figure, ABCD are the Nepali word inputs at each time step and XYZ are the English word outputs at each time step. <eos> is the end of sentence tag. On the decoder side, if the output of one-time step is used as the input to the next time step, the process is called teacher forcing. So, an optimization with both using and non-using of teacher forcing will result in the state-of-the-art project.

1.1 Motivation and Significance

Machine Translation has a history way back from world war to deep learning currently. It all started with an attempt to convert Russain Language to English Language. Then, each russian word was defined as an english word and the russian statements were converted to english with word by word transformation. However, the invention was considered useless because the device that was used for translation did not convey any meaning on the output. In 1966, there was another attempt by the US ALPAC committee but it was also useless. So, focusing on the development of a dictionary was more taken into account. It was a different condition then. But, currently, Machine Translation has developed on an extensive level. Especially, with the use of Natural Processing techniques and Deep Learning, machine translation has become easy and also more convincing. However, it feels really bad to see that the machine translation in our own native language has not developed it. There are still very few contributions. There are still no translation datasets in TensorFlow datasets. And, google translate also does not provide completely accurate results. So, this really motivated us to take this project.

The significance of Machine Translations is extensive. English to Nepali machine translation can be used for simplification of conversation between people speaking different languages. The search engines are poor and do not provide accurate results. So, they can also be used for search engine optimizations. They are cheaper and better to use instead of having a personal translator to translate languages for us. They are used for marketing to communicate with customers of different natives for buying and selling of the products too. Moreover, with the advancement of machine translation, every human will have a common language to speak and communicate in a better way.

1.2 Objectives

The objectives of this project are:

- To translate English Nepali language to English in a better and efficient way.
- To implement this concept of machine translation for the Search Engine Optimizations
- To learn how encoder-decoder models with attention are used to learn the relationship between words in two different languages.
- To obtain a bleu score of at least 25% and interpret the results.

2. Literature Review

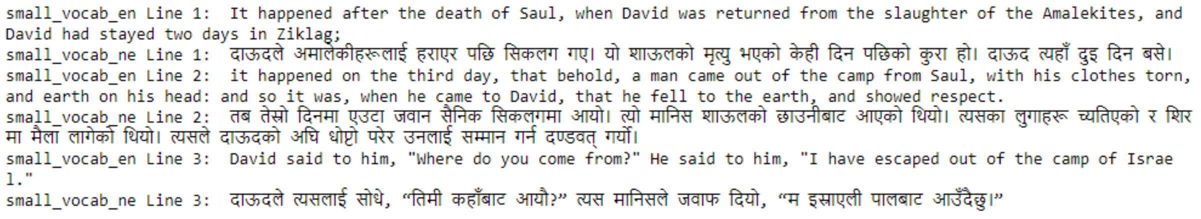
Statistical Machine Translation (SMT) has been the dominant translation paradigm for decades. Practical implementations of SMT are generally phrase-based systems (PBMT) which translate sequences of words or phrases where the lengths may differ. Even prior to the advent of direct Neural Machine Translation, neural networks have been used as a component within SMT systems with some success. Perhaps one of the most notable attempts involved the use of a joint language model to learn phrase representations which yielded an impressive improvement when combined with phrase-based translation. This approach, however, still makes use of phrase-based translation systems at its core, and therefore inherits their shortcomings. Other proposed approaches for learning phrase representations or learning end-to-end translation with neural networks offered encouraging hints, but ultimately delivered worse overall accuracy compared to standard phrase-based systems.

Neural machine translation is a recently proposed approach to machine translation. The models proposed recently for neural machine translation often belong to a family of encoder–decoders and encode a source sentence into a fixed-length vector from which a decoder generates a translation. In the paper by Dzmitry Bahdanau, they conjecture that the use of a fixed-length vector is a bottleneck in improving the performance of this basic encoder–decoder architecture, and propose to extend this by allowing a model to automatically (soft-)search for parts of a source sentence that are relevant to predicting a target word, without having to form these parts as a hard segment explicitly. With this new approach, they achieved a translation performance comparable to the existing state-of-the-art phrase-based system on the task of English-to-French translation.

