WOVEN NMT Model Tests
BLEU Scores with sacreBLEU

The evaluation of machine translation (MT) models has been challenging due to the fact that there is no *one correct way* to interpret and hence translate a text. Language is a means by which we express an idea in a mental format; The true representation of meaning exists only in the mind. Language is a crude form of this representation and can be filled with ambiguity, consider for instance the multitude of synonyms available in English with very minute differences in meaning.

However, the community has made significant progress with metrics that may not tell you if a translation is the “best” (even translators would argue over that), but will reliably indicate a difference in quality from one model to the next.

BLEU (Bilingual Evaluation Understudy)
An understudy is someone who is focused on learning on the back of another’s work. This is precisely what one very popular evaluation metric aims to do.

BLEU uses a source text, a candidate set, and a reference set. The source text is the base text to be translated. The candidate set is a one-dimensional array which comprises the output of the MT model after translating the source text. The reference set is a two-dimensional array which contains human translations of that same source text, where there can be multiple translations for each sentence in the source text.

The proportion of words (or *tokens*, more accurately) that are present in both a candidate and its respective references is calculated. If a token appears more times in the candidate than in any of the references it will be penalised. We refer to this as a *modified unigram precision.* Applying this to sets of multiple tokens will give us *modified n-gram precision*. Accounting for n-grams will penalise models for including too many or too few words which do not exist in any of the references.

To compute the BLEU score, the geometric mean of n-gram precisions is taken and multiplied by a brevity penalty, which penalises candidates that are not of the same length as any of the references. This results in a range of [0,1], where closer to one is interpreted as a higher similarity to the human translations. Note however that BLEU scores are most often expressed as a percentage. An exact score of 1 (or 100) means the candidate is identical to a reference; even a fluent human translation may not result in this score merely due to the reference not existing in the reference set.

Criticism of BLEU
If an MT model in one paper receives a score of 35 and in another a score of 37, you cannot immediately know whether one is ‘better’ than the other. In scoring translations, a tokeniser must first be used to parse the output string into individual elements, or tokens. How these tokens are formed can dramatically change the BLEU score of an MT model using the same references.

However, these difficulties have been tackled through projects which aim to standardise the tokeniser used. In this project I will take advantage of these developments with *sacreBLEU.*

SacreBLEU is a project developed with the motivation to standardise the tokeniser used in BLEU score calculations and hence standardise BLEU across the research landscape. It serves as an all-in-one scorer handling n-gram calculations, their geometric mean, and the application of the brevity penalty.

WMT20 Test Set
The model is tested on 1400 English to German translations. The source text and references used in the BLEU-score calculation were taken from the Fifth Conference on Machine Translation (WMT20). Only a single set of references is used, meaning that for each candidate translation there is one respective reference translation. These are the standards as set by the conference, and therefore very suitable for the purpose of my own tests.

Interpreting Scores
Google has created a rough guideline by which an intepretation of these scores can be made. Less than 10 is described as ‘almost useless’ and 10-19 as ‘difficult to get the gist’, while higher than 60 means the ‘quality is often better than that of a human’ and 50-60 means that translations are of very high quality. From 30 onwards, the translations are considered at the very least understandable. We can see here how high quality can mean a BLEU score which is not the highest score possible but rather reaching a certain threshold.

WOVEN NMT Model Results
The WOVEN NMT model achieved a BLEU score of 31. This falls into the range of ‘understandable to good translations’, and achieves what I set out to do in this phase of the project: create a model which achieves reasonable but not necessarily great translation accuracy using a well-defined metric. Note that this score would likely be somewhat higher if multiple reference sentences were used for each single source sentence, however in this case we limited our score in that respect in order to fall in line with WMT20 standards. The n-gram precision scores can be seen below.