Machine Translation third assignment: Inuktitut-English Translation, the importance of data quality

A. Syahdeini (s1575408), F. Rodriguez (s1670175)

April 14, 2017

Abstract

In this work we explore three different data preprocessing methods applied to a Inuktitut corpus for ulterior automatic machine translation usage. Our results shows that previous analysis and processing of linguistic data can improve the performance of a LSTM based model, and suggest that better methods could be an inexpensive and relatively fast complementary tool. Some ideas for further research are also proposed.

1 INTRODUCTION

1.1 Neural Machine translation

As a Neural network increase its popularity to solve complex computational natural language problem. Sennrich et al. (2016),Kalchbrenner and Blunsom (2013),Cho et al. (2014) proposed an Idea of using Neural Machine translation rather than phrase based translation Koehn et al. (2003). Neural Machine translation or NMT is a large neural network that consist of encoder-decoder which translate a source sentence into a correct sentence. encoder will read the input sentence, which consist of bag of words that transformed into 1-hot-encoder. and decoder will output the translation from the vector of encoder. the basic idea is to maximize the probability of corret translation from source sentence. Bahdanau et al. (2014).

Recently this [people] introduced the idea of using bidirectional NMT which use two direction of an input. Another improvement also introduced by [this stanford guys] who improve the NMT model by using attention-based model. In this assignment, we use bi-directional NMT and soft attention as our model.

1.2 The Inuktitut language

Within the wider category of **agglutinative** languages, which have a notorious tendency of composing complex meanings by combining small meaningful units into one larger linguistic element, the Inuktitut language belongs to the *extreme* **Polysynthetic** type, in which these units (or morphemes) can merge into very elaborated sequences (that an english speaker could interpret sometimes as words, sometimes as sentences). Merging is probably the best word to characterize these syntheses processes, as morphemes not only have different materializations depending on the adjacent morphemes, but also can queue different interpretations. These features, among others, pose a real challenge for Natural Language Processing (NLP), and specially for Machine Tranlation (MT) applications.

1.3 Data preprocessing

Even with its limitations, neural machine translation, is one of the best tools we currently dispose of, and in this context, any method that is able to boost its results, can be considered a complementary tool and result useful either for quick circumstances-based implementations and for future research. As it has been previously exposed, one main problem for Inuktitut-English automatic translation is the great difference between the linguistic systems. Nevertheless, spite of polysynthetic complex word-sentence synthesis, we hypothesized that Inuktitut should have some kind of regularities that could be exploited for improving the input for translation purposes. Motivated by this ideas, on the following work we describe our exploration on different data preprocessing methods that improve the translation model performance by segmenting the Inuktitut corpus.

2 Description of the experimental methods

All the statistics presented below are based on the first 50,000 lines used as input for the LSTM model.
All the code and input data that we have used in this project is available in github.

2.1 Preliminary data analysis

Initial type/token ratios for the original data is: English: 2.174%, Inuktitut: 32.652% (For more statistical data see table 1), this will have some negative effects in our model:
Firstly, the high type/token ratio (given that types can convey sentence-like meanings) of Inuktitut language, and the word-processing of linguistic elements by our model will result in very high rate of _UNK tokens. Also, because of the type/token disparity between both languages, the encoded states will probably will cause unbalanced associations between one English and many Inuktitut types, when the prediction function is called, the probability distribution for many Inuktitut tokens will contain the same indexes as possible outputs, and even if their specific probabilities won't be the same, the overall chance of yielding repeated English tokens (many of the _UNK because of the same) will be very high. The three following presented methods are attempts of tackling these problems.

2.2 Filtered Byte Pair Encoding (BPE)

Our first experiment is based on Byte Pair Encoding (BPE) compression technique. BPE is a word-segmentation adaptation from a compression technique that iteratively replace most frequent pair of character into single unused character for each iteration Sennrich et al. (2016). It works like a counter and combiner that combine characters/word if they occur frequently. It works based on character at a first stage and then start combining the most frequent character into words and replace them as characters. The final result is each character will be decoded into real character/word. At the end we will segment a word by looking at the frequency of the character/word, in our code if the frequency is more than 2 it will get combined into one character. An example of BPE Algorithm can be see on Figure 1.
We applied the BPE segmentation through three steps:
1) Because the the BPE algorithm look for frequency patterns of occurrences for posterior segmentation, we didn't want that proper names, or their phonetic translations (*Mr. Peter Kattuk* -> *pita kattuq*), that shouldn't be segmented, contaminated this frequencies. For this purpose we filtered the lines including any proper names (institutions, people, locations, etc) on it.
2) After this we fed the BPE learning algorithm with the previously filtered data as input.
3) Then we applied the learned patterns to the full Inuktitut text, the result of this process is the segmented Inuktitut text.
In the github repository there's also a very simple script that realigns the texts after the BPE application (a minor, but important bug that corrupts the alignment of the sentences when a non UTF-8 character appears).

2.3 The Uqailaut project

The Uqailaut project is a morpheme based analyzer application developed by the research office at the Institute of Information Technology of the National Research Council of Canada (NRC). The project use Java programming language and it contains useful files for different types of applications.
We used the Uqailaut morpheme analyzer to segment 20,000 sentences in our corpus data. Our intention was to segment the same 50,000 lines using this algorithm as with the other segmentation methods; Unfortunately, our implementation of this segmentator is too slow and after 3 days we only could segment 20,000 sentences. Probably, the problem was that we used the Java executable that they provide executing it in python, as it can be seen in the code.

