# Paraphrase Extraction with Neural Machine Translation


## Abstract

Neural machine translation is a recent and promising approach to solve machine translation problem. A major weakness of typical NMT systems is their inability to translate rare words that occur few times in the training corpus. One potential way to overcome this is through the use paraphrases of rare words, which the NMT system might be able to translate. In this work, we address the first step in the task - extracting prospective paraphrases. We propose a method to extract paraphrases for a source language using its translation by an NMT system. We follow the process of bilingual pivoting across the source-target language pair for extracting paraphrases.


## 1 Introduction

Paraphrases are alternate ways of conveying the same information. They are instrumental in tasks like sentiment analysis, sarcasm detection, question-answering systems and abstractive summarisation. Paraphrases have also been shown to be useful for improving performance of statistical machine translation [3]. Neural machine translation(NMT), a more recent approach in machine translation, can benefit from the use of paraphrases. In particular, paraphrasing can address the rare word problem that is prevalent in the NMT paradigm.

NMT systems are trained using a parallel corpus of source and target language sentences. Ability of a trained NMT system to translate a source word depends on whether that word was encountered sufficiently large number of times during training. This leads to the ”rare word” problem where NMT system cannot make a translation, and instead copies the source word or prints *UNK* (for unknown) in the target side. In such scenario, using a paraphrase table increases the chance that NMT system will be able to translate the information to target language.

An effective method for paraphrase extraction suggested by Bannard et. al. [2] involves pivoting about a foreign language phrase to extract paraphrases in a language. An example is shown in Figure 1 where the German phrase *unter kontrolle* maps to paraphrases *under control* and *in check* in English. We follow a similar approach in our work, but with two notable differences. Firstly, foreign language is the target language of our NMT system, and therefore, not a manually verified translation of the source language text. Secondly, we do not have hard alignments between source and target sentences, and we map back to the source side through soft alignments(attention) information from the NMT system.

In [2], a probability was assigned to each paraphrase pair using the counts of source phrase-target phrase occurrences. They also carried out re-ranking of the probability scores using a language model, and final evaluation of extracted paraphrases was performed manually. We follow a slightly different method for evaluation of extracted paraphrases. As our task is linked to NMT, we envision to evaluate paraphrase quality indirectly through machine translation performance as an extension to this course project.

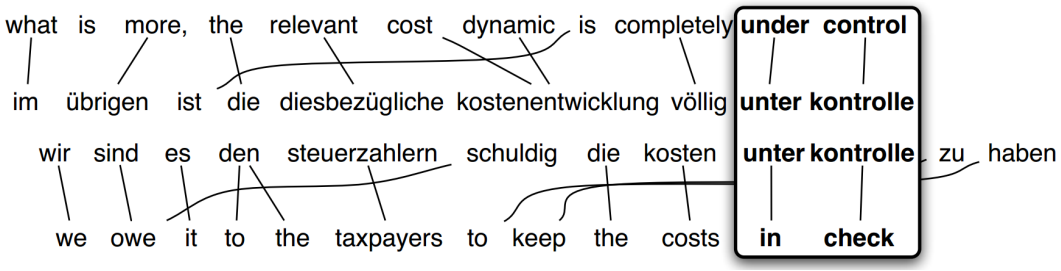


Figure 1: Pivoting about a target language(German) phrase to extract paraphrases in the source language(English). Figure from Callison-Burch et. al [3].

We organize this report as follows. In Section 2, we discuss briefly the basics of neural machine translation with attention. In Section 3, we describe our algorithm to extract paraphrases. In Section 4, we describe our experiments and illustrate some results from our work. Finally, we conclude and provide details on future work in Section 5.

## 2 NMT with attention

In neural machine translation, a source language text is translated to a target language one sentence at a time. Typical NMT systems have an encoder-decoder framework where both encoder and decoder are some kind of gated recurrent neural network(LSTM or GRU). A source sentence, tokenized into words(or sub-words or characters), is fed into the encoder one token per time step. End of the sentence is marked by a *EOS* token, after which the decoder starts producing output words (or sub-words or characters). The target sentence is completed when decoder produces *EOS* token.