3. Dataset

We begin by investigating the dataset that will be used to train and evaluate NLP pipeline. The dataset used for **Neural Machine Translation on Nepali Texts** is available at corpus.zip. However, that will take a long time to train a neural network on and also for preprocessing. Therefore, We used an identical but smaller dataset we found on GitHub: Small_Dataset.zip that contains a small vocabulary so that we will be able to train our model in a reasonable time with this dataset and will be helpful for model experimentation. Then we will use the original dataset after finding the right model for training and testing.

The data is located in data/small_vocab_en and data/small_vocab_ne. The small_vocab_en file contains English sentences with their nepali translations in the small_vocab_ne file. Each line in small_vocab_en contains an English sentence with the respective nepali translation in each line of small_vocab_ne.

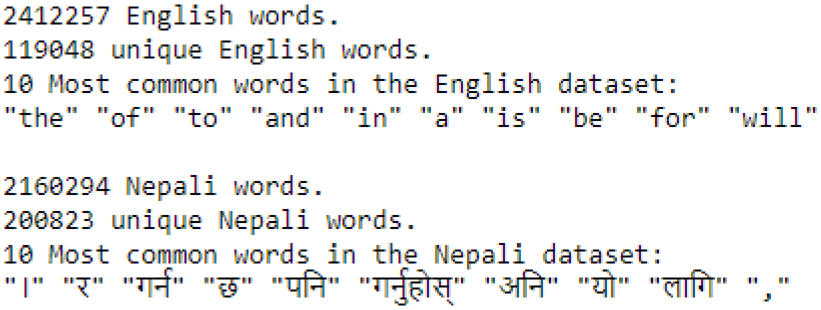


```
small_vocab_en Line 1:  It happened after the death of Saul, when David was returned from the slaughter of the Amalekites, and
David had stayed two days in Ziklag;
small_vocab_ne Line 1:  दाऊदले अमालेकीहरूलाई हराएर पछि सिकलग गए। यो शाऊलको मृत्यु भएको केही दिन पछिको कुरा हो। दाऊद त्यहाँ दुइ दिन बसे।
small_vocab_en Line 2:  it happened on the third day, that behold, a man came out of the camp from Saul, with his clothes torn,
and earth on his head: and so it was, when he came to David, that he fell to the earth, and showed respect.
small_vocab_ne Line 2:  तब तेस्रो दिनमा एउटा जवान सैनिक सिकलगमा आयो। त्यो मानिस शाऊलको छाउनीबाट आएको थियो। त्यसका लुगाहरू च्यतिएको र शिर
मा मैला लागेको थियो। त्यसले दाऊदको अघि धोप्टो परेर उनलाई सम्मान गर्न दण्डवत् गर्यो।
small_vocab_en Line 3:  David said to him, "Where do you come from?" He said to him, "I have escaped out of the camp of Israe
l."
small_vocab_ne Line 3:  दाऊदले त्यसलाई सोधे, “तिमी कहाँबाट आयौ?” त्यस मानिसले जवाफ दियो, “म इस्राएली पालबाट आउँदैछु।”
```

Figure 2. First two line of original data

3.1 Vocabulary

The complexity of the problem is determined by the complexity of the vocabulary. A more complex vocabulary is a more complex problem. Therefore, we looked at the complexity of the dataset.



```
2412257 English words.
119048 unique English words.
10 Most common words in the English dataset:
"the" "of" "to" "and" "in" "a" "is" "be" "for" "will"

2160294 Nepali words.
200823 unique Nepali words.
10 Most common words in the Nepali dataset:
"।" "र" "गर्न" "छ" "पनि" "गर्नुहोस्" "अनि" "यो" "लागि" ","
```

Figure 3. Vocabulary