Machine Translation Assignment 3: Reranking

Q1

Since we are using argmax over the sum of weighted features to find the best translation, changing the sign on one feature weight means that we are looking for the maximum of its inverse. Instead of selecting $\hat{e}$ such that maximizes a linear combination of feature values, we select one which effectively maximizes two of the features and minimizes one. If the weight of language model log likelihood is set to -1, we instruct the reranker to favour the least fluent, rather than the most, English sentences. In principle we can expect the quality of the top translation to decrease. However, the magnitude of the effect depends on considerations such as how good the language model is or how big the variation in p(e) is between the top 100 decoder outputs.

Q2

Subjectively, the translations chosen with reversed language model weight are not obviously worse, or even less fluent. They do, however seem to be longer. Since the longer the sentence the smaller the language model probability it is to be expected that minimizing this feature value will favour longer outputs. For instance, the first example in table 2 shows sentences which arguably do not differ in content, although the second uses 10 words for what the first expresses in 3..
It seems that sometimes the preference for shortness leads to loss of information, e.g. in the second example. While both are good English, the second one is closer to the reference translation which includes the phrase *"this summer's tragedy"*. In this example by reversing the p(e) weight we actually choose a better candidate. That the default reranker leaves information out is evident from comparison of sentences in the third example. The default translation simply does not include the adjectives modifying *language* and *question.*
In some cases reversing the bias for language model probability leads to choosing more comprehensible translations, as seen in the next example in table 2.
There are examples of LM likelihood value actually tracking what we'd like it to track, namely English fluency and idiomaticity. The last example in this section of table 2 shows an idiomatic expression *keep in mind* being dealt with correctly by the default, but not the LM-flipped, reranker.

Q3

Table 1 and figure 1 illustrate the BLEU scores achieved by rerankers using a variety of feature combinations.
The first thing to note is that the best BLEU score model is achieved when all three features are used, which suggests that all of them contribute useful information. We can approach the question of relative importance of the features from different angles.

- Which feature provides the best performance when used alone? : TM likelihood.
- Exclusion of which feature causes the largest performance decrease relative to the default system? : LM likelihood
- Reversal of which feature causes the largest performance decrease relative to the default system? : LM likelihood

The influence of particular features on the BLEU score is quite opaque. There is no single most informative feature. Interestingly, combining TM likelihood with one other feature has a negative influence on the score, but combining it with both brings the score to maximum.

Q4

In comparison to the default reranker, the oracle translations tend to have better BLEU scores. The oracle outputs are intuitively better, but there are still types of Russian constructions which are not dealt with well. Let's look at the first example in table 2.
The oracle translation is more precise and accounts for the English distinction between *house* and *home* by choosing the phrase *residents of the house*, rather than *home*. However, the preferred preposition of *the stairwell* was correctly chosen by the default reranker and not by the oracle. Similarly in the second and third example, the oracle provides a more fluent translation, being able to create a fully formed sentence made up of complex clause types - relatively rare or infrequent syntax.
However, aside from these particular improvements, there are many sentences that are still not good translations. These occur particularly with sentences with rare special characters, as in *"including an option to buy 20 more planes , total volume , in fact , the $ 26 billion ."*.
Further, Russian conventions for comparative constructions are retained in the translation, resulting in poorly-formed English sentences, e.g. *"but residents complaining of the dirt and noise , in many communities is becoming more and more ."*
Similarly, longer and more complex multi-clausal sentences tend to be problematic: *"all four birds used the wire to make and used hooks hooks , to a bucket for pen and pull him out of the cylinder"*
Since the oracle chooses translations out of the 100 provided candidates so that the BLUE score for the whole set is maximized, the shortcomings of oracle's outputs likely reflect the shortcomings of the decoder, translation models, and language model used to produce the translations. For instance, no good translation ca be found among the top 100 for the sentence describing the $26 billion plane-buying event. Alternatively, we might interpret the badness of the chosen translations as indicating that the BLUE metric is not that well aligned with actual translation quality. An illustration of this point is offered by example four in table 2. Even though the choice of the default reranker is clearly better, its BLUE score is lower.
To decide how much blame to ascribe to each explanation we can subjectively pick the best candidate translations and see in how many cases does the oracle agree with our choice.