2.4 Self-defined morpheme segmentation

Because our implementation of the Uqailaut project Java segmentator was taking so long, we decided to build our own morpheme analyzer, which is available in word_segmentation.py. For this we used the Inuktitut set of morphemes provided by the Uqailaut project in which are listed most of the roots and suffixes of Inuktitut. We also exploited the fact that the root is always the first morpheme of the sequence and the attachment of new morphemes occurs only through suffixation. Nevertheless, and as mentioned in the previous section, a distinctive feature of polysynthetic languages (so Inuktitut) is that morphemes can change their realization (*ut* -> *utiq*) depending on the adjacent morphemes, and even if we used the core rules by which the morphemes change (appereance of one of four types of sounds in the last position of the previous morpheme: vowels, t, k, q), some variations are the product of purely phonological variables, like the nasalization of sounds, what introduced some errors on segmentations.

- *root.en.txt* file containing the list of Inuktitut roots
- *suffixes.en.txt* file containing the list of Inuktitut suffixes
- and the *allSuffixes.txt* file containing the rules by which morphemes vary.

	Types	Tokens	Ratio	UNK
Baseline - English	9,996	459,895	2.174%	637
Baseline - Inuk	72,734	222,757	32.652%	7,880
BPE - Inuk	29,870	397,425	7.516%	1,670
Self-defined - Inuk	12,615	659,136	1.914%	878
English (20,000)	6,667	187,738	3.551%	364
Uqailaut - Inuk	19.646	190.742	10.3%	2,131

Table 1: English and Inuktitut word distributions.

We used these files to segment the inuktitut *words* by creating a dictionary with the different variations as keys that returned the original morphemes as their values. We included roots and suffixes into the dictionary after deleting the last letter of morphemes that could be changed. The Algorithm 1. show the pseudocode about our self-defined morpheme analyzer.

3 Results

As we can see in Table 2. the results of all the above described experiments surpass the results of the baseline model. This corroborates the importance of, when is possible, preprocessing the input data before starting to properly work with it.

In table 1. we present the better perplexity/BLEU-score trade-off encountered for each model, this means that even if we found some slightly better BLEU or perplexity values, these values are the the ones in which both measure criteria had a good score.
We can see how all the preprocessing methods have a large impact on the number of Inuktitut types, tokens, UNK tokens and in the type/token ratio. In all of them, the number of UNK goes down drastically, being the the case of the self-defined segmentator the one with less UNK tokens. Because of the same, the self-defined segmentator is also the one with the lowest type/token ratio (with a ratio even lower than the English ratio). Nevertheless, even if we can attribute the improvement on the performance to the data preprocessing, it seems that there's no direct or lineal correlation between the type/token ratio and the results we obtained after we trained our model: While the Uqailaut has a type/token ratio near 10%, the results of the model after its implementation are closely similar to the ones of the self-defined experiment. Being this the case, perhaps we could hypothesize that after some critical point the benefit of keep reducing the ratio is not as effective as with the first reductions. It is important to note that the model based on the Uqailaut morpheme segmentation has run using just half of the data compared with the other 3 models, so we should expect different results (probably better, because of more examples but we can't be certain) with more training data. Below we present the first translation of the development set for each method, for more examples please refer to the appendix or/and to the github repository.

BPE segmentation:
sentence: 45000
Src | isumajunga minisitaujuq tunisijunnar pa uvattinnut ilanginnik uattiarurni savi nirnik ammalu qanuili jjuti gijanginnik tamatumunga ilag uttiuti jau simajumut akilirsu innariaq aruti nginnut tamanna pitaqariaqa laurmat kiinaujanik ammalu aulatti utinginnik titirarvi ngmi
Ref | i wonder if the minister can give us some background and rationale for this additional expenditure that is required for finance and administration
Hyp | i think the minister can already they give many many of many many many and and and and this and

Uqailaut segmentation: sentence: 18000
Src | iqqanaija qati gi guma ttia ta kka asi ngi katujji qati gii t ajji ngi nngi ta ngani aturialaungin ittinni
Ref | i would like to work closely together with the other organizations before we come with a different approach
Hyp | i want like work to together together organizations organizations organizations organizations orga nizations organizations not not different different have have _EOS

Self-defined segmentation:
sentence: 45000
Src | isuma juunga minisita utjuq tunit liq juu naq gik uva tut tit gik ila nguq nnik uattiaruq niq saq viniq niq ammalu qanuit liq jjut gik jaq ngau nnik tamatu miik nguq ila uti tit tit jaq sima jumaut aki liri suq naq giaq pillaq guq ngau ut gik tamanna pitat pillaq giaq qalauq miat kiinaujanik ammalu aulat liuq tit ngau nnik

	Epoch	Perplexity	BLEU scores
Baseline	21	1341.3691	6.43
BPE	43	146.3434	11.01
Self-defined	50	194.2248	12.80
Uqailaut	50	138.4691	13.89

Table 2: Best reported perplexity/BLEU trade-off for each model (Full DEV SET) .

titirarvik miik
Ref | i wonder if the minister can give us some background and rationale for this additional expenditure that is required for finance and administration
Hyp | i wonder that the minister is provide us to to some and and and included and into and and and

file for BRP : translation_brp.txt, Uqailaut : translation_uqi.txt, self defined morphine analizer : translated_self_defined.txt.