In the system described above, source sentence is encoded into a fixed length vector - context vector of the last encoder unit, which is then decoded by the decoder. This limitation of having fixed size encoding for every sentence was addressed in Bahdanau et. al. [1] by using an attention mechanism for decoding. Attention model provides soft alignments by using a weighted sum of all encoder context vectors for predicting each target token, rather than using context vector of only the last encoder unit. While giving a variable length representation for a sentence, attention model also allows to selectively attend to source words that are relevant to predict each target word. This model gives an attention matrix for each source-target sentence pair.

Figure 2 shows the attention mechanism used in [1]. Note that the encoder uses forward and reverse GRU in this work, and context vector for each input token is concatenation of forward and reverse context vectors. $\alpha$ values shown in the figure are the weights (in attention matrix) for predicting $y_t$ using all the input context vectors.

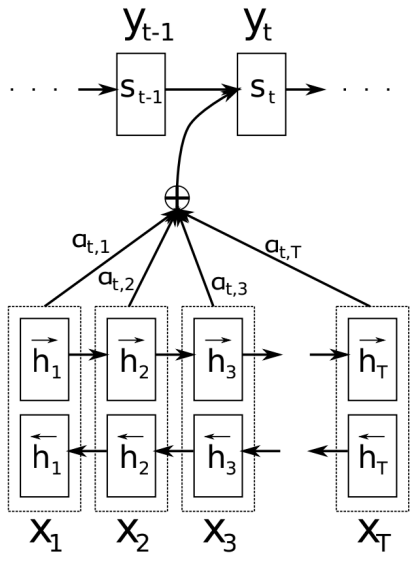


Figure 2: A source sentence $x_1, x_2, ..., x_T$ is being translated at the $t^{th}$ step of decoding. Figure from Bahdanau et. al. [1].

## 3 Paraphrase extraction algorithm

We use Edinburgh Neural Machine Translation System for WMT16 [7] as a backbone for our paraphrase extraction mechanism. First, we translate the source language using our NMT system. Next, we find identical words on the target side, after cleaning it by removing stopwords. We create a word

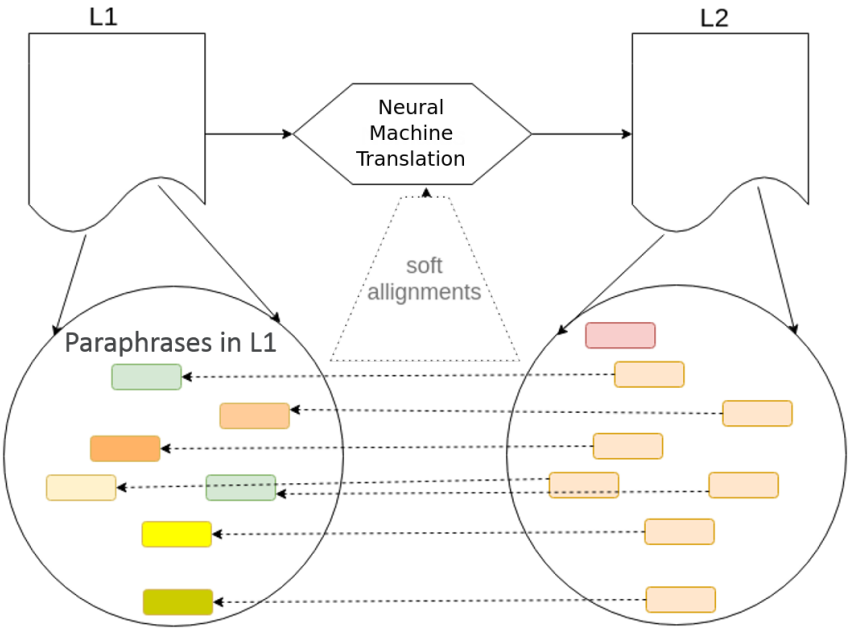


Figure 3: A graphical representation of paraphrase extraction algorithm, where L1 is the source language and L2 is the target language.

dictionary on target side, `word-dict`, with the word addresses, arranged in the order of word-level frequency distribution. We then map back to the source side using the alignment matrices that we save during the translation task, and extract one or more source words corresponding to each target word. Figure 3 illustrates this process graphically.

