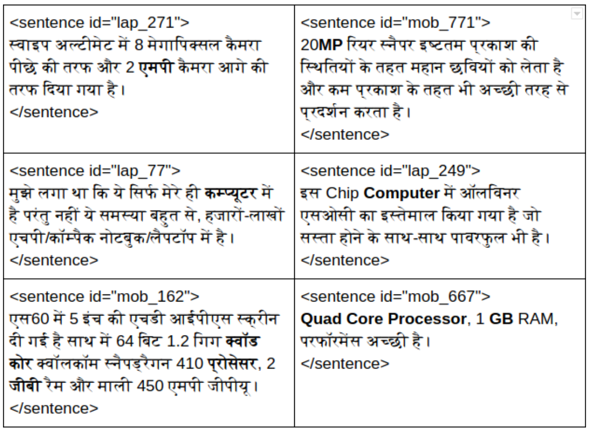
1. While the percentage of marked sentences (sentences with one or more aspects) is higher in the Hindi dataset than in the English one, there is a noticeable difference between the average number of aspects per sentence (both for marked sentences and overall).

2. The percentage of `I`s in the Hindi dataset (3.26%, $3,135$ out of $96,140$ words) are higher than the English dataset (2.96%, $2,564$ in $86,552$ words), while the number of `B`s in the Hindi dataset are lower. This implies that in multi-word aspects are far more common in Hindi than they are in English. Further, the percentage of `O`s is higher in the Hindi dataset as well, so there are not as many words which are aspect terms either.

3. The data in Hindi corpus is from *12* different domains, with some domains having less than 50 sentences. So, not only is there a large variety in topics and aspects per topic, there is also a high disparity in the number of samples per topic. In contrast, the English dataset is derived from only two different domains, with over 2000 sentences per domain.

This disparity in the number of aspects per topic as well as the noticeable difference in the number of multi-word aspect terms implies that corpus developed by Akhtar et al. (2016) is sparse with very few examples of the syntactic features, aspects and their categories.

[1] https://drive.google.com/file/d/1wrCHi3VbQjosvhmpfS577TXZL-O6UbUo/view?usp=sharing



<sentence id="lap_271"> स्वाइप अल्टीमेट में 8 मेगापिक्सल कैमरा पीछे की तरफ और 2 **एमपी** कैमरा आगे की तरफ दिया गया है। </sentence>	<sentence id="mob_771"> 20**MP** रियर स्नैपर इष्टतम प्रकाश की स्थितियों के तहत महान छवियों को लेता है और कम प्रकाश के तहत भी अच्छी तरह से प्रदर्शन करता है। </sentence>
<sentence id="lap_77"> मुझे लगा था कि ये सिर्फ मेरे ही **कम्प्यूटर** में है परंतु नहीं ये समस्या बहुत से, हजारों-लाखों एचपी/कॉम्पैक नोटबुक/लैपटॉप में है। </sentence>	<sentence id="lap_249"> इस Chip **Computer** में ऑलविनर एसओसी का इस्तेमाल किया गया है जो सस्ता होने के साथ-साथ पावरफुल भी है। </sentence>
<sentence id="mob_162"> एस60 में 5 इंच की एचडी आईपीएस स्क्रीन दी गई है साथ में 64 बिट 1.2 गिग **क्वॉड कोर** क्वॉलकॉम स्नैपड्रैगन 410 **प्रोसेसर**, 2 **जीबी** रैम और माली 450 एमपी जीपीयू। </sentence>	<sentence id="mob_667"> **Quad Core Processor**, 1 **GB** RAM, परफॉरमेंस अच्छी है। </sentence>

Figure 1: Some examples of inconsistent samples in the Hindi dataset. The words in bold face are the same in both examples, transliterated into Devnagari on the left and left Romanized on the right in different training samples.

Further qualitative analysis of the data reveals discrepancies in the data creation methodology, particularly surrounding technical terms which do not have commonly used translations. Terms such as 'computer', 'megapixel', 'quad core' and 'processor' have been transliterated in some examples and have been left Romanized in others. Given the low number of examples per category, this inconsistency contributes to the data sparsity. Figure 1 shows a few examples of such inconsistencies.