4 Discussion and future work ideas

Probably a discussion about why some model is better than the other won't be very productive for now, as the values for the three models are fairly similar. A more interesting topic could be why the have these similar values even when they have segmented the data to different number of types. One of the evident limitations of this study is that we weren't able of deepening on some qualitative results of the segmentation, besides the purely quantitative ones.

An important matter in informatics is efficiency, in this sense the preprocessing based on the Uqaluit morpheme analyzer Java segmentator seems to be highly inefficient, but as we already noted, the model using its preprocessing method was only using less of half of the data and the slow running is because our lack of a better implementation that could have been done with more time. On the same lines, the advantage of using the self defined segmentator and the BPE based one is that they run faster the Uqailaut project based, nevertheless in these cases given that not all Inuktitut words are available in the provided lists, the algorithms won't segment rare words correctly. Another disadvantage we need to mention is that words that should not be segmented (like proper names) are sometimes segmented if there is a matching segmentation pattern in the dictionary; For example, in sentence 1. *Hansard* is segmented into *Hat saqrd*, because "hat" and "saqrd" are available in our dictionary, while "Hansard" is not. The same happens in line 8 with *sitamiq* (which mean Thursday) which is segmented into *sitaq miq*.

One qualitative comment we can do, and that results evident when looking at the translations, is that even using these methods a recurrent error in the translation is the repetition of the same word a lot of times and that translation hypotheses length don't fit with the length of the real translations. One of the matters that we couldn't undertake because of time limitations was this problem, and we think this could be an interesting research problem to investigate.

A final topic we will mention is that these methods, by reducing the difference between the type/token ratio for both languages, are doing something that is very intuitive: after all, meaning represent cognitive distinctions, and we should expect of these to be quite similar across all languages. Being like this, maybe as we can reduce the number of polysynthetic types by splitting them in smaller units, we could expect of isolated languages to have meanings represented by sequences of words that are fixed in form (like *kick the bucket* or other idioms). This will surely be a matter of future research, as well as the need of use different preprocessing methods for different linguistic system types.

References

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. In *ICLR 2015*.

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Kalchbrenner, N. and Blunsom, P. (2013). Recurrent continuous translation models. Seattle. Association for Computational Linguistics.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 48–54, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

5 Appendix

Algorithm 1 Self-defined Segmentation

```
procedure SentenceSegmentation(sentence)
    for each word in sentence do
        segmented words = segmentWord(word)
    return segmented words

procedure segmentWord(word)
    while there is character in word do
        segment,rest, melted char = segmenting the word based on suffix and root dict
        if if word is segmented then
            if (last segmented word + melted char) in allSuffixes dict then
                last segment words += melted char
                push segment into segmented words
        else
            word=word[1:]
    return segmented words

procedure Segment word based on dictionary(word)
    last_idx = length of word
    while last_idx > 0 do
        token, after_token = word[:last_idx], word[last_idx:]
        if token in all suffixes_dict then
            return suffixes, melted_char = word_dict[token]
        return word
        last_idx-=1
```

Algorithm 1 Learn BPE operations

```
import re, collections

def get_stats(vocab):
  pairs = collections.defaultdict(int)
  for word, freq in vocab.items():
    symbols = word.split()
    for i in range(len(symbols)-1):
      pairs[symbols[i],symbols[i+1]] += freq
  return pairs

def merge_vocab(pair, v_in):
  v_out = {}
  bigram = re.escape(' '.join(pair))
  p = re.compile(r'(?<!\S)' + bigram + r'(?!\S)')
  for word in v_in:
    w_out = p.sub(''.join(pair), word)
    v_out[w_out] = v_in[word]
  return v_out

vocab = {'l o w </w>' : 5, 'l o w e r </w>' : 2,
         'n e w e s t </w>':6, 'w i d e s t </w>':3}
num_merges = 10
for i in range(num_merges):
  pairs = get_stats(vocab)
  best = max(pairs, key=pairs.get)
  vocab = merge_vocab(best, vocab)
  print(best)
```

r ·	$\longrightarrow$	r·
l o	$\longrightarrow$	lo
lo w	$\longrightarrow$	low
e r·	$\longrightarrow$	er·

Figure 1: Example of BPE algorithm (Sennrich et al., 2016)

Predicted sentence for self-defined Morpheme Segmentation

---

sentence: 45000
Src | isuma juunga minisita utjuq tunit liq juu naq gik uva tut tit gik ila nguq nnik uattiaruq niq saq viniq niq ammalu qanuit liq jjut gik jaq ngau nnik tamatu miik nguq ila uti tit tit jaq sima jumaut aki liri suq naq giaq pillaq guq ngau ut gik tamanna pitat pillaq giaq qalauq miat kiinaujanik ammalu aulat liuq tit ngau nnik titirarvik miik
Ref | i wonder if the minister can give us some background and rationale for this additional expenditure that is required for finance and administration
Hyp | i wonder that the minister is provide us to to some and and and included and into and and and

---

precision | 0.5000
recall | 0.4348

---

sentence: 45001
Src | qujannamiiq guq juu tiqt
Ref | thank you
Hyp | thank you mr ng _EOS

