USING TEXT SUMMARIES FOR PREDICTING RATING SCALES

Horacio Saggion
Grupo TALN
Universitat Pompeu Fabra
Horacio.saggion@upf.edu

Elena Lloret and Manuel Palomar
Grupo GPLSI
Universidad de Alicante
elloret@dlsi.ua.es
mpalomar@dlsi.ua.es

Overview

- Introduction
- Related Work
- Text Summarization Approaches
- The Rating Inference Process
- Experimental Setup
  - Dataset
  - Experiments and evaluation
  - Results
  - Discussion
- Conclusion and Future Work

Introduction

CONTEXT

- Web 2.0 ➔ emergence of new types of websites (wikis, blogs, social networks, …)
  - Users express their opinions
  - Studies show that 73%-87% of users take into account the Internet reviews before making any decision (Pang and Lee, 2008)
  - Companies are interested in monitoring public perceptions on products, services, policies, etc.
- Problem ➔ exponential and fast growth of information

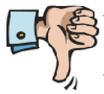
Introduction

CONTEXT

- NLP groups a series of tasks capable of providing methods and tools to efficiently deal with all this available information
  - ***Sentiment Analysis*** detects and classifies opinions in text
  - ***Text summarization*** provides a brief but accurate representation of the contents of a document
- <u>Rating Inference</u> is the task of identifying the author's textual evaluation of an entity
  - Users normally rate a service, product, ... (1=worst,...5=best)

Introduction

CONTEXT

- Rating Inference
  - Subtask of sentiment analysis
  - More challenging than binary classification (e.g. thumbs up , thumbs down )
    - Fine-grained granularity (e.g. 5-star rating)
    - With short documents the results are quite good (74% of accuracy), but not for long documents (40%)
- Text Summarization has been proven to be useful to other NLP tasks
  - Information Retrieval
  - Question Answering
  - Text Classification

Introduction

OBJECTIVES

- Preliminary analysis of the usefulness of different types of text summaries for the rating inference task
  - What type of summary and compression rate could be used instead of the full document for the rating inference task?
- We present a limited set of extractive summary types that have a widespread use in the NLP community

Related Work

- Previous work on Rating Inference
  - 1-4 star rating prediction for movie reviews (Pang and Lee, 2005)
  - 1-5 star rating prediction for product reviews (very short texts) (Saggion and Funk, 2009)
  - Rating of different aspects of hotels given in reviews (Baccianella et al., 2009)
  - Simplification of the task into binary classification (Devitt and Ahmad, 2007)
- To the best of our knowledge Text Summarization has not been analyzed within this task

Text Summarization

TECHNIQUES

- Techniques for detecting relevant information
  - **Term Frequency** (TF) → the most frequent words of a document are indicative of its topic
  - **Code Quantity Principle** (CQP) → the length of a noun-phrase may indicate important information
  - **Textual Entailment** (TE) → avoids incorporating redundancy in summaries
- Different combinations of techniques are analyzed
  - TF
  - TE+TF
  - CQP+TF
  - TE+CQP+TF

Text Summarization

APPROACHES

- Summarization types
  - **Generic** → the summary contains the main ideas
  - **Query-focused** → the summary contains the most important facts related to a query/topic/entity
  - **Sentiment-based** → the summary contains the most relevant subjective information of a document
- Baselines
  - ***LEAD*** → extracts the first sentences
  - ***FINAL*** → extracts the last sentences
  - ***QF*** → extracts the most similar sentences to a query
    - using SUMMA toolkit (Saggion, 2008)
  - ***SENT*** → extracts the strongest opinionated sentences

The Rating Inference Process

TOOLS

- General Archiecture for Text Engineering (GATE)
  - Tokenization and lemmatization (root of the word)
  - Part of speech tagging (pos)
  - Morphological analysis
  - Sentiment features using SentiWordnet (Esuli and Sebastiani, 2006) (sentiWN)
  - SVM
    - 10-fold cross validation
    - Features to train the full review and summaries
      - Root
      - Root+pos+sentiWN

Experimental Setup

DATASET

- Small corpus
  - 89 bank reviews rated from 1 to 5 gathered from *Ciao* Website (*http://www.ciao.co.uk*)
- Corpus statistics

# Reviews	89
Avg. Length	2,603
Max. Length	5,730
Min. length	1,491

- Class distribution

Star-rating	#Reviews	%
1-star	17	19
2-star	11	12
3-star	9	10
4-star	28	32
5-star	24	27

Experimental Setup

EXPERIMENTS & EVALUATION

- Summaries vs. Full review
- 7,120 summaries
  - 3 approaches (generic, query-focused, sentiment-based)
  - 4 combination of techniques (tf, cqp+tf, te+tf, te+cqp+tf)
  - 4 baselines (lead, final, qf, sent)
  - 5 compression rates (10%, 20%, 30%, 40%, 50%)
- Evaluation
  - Mean Squared Error (MSE)