Finally, we see examples of incorrect annotation which also contributes to the dataset quality in terms of performance in machine learning models. These incorrect annotations include incorrect spacing between words in the original review text, incomplete aspect annotation where the last character of the last word of the aspect was not a part of the aspect span, and subword level aspects due to stemming, lemmatization and dehyphenation.

There are two available task performance measures for the term of aspect extraction in the Hindi dataset:

- Akhtar et al. (2016) analyzed aspect term extraction using the `BIO` annotation using conditional random fields (CRFs) for sequence labelling. They report an average F1 score of just 41.07%. The CRFs used were heavily feature engineered to use features such as semantic orientation, local context tagging and bigram specific features.

- Akhtar et al. (2020) performed joint modeling and end-to-end aspect extraction on both the Hindi as well as the Pontiki et al. (2014) English dataset. They reported a maximum F1 score of 83.36% for the English dataset using an end-to-end architecture, while the maximum F1 score for Hindi using the same architecture was 52.03%. Other experiments also show this vast disparity.

These discrepancies show that even heavily feature-engineered statistical models as well as neural models do not perform well on the existing Hindi dataset and the neural models seem to perform a lot better on the SemEval 2014 dataset. An aspect term extraction task comparison for various models can be found in table 3 for a number of models described in section 3.1.

Table 3 shows that the discrepancies noted by Akhtar et al. (2020) continue to hold across multiple neural models. The difference between the F1 scores between the two datasets is nearly 40% for all three models, with the maximum F1 score in the Hindi dataset being a mere 38.21% for the DeCNN model. We conclude through this thorough analysis that the Akhtar et al. (2016) dataset is inadequate as a benchmark dataset for aspect extraction in Hindi.

2.2 Constructing the Parallel Corpus

We construct a parallel corpus by translating the SemEval 2014 English aspect based sentiment analysis dataset of restaurant and laptop reviews (Pontiki et al., 2014). The dataset constructed by this translation can be used as an independent Hindi dataset, or can be used such that it leverages the English dataset for aspect extraction. By using the guidelines provided below, we are able to preserve the diversity of syntactic constructions from the original dataset, making the quantitative comparisons more representative.

The final dataset constructed by this methodology consists of 5989 sentences with 5864 aspects. Not all the sentences could be translated based on our guidelines which aim at maintaining naturalness and fluency. The guidelines pertaining to the translation and aspect extraction have been discussed below, followed by the methodology of annotation. The comparative statistics of this dataset can be found in 1, when compared to Akhtar et al. (2016) and Pontiki et al. (2014).

Annotation Guidelines The guidelines for creating this parallel corpus were twofold, translating the dataset into Hindi and identifying the aspect terms in the translations.

The translation methodology adopted for this task had to account for fluency, accuracy and style. Not only did the translated reviews had to be as semantically similar to the original review as possible, but they also had to be faithful to the style of restaurant and technology reviews in Hindi. In order to achieve a natural translation true to this style, we propose the following translation guidelines.

1. For proper nouns and other names in English, such as locations, company names and other named entities, annotators were asked to directly use Roman script. For example: *Brooklyn, 2nd Street, Sony*. We found that proper nouns in both domains indicated a property of the main topic and rarely that of an aspect, so using Roman script could aid in attribute extraction without being a problem in aspect extraction or other downstream tasks.

2. For common nouns without Hindi translations, or with very obscure translations which are not commonly used, annotators were asked to transliterate these nouns into Hindi. This was done in order to maintain consistency in the use of technical terms which could act as aspects in the Hindi sentence, while maintaining the domain-specific naturalness and fluency of the translated sentence. Word such as *keyboard, bluetooth, monitor, sake* and *soy sauce* were transliterated into Hindi.