---

precision | 0.5000
recall | 1.0000

---

sentence: 45002
Src | iksiva taq tusaajikkut
Ref | chairperson interpretation
Hyp | chairperson interpretation _EOS

---

precision | 1.0000
recall | 1.0000

---

sentence: 45003
Src | qujana mi nguq juutit ut raq jaq
Ref | thank you mr o brien
Hyp | thank you mr o brien _EOS

---

precision | 1.0000
recall | 1.0000

---

sentence: 45104
Src | iksiva taqa qu apiri nasuk juqnga marruut sima liri mi tit u kaat tut laaq siq qattaq sima suuq
Ref | mr chairman i guess i m asking there s been two different contractors being used
Hyp | mr chairman i am asking in this question out a a service mr north north _EOS

---

precision | 0.2667
recall | 0.2667

Predicted sentence for Uqilaut

English predictions, s=18000, num=10:

sentence: 18000
Src | iqqanaija qati gi guma ttia ta kka asi ngi katujji qati gii t ajji ngi nngi ta ngani aturialaungin ittinni
Ref | i would like to work closely together with the other organizations before we come with a different approach
Hyp | i want like work to together together organizations organizations organizations organizations organizations organizations not not different different have have _EOS

precision | 0.3684
recall | 0.3889

sentence: 18001
Src | amma lu taku nna ria qa ri vugut umajuit qauji saq ta u vam mata qauji saq ti nut amma lu umajulir ijinut amma lu asinginnu
Ref | we should also look at the ways the wildlife are studied by scientists and biologists and so on
Hyp | we i are and and and and _UNK and on and and and and _UNK _EOS

precision | 0.3333
recall | 0.2778

sentence: 18003
Src | qujannamii uqaqtii
Ref | thank you mr speaker
Hyp | thank you mr speaker _EOS

precision | 1.0000
recall | 1.0000

sentence: 18005
198 Src | qujannamiik mista qilavvaq
Ref | thank you mr kilabuk
Hyp | thank you mr kilabuk _EOS

precision | 1.0000
recall | 1.0000

sentence: 18007
Src | mista ikkarrialu tusaajitigut
Ref | mr iqaqrialu interpretation
Hyp | mr iqaqrialu interpretation _EOS

precision | 1.0000
recall | 1.0000

sentence: 18008
Src | qujannamii uqaqtii
Ref | thank you mr speaker
Hyp | thank you mr speaker _EOS

precision | 1.0000
recall | 1.0000

sentence: 18009
Src | isuma gi jara tanna kama gi jaria lik isuma na mmarim mat amma lu minista u juu p kiujutinga piugijara
Ref | i think this issue is very important and i like the ministers response
Hyp | i think this issue important important important important important and and ministers response response response _EOS

precision | 0.5333
recall | 0.6154
sentences matching filter = 10

Predicted sentence for BPE

English predictions, s=45000, num=10:

---

sentence: 45000
Src | isumajunga minisitaujuq tunisijunnar pa uvattinnut ilanginnik uattiarurni savi nirnik ammalu qanuili jjuti gijanginnik tamatumunga ilag uttiuti jau simajumut akilirsu innariaq aruti nginnut tamanna pitaqariaqa laurmat kiinaujanik ammalu aulatti utinginnik titirarvi ngmi
Ref | i wonder if the minister can give us some background and rationale for this additional expenditure that is required for finance and administration
Hyp | i think the minister can already they give many many of many many many and and and and this and

---

precision | 0.4000
recall | 0.3478

---

sentence: 45001
Src | qujannamiinguj uti t
Ref | thank you
Hyp | thank you mr you _EOS

---

precision | 0.5000
recall | 1.0000

---

sentence: 45002
Src | iksivautaq tusaajikkut
Ref | chairperson interpretation
Hyp | chairperson interpretation interpretation interpretation interpretation interpretation interpretation _EOS

---

199 precision | 0.2857
recall | 1.0000

---

sentence: 45003
Src | qujannangmingujutit uu purai jan

Ref | thank you mr o brien
Hyp | thank you mr o brien _EOS

---

precision | 1.0000
recall | 1.0000

---

sentence: 45008
Src | piliriangujut angijuutiit ilangi pijariiqtau laungittut tungaani airri ili 1 1 9 9 9 akiliga ksau lilauqput tamatumani arraagugi liqta ttinni
Ref | certain major projects that weren t completed prior to april 1 1999 became an expense for this current year
Hyp | the department of of is on on april april april april april april april 1999 1999 1999 in standing standing

---

precision | 0.1000
recall | 0.1053

---

sentence: 45009
Src | piluaqtumit iqqanaijaqtulirijikkut sikkiliur utit amma inuliriji kkuur uta uqattaqtu n nut
Ref | primarily in those areas of human resources and payroll systems and income support systems
Hyp | the health health of of of and social services _EOS

---

precision | 0.2222
recall | 0.1429
sentences matching filter = 10

sentence no	original	translation
1	Hansard	Hansard
2	nunavut kanata	Nunavut Canada
3	nunavut maligaliurvia	LEGISLATIVE ASSEMBLY OF NUNAVUT
4	sivuliqpaat katimaniq	1st Session
5	sivuliqpaat maligaliurvik	1st Assembly
6	maligaliuqtiit katimautigisimajangitta	HANSARD
7	titiraqsimaningit	Official Report
8	sitamiq, ipuru 1, 1999	THURSDAY, APRIL 1, 1999
9	nunavut maligaliurvia	Legislative Assembly of Nunavut
10	maligaliurtiit	Members of the Legislative Assembly