The NMT system we use operates at the level of sub-word units [8], hence a direct thresholding in the attention matrix is not meaningful for word extraction. Figure 4 shows an example of attention matrix that we obtained for translation from English to German. Using word address information from `word-dict`, we select attention matrices and corresponding source and target sentences. To extract complete words from source language for a given sentence pair, we first group together sub-word units of words on both sides. We apply a threshold on attention matrix rows for each sub-word unit of the target word from `word-dict`, and select source word(s) that contain mapping sub-word unit(s). The process is repeated for all the word addresses for a given target word in `word-dict` to produce a word list in source language. A paraphrase list corresponding to the target word is obtained by removing duplicates from this word list.

## 4 Experiments and Results

To ensure productivity, we restrict our algorithm to extract unigram paraphrases. For conducting experiments, we use the NMT system mentioned in Section 3 trained on the English-German parallel corpus from the WMT 16 dataset [4]. We use this NMT system to produce translations of the first 29500 lines from En-De (English-German) parallel corpus from *Europarl v7* [5], English being the source language.

The first set of experiments were fully unfiltered. The target side `word-dict` is used as is without any pruning or a threshold value. For the second set of experiments, we applied filters to the target side `word-dict` and the soft-alignments attained through the attention weights. Minimum frequency of target word is set to 5, with the motivation that given a corpus size of 30,000 if a word doesn't appear at least 5 times, its quite infrequent and likely to be unyielding to expend computing on. A second filter is applied to the upper-level of frequency. Words which appeared too frequently in the dictionary, e.g. 4000 times in a corpus of 30,000, were pruned. The intuition being words of such high frequency might as well be stopwords (e.g. the target language equivalents translations of

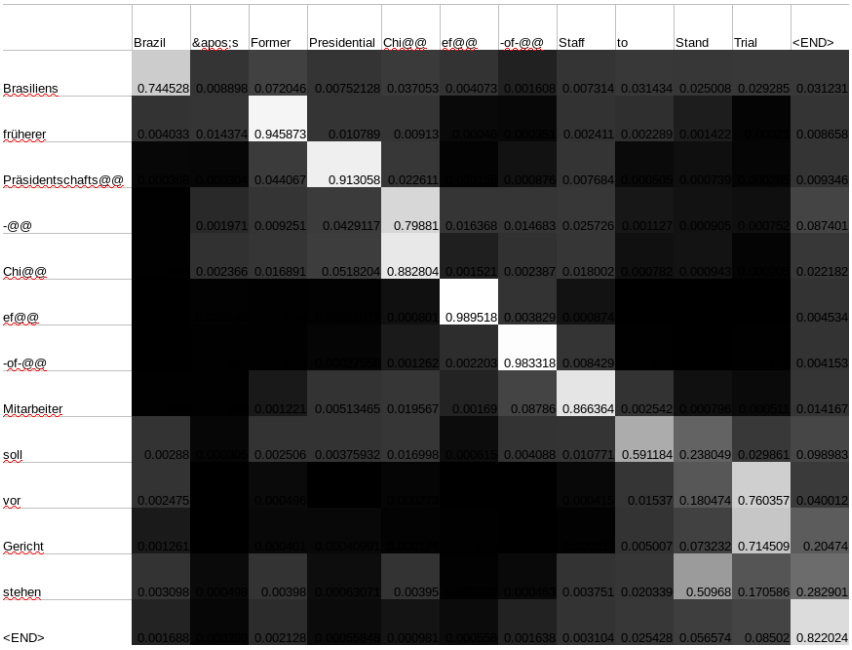


Figure 4: Attention matrix with soft alignment weights used in translation. Sub-word units of source sentence in English is given along the row on top of the matrix, and sub-word units of target translation in German is given along the column.