3. Aspect descriptions often contain idiomatic constructions or other compositional phrases. Translators were asked to simplify these phrases to their meaning rather than translate word for word. Therefore, for phrases such as '*on the nose*' was translated to *yathaarth* (meaning 'obvious') rather than *naak ke upar* (literally meaning 'on or over the nose')

4. For common nouns with gender and number inflections, annotators were asked to transliterate the root word (as mentioned in rule (2)) but use the Hindi inflection markers. As English pronouns and nouns are not gender marked, the default male inflection is used whenever applicable.

5. For all other words, aspects and aspect descriptions, translate into Hindi using the most commonly used words given the appropriate context. In the case where the context is so scarce that there is no way to translate the sentence in a way that preserves meaning, do not translate the sentence.

After the translation, a different group of annotators were asked to identify aspect terms. Aspect term identification guidelines were the same as those used in the SemEval-2014 ABSA task[2] (Pontiki et al., 2014). The annotators were asked to annotate all single or multiword terms which were a particular aspect of the target entity (i.e. 'Restaurant' or 'Laptop').

Annotation Methodology Each sentence in the Pontiki et al. (2014) dataset was translated by four translators, two undergraduate and two graduate students. All translators are bilingual speakers of Hindi and English and are between the ages of 18 and 22. The translated sentences were then provided to two other annotators for the aspect extraction task. These annotators were in the same age group and of the same composition in terms of expertise in Hindi and English.

Translation was performed in two phases: *aspect-aware* and *aspect-blind* translations. In aspect-aware translation, the translator were provided the aspect terms while translating the sentence and were to retain as many aspects in the translated sentence as possible while maintaining the rules of translation mentioned above. In the aspect-blind translation, the translators were provided just the sentence to translate with no additional instructions. This two-phase translation was done to determine the fluency and naturalness of the translations with respect to one another with and without the constraint of maintaining aspects. The dataset contains the most fluent version of the annotations and those which maintain the most aspects from the source sentences in the SemEval dataset.

These translated sentences were provided to the final annotators, who were asked to identify the aspects in these sentences based on the guidelines provided above. This was compared to a direct translation of the extracted aspects in the source sentence (which were provided in the dataset).

Challenges in Annotation Some of the main challenges in translating the data are detailed below.

[2] http://alt.qcri.org/semeval2014/task4/data/uploads/semeval14_absa_annotationguidelines.pdf

Metrics	Akhtar et al. (2016)	Pontiki et al. (2014)	Our Dataset
Total # Sentences	5417	**6092**	5989
Total # Aspects	4509	**6072**	5864
Total # Tokens	96140	86552	**104618**
% Sentences marked with Aspects	**61.5%**	57.7%	57.9%
Avg. # of aspects per sentence	0.81	**0.99**	0.98
Avg. # of aspects per marked	1.32	**1.72**	1.69
% of Bs	4.69%	**7.01%**	5.60%
% of Is	**3.26%**	2.96%	2.80%
% of Os	**92.15%**	90.22%	91.60%
No. of Domains	**12**	2	2

Table 1: Some basic comparative statistics between the aspect extraction and aspect based sentiment analysis datasets. We see that while the Hindi dataset has lower number of samples, fewer aspects, lower ratio of aspects per sentence and lower number of sentences with aspects. Interestingly, however, these words have been added to a much larger number of domains in Hindi and there are higher number of words with the I and O tags.

1. The most common problem in translation was semantically compositional constructions such as idiomatic phrases. Phrases such as "*boy oh boy*", "*don't look down your nose*" etc. were descriptive of a given aspect in the corpus, but could not be easily translated due to a lack of natural corollaries for these phrases in Hindi.

2. Constructions with puns and aspects embedded in the compositional constructions were the biggest challenge to the translation. For example: '*But that wasn't the icing on the cake: a tiramisu that resembled nothing I have ever had*' had the aspect '*icing on the cake*' which is both literal and metaphorical in this sentence. In the final version of the data, these sentences have not been included due to very high disparity between the translations and the difficulty in extracting aspects.

3. Elided references were a concern for translators. For example, a sentence such as '*A cheap eat for NYC, but not for dosa.*' uses the term '*eat*' to refer to a '*place to eat*' which is also an aspect in this sentence. A direct translation forces this elision to be explicit, which also changes the aspect term.