Table 3: source and target sentence.

no sentence	self-defined Morphine segmentation	Uqailaut segmentation	BPE
1	Hat saqrd	Hansard	han sard
2	nunavut kanata	nunavut kanata	nunavut kanata
3	nunavut maligaliurvik	nunavut maligaliurvi a	nunavut maligaliurvia
4	sivu liq paaq katit mi niq	sivu liq paat kati ma niq	sivuliqpaat katimaniq
5	sivu liq paaq maligaliurvi	sivu liq paat maligaliurvik	sivuliqpaat maligaliurvik
6	maligat liuq tiqit katit mi tit gik sima jaq ngau tuq	maliga liuq ti it kati ma uti gi sima ja ngit ta	maligaliuqtiit katimautigisimajangitta
7	titiraq sima niq ngau	titiraq sima ni ngit	titiraqsimaningit
8	sitaq miq, ipuk 1, 1999	sitamiq, ipuru 1, 1999	sita miq ipuru 1 1 9 9 9
9	nunavut maligaliurvik	nunavut maligaliurvi a	nunavut maligaliurvia
10	maligat liuq gik it	maliga liur ti it	maligaliurtiit

Table 4: Segmented words.

Epoch	Perplexity	BLEU
1	65.74173914104252	1.2164087976074296
5	34.877712978583546	3.50707973832117
10	34.47087712305048	4.118952372420291
15	38.306090407176825	6.224426898946241
20	47.280715603730066	8.580795427120613
25	55.22199641368928	10.341212463850422
30	60.116457857999	10.2088787268695
35	64.03285727104758,	11.596703972189113
40	76.5174937549621	12.131409964160643
45	85.65961149685397	12.850738818060492
50	87.6732553115107	14.01207278634293

Table 5: BLEU and perplexity using self-defined MA.

Epoch	Perplexity	BLEU
1	74.27234364753733e	0.00
5	53.193082849081684	4.22
10	39.659788672130745	5.00
15	48.90469051559858	6.02
20	44.198409188391814	8.057643435647051
25	50.055328247287775	9.41
30	47.27672810474158	12.528944349322662
35	54.22278004150089	15.467546776512359
40	69.08218701863932	17.73395807471084

Table 6: BLEU and perplexity using Uqilaut.

Epoch	Perplexity	BLEU
1	79.50819333312742	1.6871372362964394
5	46.0141672889105	2.738530699054858
10	49.498279266828774	3.977359709670532
15	56.24593987941152	4.090657730428738
20	72.2894064487986	4.8232111495314225
25	83.88830037070403	7.812653214050136
30	109.96170715020641	10.737135797515265
35	122.36330519847414	13.849140938694312
40	149.11996233068768	13.90269869745394
43	149.11996233068768	14.45502313053065

Table 7: BLEU and perplexity using BPE.

