CS388 Notes

Zeyuan Hu
Department of Computer Science
University of Texas at Austin

Contents

1 Perplexity

In the lecture slide, the perplexity is calculated as follows given a test set $\mathbf{W} = w_1 w_2 \dots w_N$:

$$PP(W) = \sqrt[N]{\frac{1}{P(w_1 w_2 \dots w_N)}} \tag{1}$$

However, in Ray's code of P1 (i.e., `public void test(List<List<String>> sentences)`), the perplexity is calculated as:

$$PP(W) = \exp\left\{ -\frac{\sum_{d=1}^{M} \log P(W_d)}{\sum_{d=1}^{M} N_d} \right\} \tag{2}$$

Here, M represent the number of sentences in the test set and N_d represents the number of words in sentence d [1]. We can manually verify that those two formulas are equivalent [2] but equation 2 is actually what we used in the implementation.

[1] The above formula takes from [1]

[2] Use the fact that $p_1 \times p_2 \times p_3 \times p_4 = \exp(\log p_1 + \log p_2 + \log p_3 + \log p_4)$

2 Precision & Recall

Precision & recall are hard to remember about the exact formula. So, I list out two examples:

- In statistical parsing, if P is the system's parse tree and T is the human parse tree (the "gold standard"):
  - *Recall* = (# correct constituents in P)/(# constituents in T)
  - *Precision* = (# correct constituents in P)/(# constituents in P)
- In machine translation,

$$Precision = \frac{\text{\# candidate translation words (unigrams) which occur in any reference translation}}{\text{the total number of words in the candidate translation}}$$

3 Meta knowledge

3.1 Type of ambiguity in language

lexical ambiguity (word sense ambiguity).The lexical ambiguity of a word or phrase pertains to its having more than one meaning in the language to which the word belongs. ”Meaning” here refers to whatever should be captured by a good dictionary. One example would be:

1. **Consider the following joke: There are two fish in a tank. One says to the other, "How do you drive this thing?" Explain what specific type of ambiguity in language understanding makes this humorous.**

   Word sense ambiguity, first you think the sense of ”tank” is ”large container of water” and the the punch line makes you realize it could also mean ”armored military vehicle.”

co-reference ambiguity (anaphora ambiguity).

4 How do we evaluate the caption in VQA?

In "Sentence Quality Evaluation" section [2], the caption is evaluated from two perspectives:

4.1 Accuracy

An average fusion of four widely used metrics: `BLEU@N`, `METEOR`, `ROUGE-L`, `CIDEr-D`, which try to consider the accuracy of the generated sentence from different perspectives [3]:

4.1.1 `BLEU@N`

See my post [4] on BLEU [5].

4.1.2 ROUGE-L

ROUGE-L

4.2 Relevance

References

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[2] Q. Li, J. Fu, D. Yu, T. Mei, and J. Luo, "Tell-and-answer: Towards explainable visual question answering using attributes and captions," *CoRR*, vol. abs/1801.09041, 2018.

[3] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick, "Microsoft coco captions: Data collection and evaluation server," *CoRR*, vol. abs/1504.00325, 2015.

[4] "Bleu post." http://zhu45.org/posts/2018/Mar/28/bleu-a-method-for-automatic-evaluation-of-machine-translation/, 2018.

[5] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: A method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, (Stroudsburg, PA, USA), pp. 311–318, Association for Computational Linguistics, 2002.