Table 1: Paraphrases obtained in the two experiments.

| **German Word** | **English Words** |
|---|---|
| **Experiment 1 (Unfiltered)** | |
| ein | a, a,a ,a ,a ,a ,a, ... |
| das | the, the, the, ... |
| im | in, in, in, into, in, .... |
| **Experiment 2 (Filtered)** | |
| Mensch | people, humans, person, persons, individual, human, man, Man |
| Sprecher | spokesperson, spokesman, speakers |
| Dauer | duration, length, period, lasting |
| vermittelt | brokered, conveyed, mediated |
| falschen | incorrect, secondly, erroneous, false, misinterpretation, misconceptions , wrong, misconception |
| Katastrophe | disaster, maritime, Amoco, catastrophes, Erika, widespread, Chernobyl, catastrophe |

*a, an, the, of, an, if* etc) which contribute absolutely nothing towards the distributed word-similarity we aim to leverage. Finally, the threshold for attention matrix value was set to 0.5.

As an overall quality filter, we remove any duplicates within a set of paraphrases to avoid any kind of noise of false results.

From the corpus of 29500 lines which we used for our experiments, our model extracted 14866 set of paraphrases in Experiment 1 and 11493 set of paraphrases in Experiment 2 as shown in Table 1. Although the extracted paraphrases look appealing and not very skewed, an evaluation measure can help understand how coherent the mechanism is under the hood. For evaluating, we trained word2vec [6] on the Common Crawl (840B tokens, 2.2M vocab, cased) and English Gigaword 5 dataset. This produced a 300 dimension vector representations of English words in a 10 GigaBytes raw file.

For next step in the evaluation, we replaced candidate words into each of the respective sentences across the entire dataset with all the paraphrases obtained.

| |
|---|
| The doctor believed that Jack lacked **attention**. |
| ⏐ |
| The doctor believed that Jack lacked **focus**. |

After replacing paraphrastic words across the corpus, we compute pairwise cosine distance $\cos\theta(h, h')$ for each sentence pair, where $h$ is the vector representation of a sentence and $h'$ is the vector representation of its variant with a word replaced from the paraphrase list. The vector representations are obtained using word2vec as mentioned earlier. Using cosine-distance, we produced a t-SNE plot of 11 randomly selected paraphrases as shown in Figure 5.

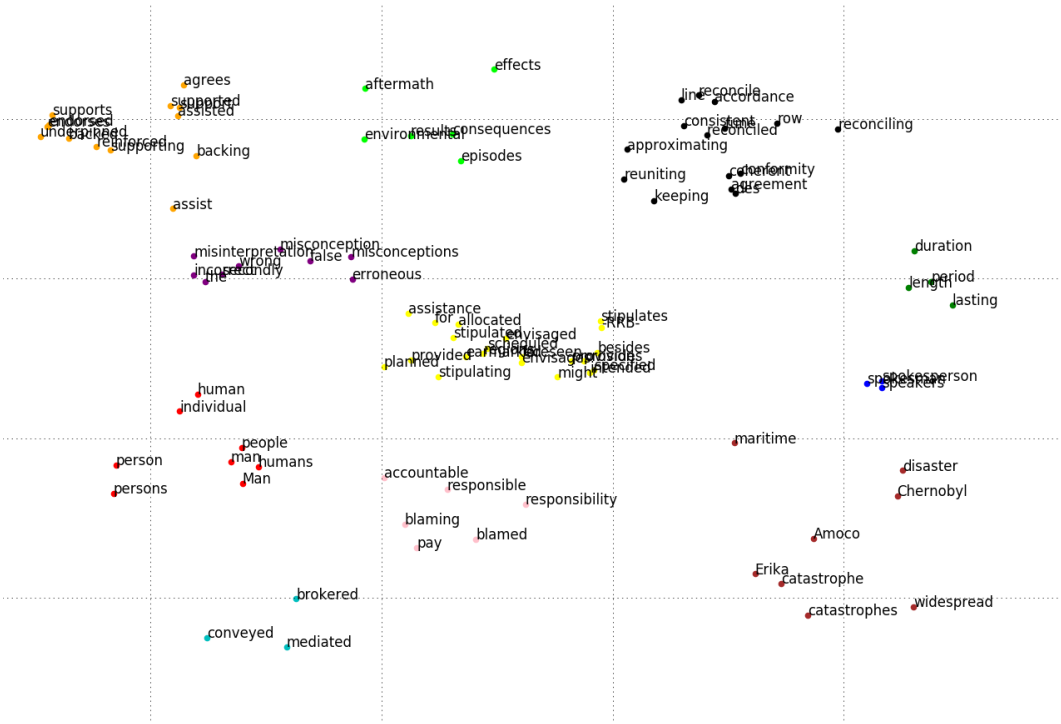


Figure 5: t-SNE plot for the paraphrase lists extracted for 11 different target words. Same color is used for paraphrases from a given list.