Q5

A grid search through all the three features (with weights between -5 and 5) was carried out. The highest BLEU score (28.16, 0.81 higher than default) was achieved with the following configuration of feature weights:

p(e) : 2, p(e—f) : 3, p_lex(f—e) : 5

Qualitatively, the translations chosen with the optimal parameter configuration are generally of better fluency and accuracy. We observe that the translations are more verbose, which is a good thing, given the previously mentioned tendency of the default model to favour too short sentences. Limiting the relative weight of the English language model feature lets the intended meaning of the Russian sentence to be preserved. This is seen in the first example in section *Best parameters* in table 2. The optimal configuration allows the adverb *latest* to be preserved in translation, whereas the default reranker favours an adverb-less version, probably due to it being shorter and having a higher likelihood according to the language model.
The second example shows how a more precise translation is created in the following case as a more fluent syntax is produced by allowing larger but less frequent clause types.
However, it does happen that the default reranker chooses a better translation. We noticed that in many cases the outputs of the two rerankers differ only in one synonymous word or phrase. In some cases optimal parameter setting leads to choosing a less English-like paraphrased equivalent of the default translation. This could be a closer match to the original Russian, caused by the heavy weighting of $p(e|f)$ and $p\_lex(f|e)$, and still reasonably good English due to being in a top 100 candidates. The final example in table 2 shows the difference an an example of an inappropriate,

although semantically justifiable, verb choice.

Q6

When other parameters are kept at 1, the best BLEU (28.57) is achieved with length feature weighted by 2.4. When the optimum weights from Q5 are used, the best choice for the length weight is 4.1, with BLEU 28.50.
However, when features are added, optimum parameter values are bound to change. According to our modified reranker from Q7, good settings of the four parameters are: p(e) : 2 p(e—
The lengths of the chosen translations, both when other parameters were set to 1 and when they had their optimum values, are quite close to reference: 11171 and 11187 words as compared to the reference length of 11280.

Q7

We decided to implement the Minimum Error Training Rate algorithm for choosing optimal model parameter values, described in **?**. While with only 3 (or 4, counting translation length) features tuning the weights by hand is feasible, in principle the number of features could be much larger. Since we want to set the model weights so that the best translations achieve the highest model scores, we need a criterion for goodness of translation, aka an error function. In essence we want to tune the weights so that highest scoring translations contain the least errors as compared to reference. We chose the use BLEU as the error function. We apply the METR implementation to learn the optimal parameters on the training data, and supply these parameters to the reranker for use on the test data. For the MERT algorithm itself two variables are important: the initial parameter values and the stopping criterion. We start at a point whose coordinates are randomly chosen from the range -10 to 10. At each iteration we try optimising each of the parameters and change the one whose optimisation improves BLEU the most. With a fixed iteration order over parameters we faced situations where no improvement in BLEU could be obtained by optimising one parameter, and the algorithm stopped, even though improvements would be achieved if other parameters were tried. Our strategy avoids that, however we recognize that it would not be prudent in case of the parameters being more numerous.
When training and testing on the development set, the best BLEU score achieved was 28.35, which constitutes an improvement of 0.19 compared to when parameters were picked by hand. The algorithm converges in less than 10 iterations. However, due to random initialisation of parameter values, sometime the combination yielding the best BLEU score is not found. Therefore, our reranker involves 10 runs of MERT algorithm, out of which we choose the best.
When trained on the training set and tested on development set, our reranker achieves 27.41 BLEU score.
When reranker is trained on the training set extended with the length feature the BLUE score is 28.44.
The command to run our reranker is

```
python my_rerank −t <training data> −r <training references> −k <test data>
```

References

Helsgaun, K. (2006). An effective implementation of k-opt moves for the lin-kernighan tsp heuristic. *Datalogiske Skrifter (Writings on Computer Science)*, 109.

Zaslavskiy, M., Dymetman, M., and Cancedda, N. (2009). Phrase-based statistical machine translation as a traveling salesman problem. *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, 1:333–341.

