Statistical Machine Translation

David Barbas Rebollo

Automatic Translation
January, 2021

1 Introduction

The objective of this laboratory assignment is to use Moses[1], a statistical machine translation toolkit, to train models for automatic translation with pairs of bilingual phrases. The results are then evaluated using the BLEU[2] metric.

2 Experiments

The are seven different tasks in this lab assignment which involve modifying different parameters and comparing its effect on the performance with respect to the initial model, the one develop during the laboratory class.

2.1 Task 1

This exercise compares if adjusting the weights of the model with MERT affects its performance.

Table 1: Results comparision of task 1, using or not weight adjustment.

Experiment	Weights Adj.	Max It.	N-grams	Discount	Monotone	BLEU
Initial	MERT	5	Tri-grams	Kneser-ney	No	91.97
Exp1	None	5	Tri-grams	Kneser-ney	No	88.42

As expected, a higher performance is achieved adjusting the model's weights with MERT, via a log-linear adjustment.

2.2 Task 2

In the second task, the number of maximum iteracions of MERT is gradually increased.

Table 2: Results comparision of task 2, varying the number of maximum iterations.

Experiment	Weights Adj.	Max It.	N-grams	Discount	Monotone	BLEU
Initial	MERT	5	Tri-grams	Kneser-ney	No	91.97
Exp2.1	MERT	6	Tri-grams	Kneser-ney	No	92.12
Exp2.2	MERT	7	Tri-grams	Kneser-ney	No	92.05
Exp2.3	MERT	8	Tri-grams	Kneser-ney	No	91.97

The results suggest that the greater the maximum iterations, the better the adjustment of the weights would be, leading to a better performance. Despite this, the experiment with highest number of max iterations obtained the same result as the model limited at 5 iterations.

2.3 Task 3

In this task, the objective of the experiment is to vary the n-grams to see if a different size context enables to the model to achieve a better performance.

Table 3: Results comparison of task 3, varying the n-gram size.

Experiment	Weights Adj.	Max It.	N-grams	Discount	Monotone	BLEU
Exp3.1	MERT	5	Bi-grams	Kneser-ney	No	91.25
Initial	MERT	5	Tri-grams	Kneser-ney	No	91.97
Exp3.2	MERT	5	4-grams	Kneser-ney	No	91.32
Exp3.3	MERT	5	5-grams	Kneser-ney	No	90.99

Unexpectedly, every other value of n worsens the BLEU value of the models. Meaning the tri-grams is the best option for this problem. This could occur because the training corpus has not enough samples to train properly the model with 4-grams and 5-grams.

2.4 Task 4

Task 4 changes the MERT for MIRA to perform the log-linear adjustment of the weights of the model.

Table 4: Results comparison of task 4, varying the weight adjustment algorithm.

Experiment	Weights Adj.	Max It.	N-grams	Discount	Monotone	BLEU
Initial	MERT	5	Tri-grams	Kneser-ney	No	91.97
Exp4	MIRA	5	Tri-grams	Kneser-ney	No	90.81

After the evaluation we can see MIRA is not able to achieve the performance obtained with MERT.

2.5 Task 5

This experiment consists in comparing different common discount methods used to build language models. The 3 discount methods compared are Kneser-ney, Good-Turing and Witten-Bell.

Table 5: Results comparison of task 5, varying the discount method.

Experiment	Weights Adj.	Max It.	N-grams	Discount	Monotone	BLEU
Initial	MERT	5	Tri-grams	Kneser-ney	No	91.97
Exp5.1	MERT	5	Tri-grams	Good-Turing	No	91.39
Exp5.2	MERT	5	Tri-grams	Witten-Bell	No	90.11

On the table above, it is possible to observe Good-Turing, even being the simplest discount method, achieves a better performance than Witten-Bell. However, Kneser-ney obtains the best BLEU value out of the 3.

2.6 Task 6

The last experiment examines the effect on the use of a monotone alignment or not.

Table 6: Results comparison of task 6, using or not monotone alignment.

Experiment	Weights Adj.	Max It.	N-grams	Discount	Monotone	BLEU
Initial	MERT	5	Tri-grams	Kneser-ney	No	91.97
Exp6	MERT	5	Tri-grams	Kneser-ney	Yes	90.11

Being completed the experiment it is clear that a monotone alignment worsens the performce of the model. The reason behind this could be that the

English and Spanish language structure phrases differently. So a monotone aligment would not take this into consideration.

3 Conclusion

In this lab assigment, there have been serveral experiments conducted with Moses[1], not only to see how this toolkit functinos but how to experiment with different parameters and analyze their effect on a model's performance. Additionally, all experiments were done in a virtual machine, so skipping the installation of this ecosystem shrotened the time to start the experimentation.

4 Bibliografía

References

[1] Koehn, Philipp and Hoang, Hieu and Birch, Alexandra and Callison-Burch, Chris and Federico, Marcello and Bertoldi, Nicola and Cowan, Brooke and Shen, Wade and Moran, Christine and Zens, Richard and Dyer, Chris and Bojar, Ondřej and Constantin, Alex and Herbst, Evan *Moses: Open Source Toolkit for Statistical Machine Translation.* 2007.

[2] Papineni, Roukos, Ward, Zhu, W. J. . *BLEU: a method for automatic evaluation of machine translation.* ACL-2002: 40th Annual meeting of the Association for Computational Linguistics. pp. 311–318. 2002.