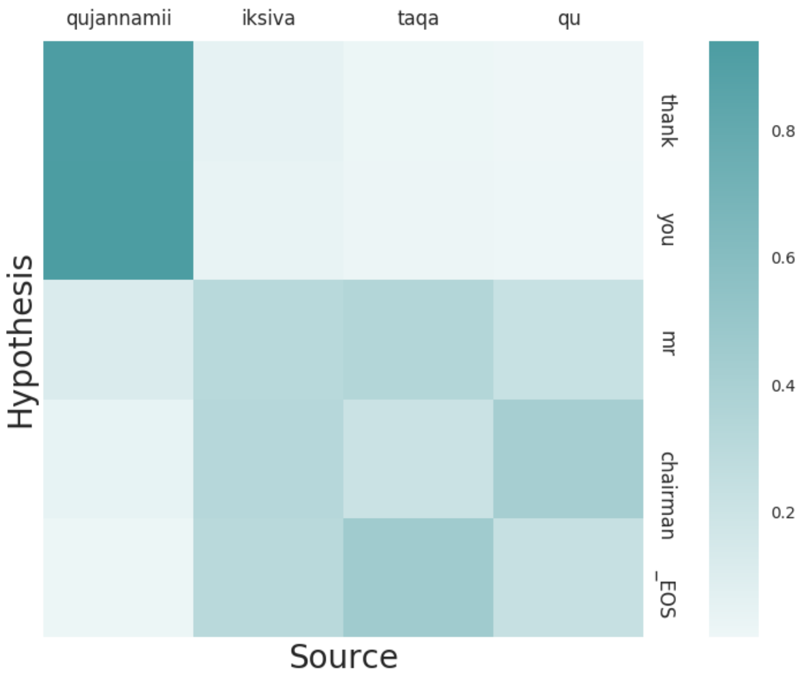

Figure 2: self-defined attention

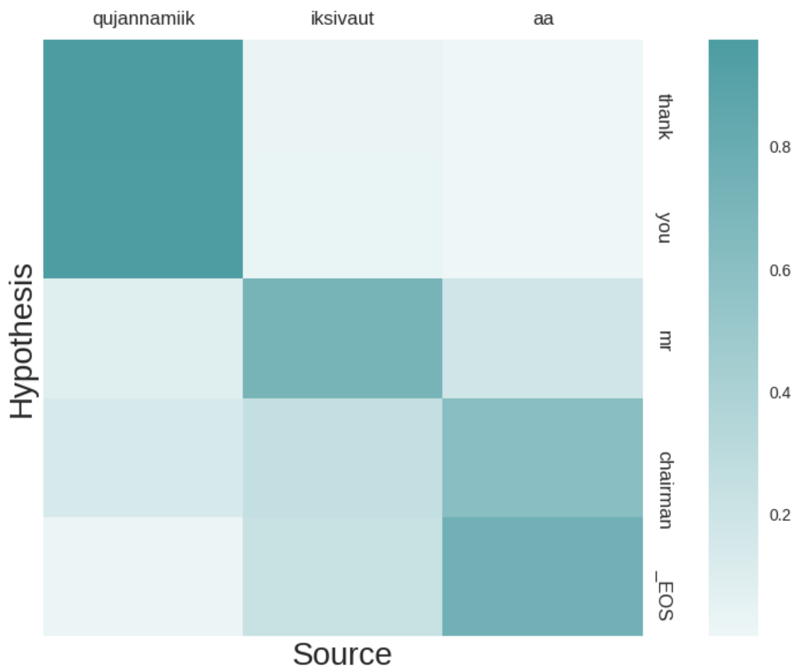

Figure 3: BPE attention

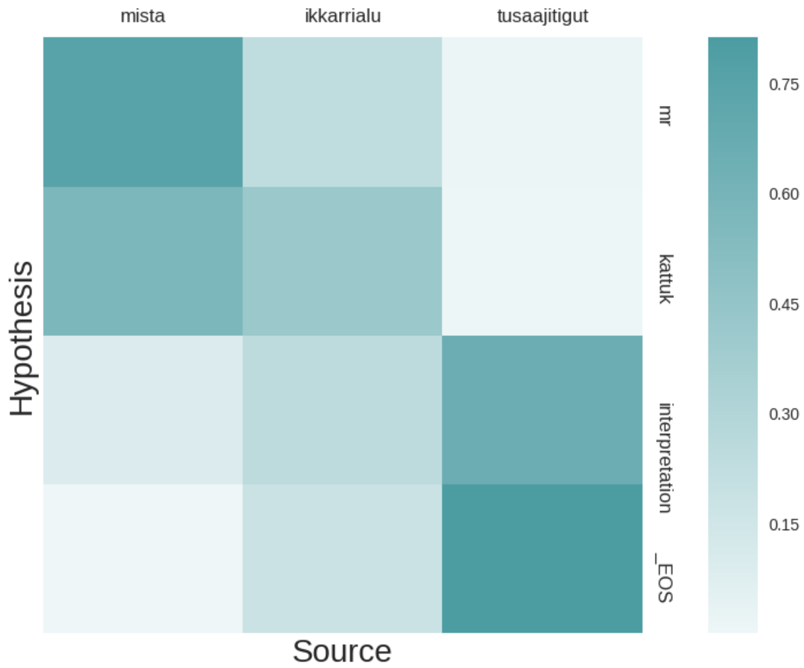

Figure 4: Uqilaut attention

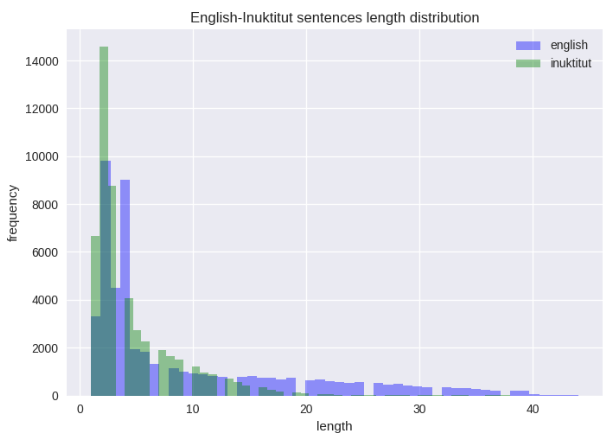

Figure 5: Original English-Inuktitut Distribution

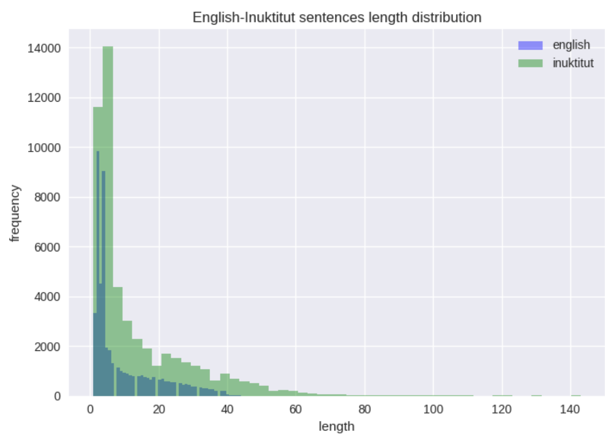