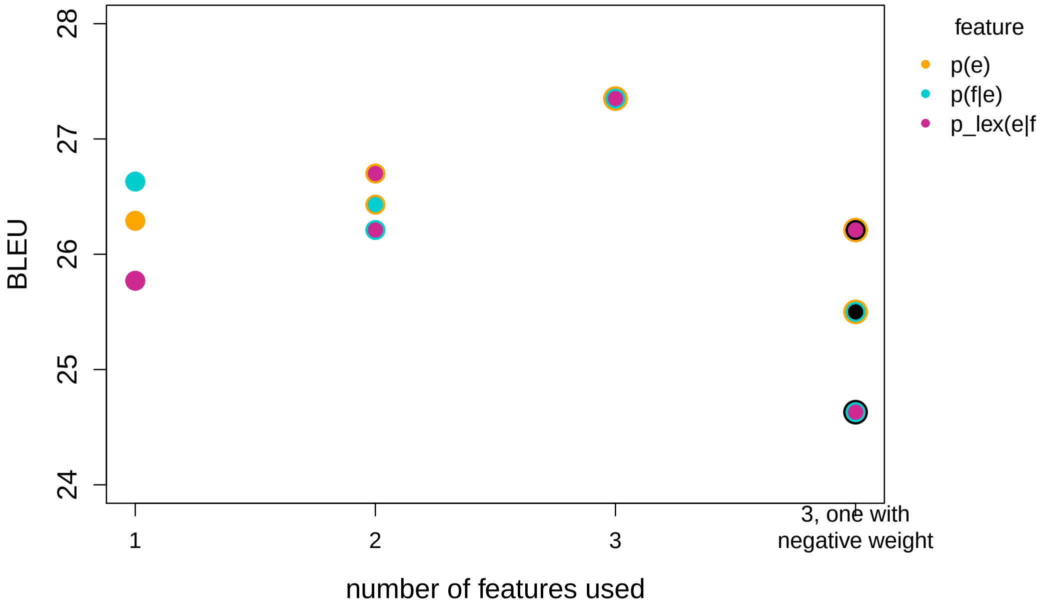

Figure 1: BLEU score as dependent on candidate translation features used for reranking.

p(e)	p(f—e)	p_lex(e—f)	BLEU
1	1	1	27.35
manipulating p(e)			
1	0	0	26.29
0	1	1	26.21
-1	1	1	24.63
-1	0	0	24.19
manipulating p(f—e)			
0	1	0	26.63
1	0	1	26.70
1	-1	1	26.21
0	-1	0	25.58
manipulating p_lex(e—f)			
0	0	1	25.77
1	1	0	26.43
1	1	-1	25.50
0	0	-1	24.84

Table 1: BLUE scores

Default	Flipped LM
”cold and inhuman”: anders breivik first publicly trial	”cold and inhuman”: anders fogh breivik publicly appears before by the court for the first time”
after the summer the researchers wanted to learn more about these people.	after summer of the tragedy the researchers want to learn more about the these people.
someone can use the ’ language ’ when answering a question : ” well... the truth... how do i know... as far as i know.”	somebody can use the ’ qualification language ’, when responds to a difficult question :” well... the truth... as far as i know... as far as i know.”
” this paper tiger, the army barracks, buildings and bombs without enough trained soldiers, to accomplish the mission,” panetta said in his introductory remarks at the pentagon.	”it ’s a paper tiger, the army with the barracks, buildings and bombs without enough trained soldiers, who can accomplish the mission,”panetta said in their introductory remarks in the pentagon.
you need to keep in mind 5,000 these concepts, ideas - them all together.	you need to keep in the head of the 5,000 of ideas - these concepts, combining them all together...
Default	Oracle
the police statement reported that the home noticed smoke in the apartment stairwell and...	police in a statement reported that residents of the house noticed smoke on the apartment stairwell and...
bella in the fourth finally able to marry her lover	bella in a fourth part finally manages to marry her beloved.
before, if you want to interview people from the british national party, would be very difficult to	previously, if you wanted to interview people from the british national party, it would be very difficult
two 19 - year - old tried to help young people, but were immediately beaten by four men.	two 19 - year - olds young people trying to help, but were just beaten by the four men.
Deafult	Best parameters
the training took place late on monday on...	the latest training took place late on monday on...
will new measures supported the council on foreign affairs or lawmakers as the june bill ?...	whether the new measures will receive full support of the council on foreign affairs or legis...
facebook also draws all kinds of data about its users.	facebook also constantly pulls all sorts of data about its users.

Table 2: Comparison of outputs of the reranker with different configurations.