4. Hindi syntax has relatively free word order, which affords fragmentation of noun and verb phrases by adjectives and adverbs respectively. The aspect-aware and aspect-blind translations often differed in such cases, as the aspect-aware translation is not fragmented, but is also generally unnatural according to the annotators. For example, the phrase "*everything bagel with lox spread*" has the annotated aspect *bagel with lox spread*, but gets translated to *lauks spred ke saath evarithing begal* (where the word "everything" fragments the aspect term).

5. Certain aspect terms translate only based on context, which is not always provided in the data. An example of this is ... *mine was well done and dry* without a subject in reference, where the term *well done* can have different translations in different contexts (such as a well-done steak versus an actual compliment).

Due to these challenges in dataset annotation and the lack of context to make an informed translation which was natural and fluent, some sentences and aspects could not be translated into Hindi. Therefore the Hindi dataset has a few fewer sentences that the English dataset. The final translated dataset consists of 5,989 sentences with 5,864 aspect terms.

2.3 Dataset Analysis

In this section, we show some basic statistical analyses of the dataset including the annotator performance in translation and aspect term extraction. For translation performance, we compare the ROUGE-L scores across the translators, while for the annotation task, we use the Fleiss' Kappa metric.

We evaluate these translation based on the ROUGE-L metrics as the average review length is no more than 15 words, and most words have only one (or very few) variations in translation. Given

Comparison	ROUGE-L	Fleiss' Kappa
Aspect-aware	0.8994	0.8961
Aspect-blind	0.8722	0.9244
Overall	0.8960	0.9130

Table 2: The average ROUGE-L and Fleiss' Kappa score in the translation and annotation tasks respectively.

the stringent translation/transliteration guidelines, lack of extensive vocabulary in the descriptions and less number of words per sentence, the ROUGE-L metric is a decent approximation of the translation quality provided by the annotators. The ROUGE-L metric also accounts for the relative free-word order nature and constituent rearrangement (Lin and Och, 2004).

ROUGE-L is the comparison of the longest common subsequence between two translated phrases. Given the translations X of length m and Y of length n, the ROUGE-L score is given by:

$$R_{lcs} = \frac{\text{LCS}(X,Y)}{m} \tag{1}$$

$$P_{lcs} = \frac{\text{LCS}(X,Y)}{n} \tag{2}$$

$$F_{lcs} = \frac{(1+\beta^2)R_{lcs}P_{lcs}}{R_{lcs}+\beta^2 P_{lcs}} \tag{3}$$

where $\beta = \frac{P_{lcs}}{R_{lcs}}$ when $\frac{\partial F_{lcs}}{\partial R_{lcs}} = \frac{\partial F_{lcs}}{\partial P_{lcs}}$. This value is an F-measure. In Table 2 we show the comparison between the ROUGE-L scores of the aspect-aware and aspect-blind translations, by taking a weighted average over the entire dataset based on the number of words in the source and target sentence. We also show the score of the translation with the highest ROUGE-L score with the rest of the translations which has been used in the dataset.

Aspect extraction is treated as a sequence labeling task and is evaluated using the Fleiss Kappa metric (Fleiss and Cohen, 1973). Fleiss' Kappa is a multiclass inter-annotator agreement score which is computed as follows:

$$\kappa = \frac{\bar{P}-\bar{P}_e}{1-\bar{P}_e} \tag{4}$$

where $P - P_e$ is the actual degree of agreement achieved and $1 - P_e$ is the degree of agreement above chance. Given N tokens to be annotated and n annotators, with k categories to annotate the data. We first calculate the proportion of annotations in the j^{th} domain as:

$$p_j = \frac{1}{Nn}\sum_{i=1}^{N} n_{ij}\ , \quad 1 = \sum_{j=1}^{k} p_j \tag{5}$$

