Machine Translation Assignment 3: Reranking

Q1

We are using argmax over the sum of weighted features, so changing the sign on one weight means looking for the maximum of its inverse. We select $\hat{e}$ which effectively maximizes two of the features and minimizes one. In this case minimizing p(e) means favouring least fluent English sentences. In principle we can expect the quality of the top translation to decrease. However, 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 p(e)=-1 are neither obviously worse or 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 p(e) 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 flipped-LM one is closer to the reference translation which includes the phrase *"this summer's tragedy"*. Similarily, in the third example the default translation simply omits 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 rearnker

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.
The influence of particular features on the BLEU score is quite opaque. There is no single most informative feature. The best one to use in isolation is p(f|e); the largest BLEU decrease is observed when p(e) is eliminated or reversed. 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

The oracle translations tend to have better BLEU scores. The oracle outputs are intuitively better, but not without exceptions. Let's look at the first example in section *Oracle* of table 2.
The oracle translation, on top of from not omitting *the residents*, accounts for the English distinction between *house* and *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.
However, there are still many less-than-good translations. These occur inter alia when sentences have rare special characters (example five); Russian conventions for comparatives are retained, resulting in poorly-formed English sentences (example six); sentence is longer, with more complex clausal structure is (example seven).
Since the oracle's aim it to maximize BLEU by choosing from the provided top candidates, the above shortcomings likely reflect both the deficiencies of the decoder (and statistical models) used

to produce the candidates and of BLEU as a metric of translation quality. Sometimes there are no good translations among the candidates, as for the sentence about 20 planes. In other cases, illustrated by example four in table 2, even though the choice of the default reranker is clearly better, its BLUE score is lower.

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:

$$\text{p(e)} : 2,\ \text{p(e|f)} : 3,\ \text{p\_lex(f|e)} : 5$$

Qualitatively, the translations chosen with the optimal parameter configuration are generally of better fluency and accuracy. They are also more verbose, which is a good thing, given the tendency of the default model to favour too short sentences. Limiting the relative weight of the p(e) lets the intended Russian meaning to be preserved. This is seen in the first example in section *Best parameters* in table 2, where the default reranker drops the adverb *latest*.
The second example shows how a more precise translation is created as a more fluent syntax is produced by allowing larger but less frequent clause types.
In many cases default and optimal reranker output differ in one choice between synonymous words or phrases. Sometimes the default reranker picks a better synonym, as in the final example in table 2. Potentially the heavy weighting of p(e|f) and p_lex(f|e) encourages a closer match to original Russian, at a cost of choosing a word less appropriate in the English context.

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.
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.
When features are added, optimum parameter values are bound to change. Using one of the good parameter combinations found by our modified reranker from Q7 the length of the translations is 11273, even closer to the reference.

Q7

We decided to implement the Minimum Error Training Rate algorithm for choosing optimal model parameter values, described in Och (2003). The aim of weight tuning is to make the best translations achieve the highest model scores. In this assignment we use BLEU as the measure of translation goodness. While with only 4 features hand-tuning is feasible, in principle the number could be much larger, requiring a principled method of searching through the parameter space. We apply MERT to learn the optimal weights of p(e), p(e|f), p_lex(f|e), and len features from the training data. These weights are supplied to the reranker for use on the test data.
For the MERT algorithm itself two variables are important: the stopping criterion and initial parameter values. We stop when improvement in BLEU from previous iteration is less than 0.0001. 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. This strategy avoids premature stopping, however we recognize that it would not necessarily be prudent in case of a large number of parameters. The algorithm converges in less than 10 iterations. However, due to random initialisation of parameter values, sometimes the

combination yielding the best BLEU score is not found. Therefore, our reranker involves 10 runs of MERT, out of which we choose the best one.
Table 3 shows how our MERT compares to other parameter-setting methods. Our reranker trained on the training set extended with length feature scored 28.67 on the development set, an improvement of 1.32 from the baseline.
Our modified reranker should be run in a directory in which reranker is saved. The command is

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

Training and dev+test files extended with len feature have been submitted.