Figure 6: Self defined English-Inuktitut Distribution

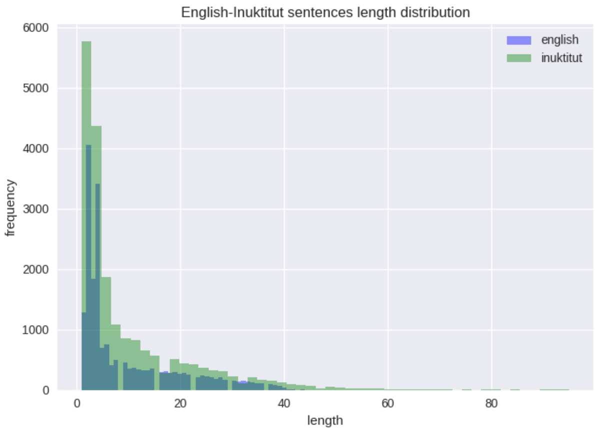

Figure 7: Uqilaut English-Inuktitut Distribution

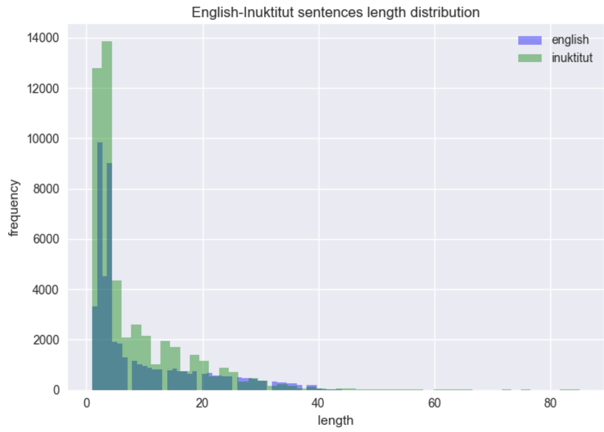

Figure 8: BPE English-Inuktitut Distribution

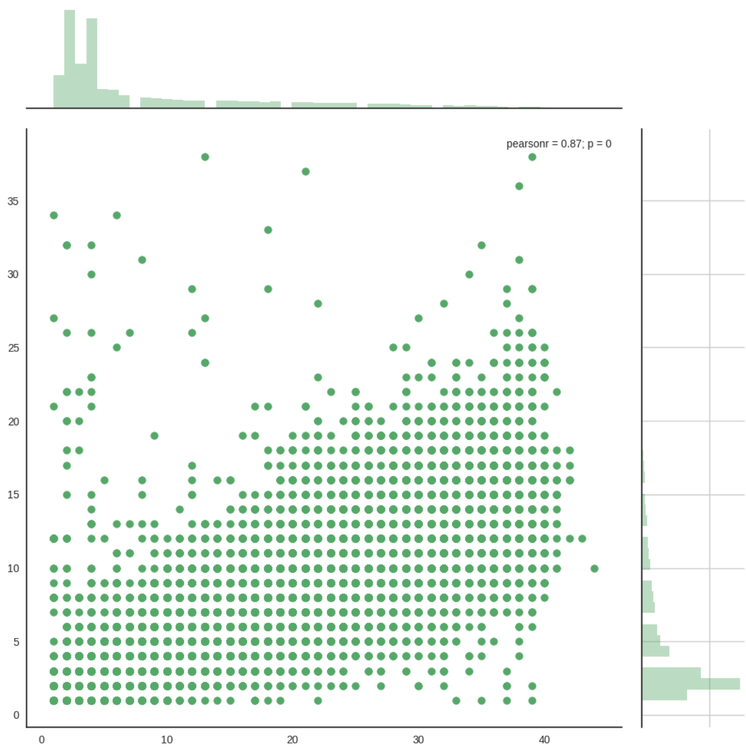

Figure 9: Original English inuktitut correlation

