CMPT 413/825 Project
Using ROSE metric for Machine Translation Evaluation

Luiz Peres de Oliveira lperesde@sfu.ca
Turash Mosharraf tmosharr@sfu.ca
Jin Jung sjjung@sfu.ca
Justin Lew jylew@sfu.ca

Abstract

1 Motivation

The motivation of using regression and ranking based optimisation in the machine translation evaluation problem is to improve the accuracy of the evaluation beyond the baseline implementation. The baseline implementation used the METEOR metric for finding precision and recall.

2 Approach

The approach taken to improve from the baseline is partially based on the implementation of the ROSE metric (Song and Cohn, 2011.) The algorithm used two feature vectors. The first feature vector consist of the parameters, preceded by it's ID:

- 1-4: n-gram precision, n=1...4
- 5-8: n-gram recall, n=1...4
- 9-12: n-gram f-measure, n=1...4
- 13: average n-gram precision for sentence
- 14: score sentence at Document level

The second feature vector consist of the parameters, preceded by it's ID:

- 1-4: n-gram precision for sentence excluding stopwords, n=1...4
- 5-8: n-gram recall for sentence excluding stopwords, n=1...4
- 9-12 n-gram f-measure for sentence excluding stopwords, n=1...4
- 13 average n-gram precision for sentence excluding stopwords, n=1...4

One way to compare the two translations is using n-gram precision and n-gram recall. n-gram precision is the ratio of the count of n-grams in the candidate translation sentence that is in the reference sentence to the counts of all n-grams in the candidate sentence. The n-gram precision is defined as

$$P_n = \frac{\sum_{ngram \in \vec{c}, ngram \in \vec{r}} Count(ngram)}{\sum_{ngram \in \vec{c}} Count(ngram)}$$

where $\vec{r}$, is defined as the (human) reference sentence, and $\vec{c}$, is defined as the (hypothesis) candidate sentence. Only 1,2,3,4-gram counts are used in the implementation of the algorithm.

3 Data

The data file to train the evaluation model is from hyp1-hyp2-ref. The file consists of a triple (hyp1, hyp2, and ref) where hyp1 and hyp2 are two translations to which is evaluated by the algorithm, along with a reference sentence of the hypothesis curated by a human translator.

The data file dev.answers contains the preference between the two hypothesis translations by a human translator. The numbers correspond to outputs of the function,

$$f(h_1, h_2, e) = \begin{cases} 1, & \text{if } h_1 \text{ is preferred to } h_2 \\ 0, & \text{if } h_1 \text{ is equally good/bad to } h_2 \\ \text{-1}, & \text{if } h_2 \text{ is preferred to } h_1 \end{cases}$$

where h_1 and h_2 are the two hypothesis translation and e is the reference translation.

4 Code

4.1 Pseudocode of modified ROSE metric evaluation

Data: $(hyp1, hyp2, ref)$ in $\mathcal{D}$
Result: output the function for each triple (hyp1, hyp2, ref) in $\mathcal{D}$
for $(h1, h2, e)$ *in* $\mathcal{D}$ **do**
    vc1, vc2 = [0] * 32, [0] * 32;
    sw1, sw2 = [0] * 13, [0] * 13;
    h1 = fix_input(h1);
    h2 = fix_input(h2);
    (vc1, vc2) = get_ngrams(e, h1, h2, vc1, vc2, TRUE);
    (sw1, sw2) = get_ngrams(rsw(e), rsw(h1), rsw(h2), sw1, sw2, FALSE);
    l1 = (sum(vc1[0:13]) + sum(sw1) * 1.1) + (vc1[13] * 0.4))/2.5;
    l2 = (sum(vc2[0:13]) + sum(sw2) * 1.1) + (vc1[13] * 0.4))/2.5;
    if *l1* == *l2* **then**
        print 0;
    else if $l1 < l2$ **then**
        print 1;
    else
        print -1;
    end
end

4.2 Miscellaneous algorithms used

The function $get\_ngrams(\cdot)$ is used to calculate the ratio of the n-gram counts P_n defined previously. The function $rsw(\cdot)$ outputs the sentence with all the stopwords removed. Stopwords are high frequency words in a given grammar and may be the set that contains the words: ”a”, ”I”, and etc.

5 Experimental Setup

6 Results

7 Analysis of the Results

8 Future Work

Acknowledgments