We then calculate P_i, the degree of agreement with the i^{th} annotator as:

$$P_i = \frac{1}{n(n-1)}\sum_{j=1}^{k} n_{ij}(n_{ij}-1) \tag{6}$$

$$= \frac{1}{n(n-1)}\left[\left(\sum_{j=1}^{k} n_{ij}^2\right) - n\right] \tag{7}$$

Finally we calculate $\bar{P}$ and $\bar{P}_e$ as:

$$\bar{P} = \frac{1}{N}\sum_{i=1}^{N} P_i \tag{8}$$

$$\bar{P}_e = \sum_{j=1}^{k} p_j^2 \tag{9}$$

The Fleiss' Kappa scores of the aspect-aware, aspect-blind and overall translations are provided in table 2. The high Fleiss' Kappa scores indicate the confidence in the aspect identification guidelines.

Note that the ROUGE-L score of the aspect-aware translation is higher than the overall as well as the aspect-blind translations, as translators often resorted to word-for-word translations in order to preserve each and every aspect of the sentence with its associated semantic information. Note that ROUGE-L is the weighted average of the F-measure taken over all the sentences in the dataset, weighted based on the number of words in the source and target sentences. For the overall ROUGE score, the weighted average was taken over the dataset, weighted based on the number of words in the sentence which gave the highest comparative score for each translation.

Another important insight into the corpus is the difference in aspect coverage between the aspect-aware and the aspect-blind translations. As mentioned in 2.2, aspect-blind translations often dropped aspects due to constraints in syntactic representation or incoherent translation due to sentence semantics, such as due to complex idiomatic phrases. The difference in aspect coverage was seen in about 6% of the corpus, specifically, 358 sentences overall.

3 Evaluating the Dataset

In this section, we detail the evaluation of our translated aspect extraction dataset. We evaluate our dataset using multiple monolingual and multilingual models. The monolingual models are trained and tested on the individual language datasets while the multilingual models involve the use of transfer learning from the SemEval-2014 dataset to the dataset we have created.

3.1 Monolingual Aspect Extraction

We evaluate our dataset against the existing Hindi dataset and the SemEval 2014 dataset using the following baselines:

- **CRF**: We use a conditional random field with basic features[3] such as word form and POS tag.

- **BiLSTM**: We use a vanilla BiLSTM as a baseline model for aspect extraction as it is an established baseline in seq2seq tasks (Liu et al., 2015).

- **BiLSTM-CRF**: We use a BiLSTM to encode the input sentence and a conditional random field for the sequence labeling. This is a commonly used baseline for sequence tagging tasks (Huang et al., 2015).

We also use the following neural models for our analysis:

- **BiLSTM-CNN-CRF**: The state-of-the-art in neural named entity recognition. The architecture uses both character and word level features in a CNN and BiLSTM respectively, and using a CRF for sequence labeling tasks (Reimers and Gurevych, 2017). We use a slightly modified version where word embeddings are generated by concatenating character embeddings, as done by Prabhu et al. (2019) for event detection in Hindi.

- **DeCNN**: The commonly adopted model for aspect extraction specifically, this model uses a combination of general and domain based embeddings in multiple convolutional layers and a fully connected layer with softmax for label prediction (Xu et al., 2018).

- **Seq2Seq4ATE**: This model is a sequence-to-sequence model for aspect terms extraction. The model uses a BiGRU encoder and a position aware attention variant of gated unit networks as a decoder with softmax for label prediction (Ma et al., 2019).

For consistency, in all the above mentioned models, we use the FastText embeddings for word as well as character embeddings for both English and Hindi (Bojanowski et al., 2017; Mikolov et al., 2018; Grave et al., 2018). For the English dataset, we use the Pontiki et al. (2014) train-test split (3045 training to 800 test sentences and 2000 training to 676 test sentences in the 'Laptop' and 'Restaurant' domains respectively). For the Hindi dataset, we use a train-test split of 4062 train to 1355 test sentences based on Akhtar et al. (2020). For the LSTM based models, we use 128 unit LSTM layers, with a hidden size of 1024, and a dropout of 0.4 over 50 epochs. For the CNN based model, we use a 128 filter network with a kernel size of 5 and hidden embeddings of size 100 and dropout of 0.4 over 50 epochs.