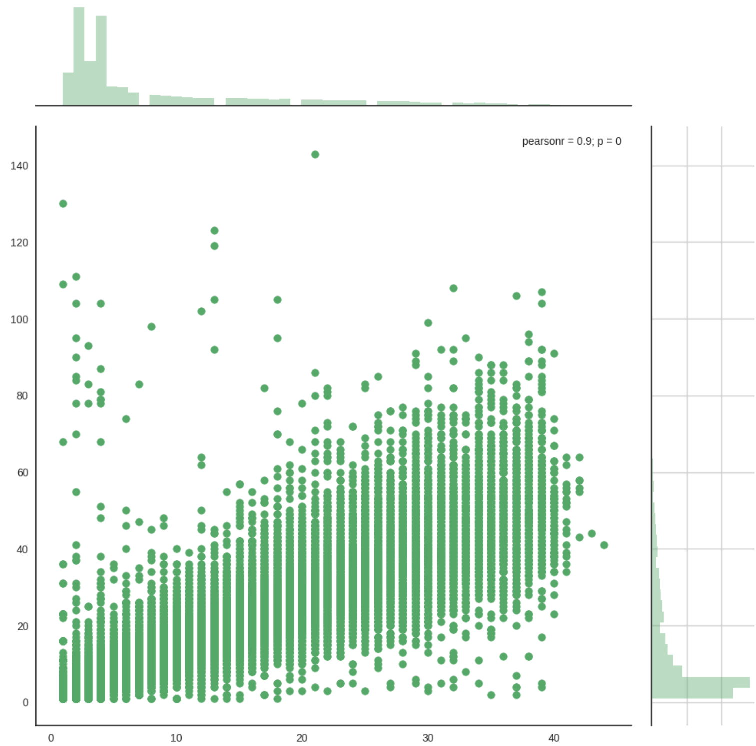

Figure 10: Self defined MA English inuktitut correlation

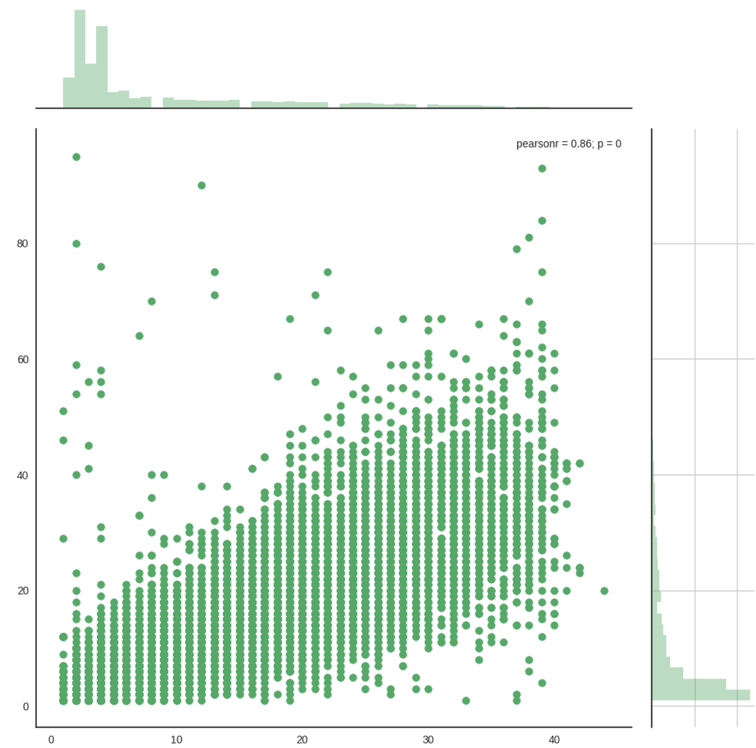

Figure 11: Uqilaut English inuktitut correlation

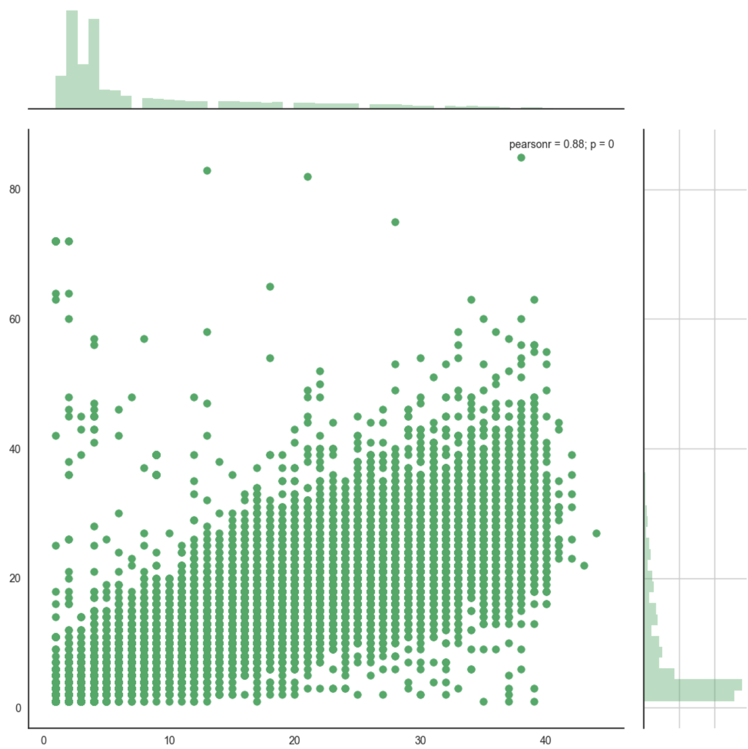

Figure 12: BPE English inuktitut correlation

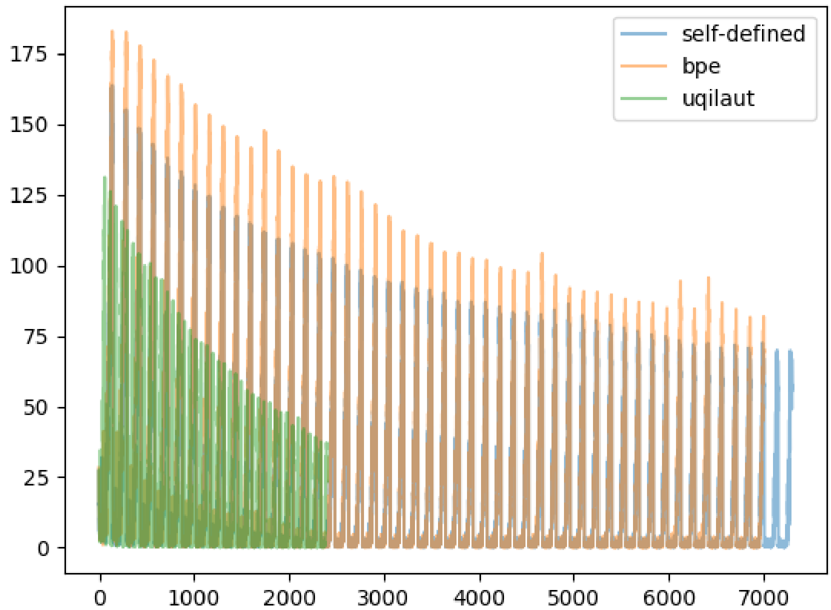

Figure 13: Loss value training