References

Och, F. (2003). Minimum error rate training in statistical machine translation. *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, 1:160–167.

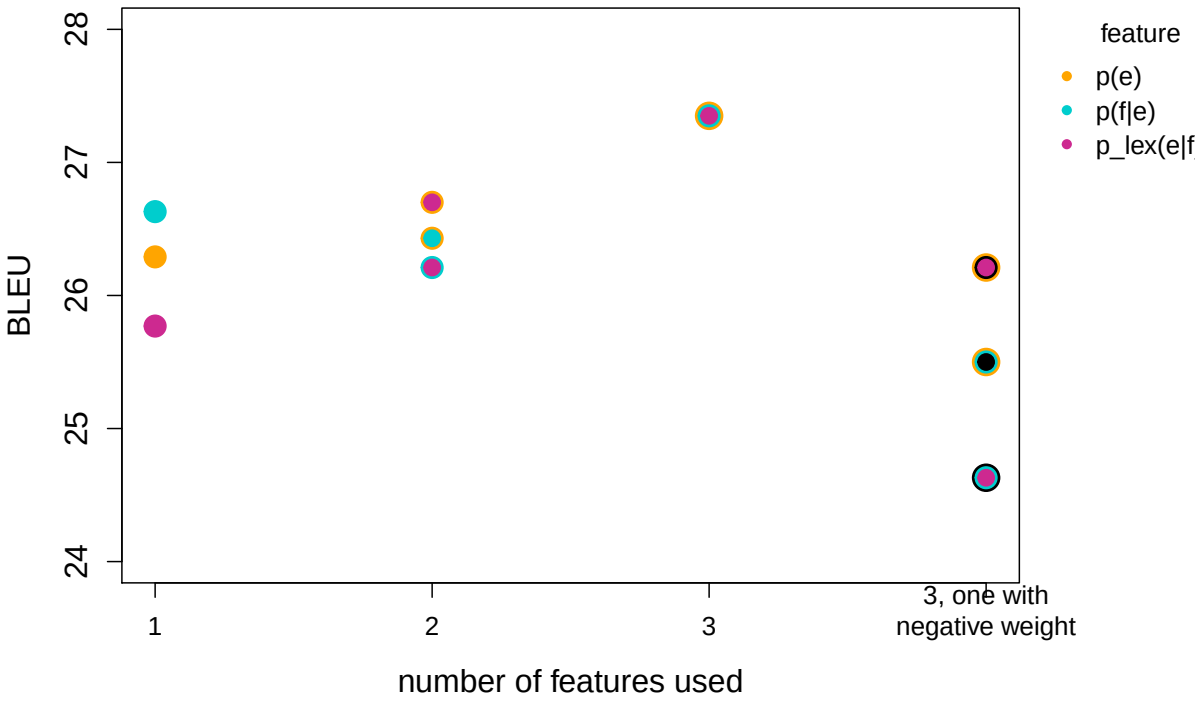

Figure 1: BLEU score as dependent on candidate translation features used for reranking.

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 achieved with different combinations of feature weights.

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.
	including an option to buy 20 more planes, total volume, in fact, the $26 billion.
	but residents complaining of the dirt and noise, in many communities is becoming more and more.
	all four birds used the wire to make and used hooks hooks, to a bucket for pen and pull him out of the cylinder

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.

no. of features	**parameter setting method**	**training set**	**testing set**	**BLEU**
3	default	-	dev	27.35
3	hand-picked	-	dev	28.16
3	MERT	dev	dev	28.25
3	MERT	train	dev	27.41
4	default	-	dev	27.99
4	hand-picked	-	dev	28.57
4	MERT	dev	dev	29.09
4	MERT	train	dev	28.67

Table 3: BLEU scores as dependent on method of feature weight setting and number of features used (4: with len feature included)