$$MSE = \frac{\sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2}{n}$$

Experimental Setup

RESULTS

- Results using only ***root*** as feature for the rating classification

Approach	Compression Rate					
Full document		**10%**	**20%**	**30%**	**40%**	**50%**
Full review	MSE	2.59	2.59	2.59	2.59	2.59
4 star-rating	MSE	2.58	2.58	2.58	2.58	2.58
Summarization strategy						
LEAD	MSE	3.10	3.00	3.10	3.30	3.10
FINAL	MSE	2.74	3.00	**2.13**	2.64	**2.48**
QF	MSE	**2.49**	2.70	3.58	3.78	3.88
SENT	MSE	3.89	3.16	3.03	2.90	2.66
Generic-TF	MSE	3.21	**2.33**	**2.39**	**2.37**	**2.44**
Generic-TE+TF	MSE	3.39	3.23	**2.52**	**2.38**	**2.29**
Generic-CQP+TF	MSE	3.01	3.34	2.61	3.17	3.03
Generic-TE+CQP+TF	MSE	2.70	2.93	3.00	3.10	2.71
QF-TF	MSE	**2.11**	**2.19**	**2.18**	**2.46**	**2.37**
QF-TE+TF	MSE	**2.44**	**2.08**	**2.30**	**2.27**	**2.42**
QF-CQP+TF	MSE	2.83	3.00	2.70	**1.80**	**2.00**
QF-TE+CQP+TF	MSE	**2.11**	**2.51**	**2.28**	**2.40**	**2.10**
SENT-TF	MSE	2.83	**2.16**	**2.47**	**2.43**	**2.29**
SENT-TE+TF	MSE	3.20	2.80	**2.40**	2.69	2.71
SENT-CQP+TF	MSE	3.01	3.27	2.62	3.21	3.10
SENT-TE+CQP+TF	MSE	2.69	3.21	3.46	2.90	2.93

Experimental Setup

RESULTS

- Results using ***root+pos+sentiWN*** as features for the rating classification

Approach	Compression Rate					
Full document		**10%**	**20%**	**30%**	**40%**	**50%**
Full review	MSE	2.59	2.59	2.59	2.59	2.59
4 star-rating	MSE	2.58	2.58	2.58	2.58	2.58
Summarization strategy						
LEAD	MSE	2.63	2.68	2.72	2.86	2.60
FINAL	MSE	3.20	2.90	**2.16**	**2.13**	**2.50**
QF	MSE	**2.45**	3.45	2.90	2.91	2.81
SENT	MSE	3.69	2.87	**2.44**	**2.34**	**2.27**
Generic-TF	MSE	3.21	2.68	**2.19**	**2.29**	**2.33**
Generic-TE+TF	MSE	3.98	2.90	2.79	3.06	2.64
Generic-CQP+TF	MSE	3.37	3.45	3.31	3.19	2.84
Generic-TE+CQP+TF	MSE	3.02	2.87	2.83	3.18	**2.53**
QF-TF	MSE	**2.44**	**2.56**	**2.37**	2.77	**2.50**
QF-TE+TF	MSE	3.21	**2.52**	**2.26**	2.62	2.72
QF-CQP+TF	MSE	2.73	3.48	**2.51**	**2.48**	**2.27**
QF-TE+CQP+TF	MSE	2.60	2.83	**2.52**	**2.44**	**2.53**
SENT-TF	MSE	3.02	**2.47**	**2.16**	**2.12**	**2.31**
SENT-TE+TF	MSE	2.99	3.07	**2.21**	2.93	2.64
SENT-CQP+TF	MSE	3.54	2.81	**2.51**	**2.43**	2.70
SENT-TE+CQP+TF	MSE	2.77	**2.31**	**2.14**	2.82	**2.47**

Experimental Setup

DISCUSSION

- Sentiment-based summaries when not using sentiment features is not of great help
- Query-focused summaries seem to be appropriate in all cases
- 30% compression rate obtains on average better MSE values
- The combination of techniques TE+CQP+TF of 10% compression rate might be useful for other NLP tasks
  - its MSE is lower than for other techniques with the same compression rate

Conclusion and Future Work

CONCLUSIONS

- **Rating inference task** → predict the correct rating associated to a document based on the language expressed
- Extensive **analysis of summarization strategies** for the rating inference task instead of using full reviews
  - Generic vs. Query-focused vs. Sentiment-based
  - Combination of techniques for detecting relevant information (term frequency, the code quantity principle, textual entailment)
  - Analysis of baselines (lead, final, qf, sent)
- Results show that **query-focused** and **sentiment-based** summaries are the **most appropriate** for tackling this task

Conclusion and Future Work

FUTURE WORK

- Extend the size of the dataset
- Replicate the experimnents in another domain
  - E.g. Movie reviews
- Analyze the problems of such a fine-granularity classification (1-5 labels)
  - Simplifying the problem considering only 3 classes instead of 5

USING TEXT SUMMARIES FOR PREDICTING RATING SCALES

Thank you very much for your attention!

UNIVERSITAT POMPEU FABRA

Universitat d'Alacant
Universidad de Alicante

DEPARTMENT OF INFORMATION AND COMMUNICATION TECHNOLOGIES

NLP NATURAL LANGUAGE PROCESSING

PLSI

USING TEXT SUMMARIES FOR PREDICTING RATING SCALES

Horacio Saggion
Grupo TALN
Universitat Pompeu Fabra
Horacio.saggion@upf.edu

Elena Lloret and Manuel Palomar
Grupo GPLSI
Universidad de Alicante
elloret@dlsi.ua.es
mpalomar@dlsi.ua.es