| **Paraphrase lists** |
|---|
| willingly, gladly, happily, happy |
| justice, condemned, condemning, denounce, condemn |
| incidentally, moreover, passing, furthermore |
| blame, fault, guilt, culpability, blaming |
| recuperation, recovery, regeneration |
| unduly, excessively, overly |
| EU, representation, female, women |
| brave, bold, courageous |
| objectives, set, targets |
| satisfaction, gratifyingly, more |
| stringent, tougher, stronger, firmer, stricter, tighter |
| still, political, continues, continue, continuing |
| stringent, tight, strict, severe, stern, rigorous |
| injured, breaches, breached, infringed, violated, contravened, openly |
| grasp, understood, understand, got |
| dispute, disagreement, contention, quarrel |
| authoritative, powerful, persuasive, convincing |
| delay, delaying, delayed, slippage |
| ask, question, enquire, wondering, wonder |
| scheduled, geplante, planned, proposed, envisaged |
| strengthened, increased, enhanced |
| Turkey, Greek, Turkish |
| assistance, remedied, remedy, remedying |
| agreeing, agreed, agree |
| ratio, relationship, percentages, proportion, relation, proportional |
| justified, justification, Rat, legitimate, justify |
| building, capacity-building, construction, build, establishment, setting |
| duly, orderly, properly |
| talked, advocated, spoke, talking, preaching, referred |
| transmission, transfer, transferring, transferred, transfers, delegating |
| promises, promise, pledges, pledge |
| Fish, fish, fishing |
| prompt, expeditiously, swiftly, speedily, quickly, swift |
| bases, builds, establishes |
| setting, set, base, establishing |
| raised, made, levelled, elevated, charged |
| threshold, dawn, maximum |
| shown, proven, unwieldy |
| ideas, imaginations, conceptions, misconceptions, beliefs |
| shipping, maritime, sea |
| alas, sadly, unfortunately, regrettably |
| frankly, outstanding, candidly, frank, openly, in, open |
| personnel, manpower, human, staff |
| strengthen, strengthened, strengthening, bolstering, reinforced |
| electorate, constituents, electors, voters |
| occurred, arose, arisen, occurring |

## 5 Conclusion and future work

We have successfully developed and implemented a paraphrase extraction model using NMT with attention model. The above table shows a few randomly chosen paraphrases that were extracted using our proposed algorithm. Apart from the very human readable and perceivable paraphrases that we have achieved, one observation we make while computing cosine-similarity of candidate paraphrases is that for potential paraphrases, the score is almost always greater than 0.7 while for non-paraphrastic words or words out of context, the score is generally less than 0.4.

In future work, there are two definite approaches we hope to try. First, to extend paraphrase extraction mechanism from unigrams to n-grams (actual phrases instead of single words). Second, we want to use the paraphrases to address the rare-word problem in a machine translation system. Another avenue to explore would be to improve paraphrase quality by using an LSTM language model in lieu of a syntax based language model to re-rank paraphrases, as mentioned in [2].

**Contributions**

**Acknowledgments**

We are grateful to Dr. Anoop Sarkar for his insightful suggestions. This work wouldn't have been possible without his supervision.

## References

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv*, 2014.

[2] Colin Bannard and Chris Callison-Burch. Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 597–604, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

[3] Chris Callison-Burch, Philipp Koehn, and Miles Osborne. Improved statistical machine translation using paraphrases. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 17–24, New York City, USA, June 2006. Association for Computational Linguistics.

[4] http://www.statmt.org/wmt16/translation task.html. Shared task: Machine translation of news. In *WMT*, 2016.

[5] Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86, 2005.

[6] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[7] Rico Sennrich, Barry Haddow, and Alexandra Birch. Edinburgh Neural Machine Translation Systems for WMT 16. In *Proceedings of First Conference on Machine Translation (WMT2016)*, 2016.

[8] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of ACL*, 2016.