COMP90049 Project1B Report

Tessa(Hyeri) Song _ 597952

1. Introduction

1.1. The major problem and data sets involved

The major goal of this report is analysing the solution of a knowledge problem which focuses on "approximate matching method". Basically, there are two data sets available introduced by Maas et al (2011) which are a list of film titles and a number of film review text files and the task is to match each review with an appropriate title in the list using a suitable approximate matching method. The chosen title should be the one which has been evaluated as "the most relevant" to the review according to the method.

1.2. Overview of the approximate matching method chosen

The fundamental approximate matching method used is "local edit distance". The system calculates the local edit distance between the given review text and each film title in the list and select the title with the highest local edit distance value.

2. Evaluating Effectiveness

2.1 Precision

The precision has been calculated to evaluate this system with randomly chosen 50 reviews.
The correct title is represented as 'O' and the wrong as 'X'

OXXXX | OXXXX | XXOOX | XOXXX | XOXOX
XXXXO | OXXOX | XXOXX | OXXOX | XXXXO

There are 14 correct titles out of 50 so the precision is 0.28

2.2 Effectiveness

This matching system using local edit distance is more effective than typical exact matching system because it makes allowance for error range so that even if only partial title is mentioned in a review, it still can be selected. For example, the review file 13251.txt is assigned the correct title 'Dollman vs. Demonic Toys' and the crucial substring within the review which leads this matching is 'dollman, and demonic toys'. This substring and the title are not exact match but they are considered as 'good approximate match' by this system and selected successfully.

2.3 Ineffectiveness

First of all, this method is not appropriate for films with reasonably short titles. To be specific, even though this matching system is able to find out the approximate matching to the correct title in a given review, if it spots another approximate matching with a longer candidate title, it will always select the latter one since it has a higher local edit distance value. For instance, the review file 35393.txt has both approximate matchings of 'Drama Queen' and 'Village of the Damned' and the latter one has been chosen as the title. However, the actual film title is 'Drama Queen'.

Additionally, this system does not have a proper ranking system which can be applied when it has more than one candidate film title with the same value of local edit distance. For example, in the case of review file 11412.txt, even if there is an exact matching to the correct title in the review, this system fails to provide the right answer. The correct title is 'Tightrope' and the chosen one is 'Love and Pain and the Whole Damn Thing'. The problematic part leading this incorrect decision is a substring 'the whole' in the review text. With this substring, 'tightrope' and 'Love and Pain and the Whole Damn Thing' turn out to have the same local edit distance value with the given review and the system does not consider the proportion of matching substring in terms of the length of each candidate film title, which leads the wrong title to be assigned. If there was a ranking system that took the proportion into consideration somehow, the correct title could have been assigned.

Lastly, this matching system entirely relies on finding out approximate matching to candidate film title and does not consider the context at all. Therefore, if a review has other wrong film title as substring, it will probably be assigned that wrong title. This is the case for review file 1518.txt that has been matched with the title 'Sorry, wrong number' just because the reviewer mentioned it as a reference film.

3. Deciding whether a film is 'good'

Aside from the main title matching problem, the approximate matching method can also be used to manipulate the datasets in other ways.

One of the examples is deciding whether a film is good. In order to achieve this, two string arrays are required and each array contains general 'good' and 'bad' words. For instance, good words can be 'good', 'nice', 'recommendable' and bad words can be 'bad', 'waste', 'disappointed' and so on. Each review now can be evaluated as either 'positive' or 'negative' to the film by calculating local edit distance between the review itself and each word in the two arrays. If it has more points for good word array, then it is a positive review and if the film has more positive reviews than negative, the film is considered good.

One of the 'good' film according this system is 'Love me or Leave me'. There are 4 reviews assigned to this film and the system says 3 out of 4 are positive reviews since they have more positive words defined in the array. In fact, all of them are quite positive about this movie, which tells that this system does reasonably good work on this film.

However, it is hard to say that this system is always working well since it does not understand the context of sentences. For instance, 'bad' and 'good' words can be used not only to evaluate films but also simply to describe characters or stories. If a reviewer more focuses on the latter part and the genre is not cheerful, the review can be considered as 'negative' even though it is not the case.

As a result, this deciding system is not perfect and sometimes can produce completely opposite results. If the 'good' and 'bad' arrays are extended with more words, this system probably can make a better decision but it will soon confront the limit as it cannot understand the context.

4. Conclusions

In conclusion, the approximate matching method of local edit distance is reasonably useful for film title matching and deciding if it is a good film, especially compared to simple exact match. However, there are still a few weaknesses which leads the system to make incorrect decisions. Some of them can be improved by developing the approximate matching logic itself, for instance, taking consideration of proportion of matching substring, but some of them are beyond the capability of the approximate matching system such as making use of context of reviews.

References

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. (2011). Learning Word Vectors for Sentiment Analysis. The 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011).