EEE 586

Statistical Foundations of Natural Language Processing Assignment 2

The Hunt for Collocations

Introduction

In this assignment, I have examined collocations in bigram form using 3 hypothesis testing methods. The methods are student's t-test, chi-square test and likelihood ratio test. In the whole work, I have used the corpus of the concatenation of seven novels of Charles Dickens: Bleak House, David Copperfield, Great Expectations, A Tale of Two Cities, Hard Times, Oliver Twist and The Pickwick Papers.

Experiments

For Part 1, I have done corpus preprocessing part. First, i have downloaded the corpus txt file and loaded it. With the help of nltk library, i have tokenized and lemmatized the corpus and find the whole number of tokens. Before lemmatization, via positional tagging function i have detected the target types and after lemmatization i have gathered the collocations with two different window size setting. First with window size 1 and next with 3. In order to gather collocation candidates, i have eliminated non noun-noun or adj-noun bigrams, bigrams with stopwords, bigrams with punctuations, and frequency less than 10.

For Part 2, i have found the collocations with three different method. I have implemented t-test, chi-squared test and likelihood ratio test for two corpora and listed the 20 greatest score for each test for each corpus in the answer sheet.

Calculations of the tests for test set "cursitor street" and "good one" are done for bigrams with window size 1. Whole critical values are for alpha=0.005. t-test score calculated as follows:

$$t = \frac{\overline{X} - \mu}{\sqrt{\frac{s^2}{N}}}$$

Figure 1: Formula used in t-test score calculation

In this calculation, sample mean resembles the p in bernoulli where p is the probability for each word to show in whole corpus. p(w1) is the probability of the first word and p(w2) is the second one. p(w1w2) is the probability of bigram. Hypothesis zero thinks they should be independent, so we multiply p(w1) with p(w2) and take difference with p(w1w2) statistically to find the t-test score.

In chi-square test:

$$X^2 = \frac{\sum_{i,j}(O_{ij} - E_{ij})^2}{E_{ij}} = \frac{N(O_{11}O_{22} - O_{12}O_{21})^2}{(O_{11} + O_{12})(O_{11} + O_{21})(O_{12} + O_{22})(O_{21} + O_{22})}$$

Figure 2: Formula used in chi-square-test score calculation

In this method we are looking for observed versus expected probability differences of the words. O terms represents observed times of the words. O11 represents both words appeared counters, O12 only word 1, O21 only word 2 and O22 none of them. This method uses marginal properties and convert into proportions.

In likelihood ratio method,

$$\log\lambda = \log\frac{L(H1)}{L(H2)} = \log\frac{b(c12, c1, p)b(c1 - c12, N - c1, p)}{b(c12, c1, p1)b(c1 - c12, N - c1, p2)}$$
$$= \log L(c12, c1, p) + \log L(c2 - c12, N - c1, p) - \log L(c12, c1, p1) -$$
$$\log L(c2 - c12, N - c1, p2)$$
$$where\, L(k, n, x) = x^k (1 - x)^{n-k}$$

Figure 3: Formula used in likelihood-ratio-test score calculation

In this last method, c12 represents encounters with both words (bigrams), c1 represents the encounters with word 1 and c2 is with word 2. These occurences are used to calculate H1 and H2 hypothesis and compared. Reported log lambda values are multiplied with -2 to test significance with chi-square score.

Conclusion

In this assignment, I have learned the importance of collocations, methods used for hunting collocations and different approaches gathered from these methods. I believe that, statistical approaches towards collocations are helpful for depthful search in machine translation problems.

References

[1] Christopher D. Manning, Hinrich Schütze 2000. Foundations of Statistical Natural Language Processing. The MIT Press Cambridge, Massachusetts London, England.

[2] Pearce, D. “Synonymy in collocation extraction.” In Proceedings of the NAACL'01 Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations. Pittsburgh, PA.