We find that the Seq2Seq4ATE model is the best performing model for this task across the datasets. We see that the model performance on our dataset is close to that on the English dataset. While the human aspect extraction baseline shows that there is a lot more work to be done in this task, our dataset provides an adequate baseline for this task, similar to those in the SemEval Aspect Extraction subtask (Pontiki et al., 2014).

3.2 Leveraging Parallel Data

As mentioned in section 2, the corpus we have developed aims to be a parallel corpus, which allows us to use language invariant, transfer learning based models for aspect extraction in Hindi. We use the BERT mutilingual sentence embeddings (Devlin et al., 2018) as the sentence representations for the English and Hindi on the (a) BiLSTM, (b) BiLSTM-CNN-CRF and (c) the Seq2Seq4ATE models, mentioned in Section 3.1. The BERT multilingual embeddings have been used for a variety of tasks in Hindi including machine comprehension (Gupta and Khade, 2020) and named entity recognition (Pires et al., 2019), among other sequence labeling tasks. Pires et al. (2019) showcases the model efficacy in using monolingual corpora for zero-shot code-mixed tasks as well, which would be useful for our corpus.

[3] https://sklearn-crfsuite.readthedocs.io/en/latest/tutorial.html

Model	Akhtar et al. (2016)	Pontiki et al. (2014)	Our Dataset
Baselines			
CRF	22.08	**54.97**	47.07
BiLSTM	20.71	**61.01**	54.77
BILSTM-CRF	34.71	**62.61**	50.26
Neural SoTA Models			
BiLSTM-CNN-CRF	36.56	**73.03**	67.08
DeCNN	38.21	**77.67**	68.35
Seq2Seq4ATE	35.04	**78.86**	68.61

Table 3: F1 scores of established models on the monolingual aspect extraction task.

Training	Model	F1-score
Baseline	BiLSTM	41.06
	BiLSTM-CNN-CRF	54.92
	Seq2Seq4ATE	43.51
Zero-shot	BiLSTM	40.72
	BiLSTM-CNN-CRF	56.16
	Seq2Seq4ATE	42.08
Fine-tuned	BiLSTM	57.37
	BiLSTM-CNN-CRF	62.12
	Seq2Seq4ATE	**66.28**

Table 4: F1-score of the models by leveraging English aspect extraction data using M-BERT. The baseline score is based on using Hindi for training as well as testing.

We design three experiments for evaluating our dataset using M-BERT, which are detailed below.

1. *M-BERT baseline* where we train and test on the Hindi sentences and aspects from our dataset directly, using the M-BERT embeddings. This has been done to establish a baseline for our experiments that follow for leveraging the English data.

2. *Zero shot aspect extraction for Hindi* where we train using the English dataset and evaluate the model performances on the Hindi data, in order to estimate how much aspect information can be extracted about aspect representation in this data which can be applied on the Hindi dataset directly.

3. *Fine tuned aspect extraction for Hindi* where we train the models on the Hindi and a small part of the English dataset and test on the translated Hindi test set. In this experiment, we augment the training data and therefore showcase the use of the English representation of aspect terms in the dataset. This is done with the motivation to boost the token representation of English tokens, as the Hindi data contains English tokens in the form of proper nouns. These tokens are aspects in a part of the corpus and therefore introducing this experiment improves the representation and extraction of these aspect tokens.

Table 4 provides the F1-scores of the various models described above. We use the pretrained BERT Mulitilingual cased model. The best performing model is the fine-tuned Seq2Seq4ATE model with an F1 of 66.28. We also see that the zero-shot performance of the BiLSTM-CNN-CRF is better than the baseline, and that fine-tuning using English data definitely helps the model.

4 Conclusion

In this paper, we detailed the state of aspect extraction in Hindi by thoroughly analyzing and evaluating the currently available baseline dataset for this task. By understanding the flaws in that dataset, we explain its inadequacy in terms of lack of uniformity, high domain sparsity and incorrect aspect annotations. We further compare its performance with existing models to show that it performs very poorly as compared to the existing English dataset.

We then explain the mechanism of creating a SemEval style corpus for aspect extraction in Hindi, by translating the English SemEval 2014 aspect based sentiment analysis corpus. We provide a detailed list of guidelines in order to make this task as replicable as possible. We also focus on maintaining the naturalness and fluency of the translations using transliteration wherever necessary. Our translation and annotation methodology is evaluated on the ROUGE-L and Fleiss' Kappa metrics respectively.

We use this dataset to show performance on basline statistical and neural sequence labeling models, as well as the current state-of-the-art models in neural aspect extraction such as DeCNN and Seq2Seq4ATE. We show that while the published Hindi dataset does not perform nearly as well, we provide comparable results to those models. Since we have a parallel corpus, we also leverage the English data for improving aspect extraction in Hindi using multilingual BERT.

Future work in this direction includes developing an aspect based sentiment analysis corpus which can be trained and tested in a multilingual manner and fine-tuning multilingual BERT for few-shot and zero-shot sequence labeling tasks.

Acknowledgements

We would like to thank the annotators of the corpora, Mukund Chaudhary, Mashrukh Islam, Ishaan Sanjeev Upadhyay, and KV Aditya Srivatsa, who took time out and volunteered to help out with this project. We would also like to thank the anonymous reviewers for their insightful comments which aided in improving this contribution.

References

Md Shad Akhtar, Asif Ekbal, and Pushpak Bhattacharyya. 2016. Aspect based sentiment analysis in hindi: resource creation and evaluation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2703–2709.

Md Shad Akhtar, Tarun Garg, and Asif Ekbal. 2020. Multi-task learning for aspect term extraction and aspect sentiment classification. *Neurocomputing*.

Muhammad Zubair Asghar, Aurangzeb Khan, Syeda Rabail Zahra, Shakeel Ahmad, and Fazal Masud Kundi. 2019. Aspect-based opinion mining framework using heuristic patterns. *Cluster Computing*, 22(3):7181–7199.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Joseph L Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3):613–619.

Lea Frermann and Alexandre Klementiev. 2019. Inducing document structure for aspect-based summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6263–6273.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Somil Gupta and Nilesh Khade. 2020. Bert based multilingual machine comprehension in english and hindi. *arXiv preprint arXiv:2006.01432*.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 605–612.

Pengfei Liu, Shafiq Joty, and Helen Meng. 2015. Fine-grained opinion mining with recurrent neural networks and word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1433–1443.

Dehong Ma, Sujian Li, Fangzhao Wu, Xing Xie, and Houfeng Wang. 2019. Exploring sequence-to-sequence learning in aspect term extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3538–3547.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Preslav Nakov, Torsten Zesch, Daniel Cer, and David Jurgens. 2015. Proceedings of the 9th international workshop on semantic evaluation (semeval 2015). In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001.

Maria Pontiki, Dimitrios Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad Al-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, et al. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In *10th International Workshop on Semantic Evaluation (SemEval 2016)*.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

Suhan Prabhu, Pranav Goel, Alok Debnath, and Manish Shrivastava. 2019. A language invariant neural method for timeml event detection. In *Proceedings of International Conference on NLP (ICON)*.

Lance A Ramshaw and Mitchell P Marcus. 1999. Text chunking using transformation-based learning. In *Natural language processing using very large corpora*, pages 157–176. Springer.

Yashwanth Reddy Regatte, Rama Rohit Reddy Gangula, and Radhika Mamidi. 2020. Dataset creation and evaluation of aspect based sentiment analysis in telugu, a low resource language. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5017–5024.

Nils Reimers and Iryna Gurevych. 2017. Reporting score distributions makes a difference: Performance study of lstm-networks for sequence tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 338–348.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, pages 502–518.

Hu Xu, Bing Liu, Lei Shu, and S Yu Philip. 2018. Double embeddings and cnn-based sequence labeling for aspect extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 592–598.