Grégoire JADI

gregoire.jadi@univ-nantes.fr

Spécialité : Informatique
Laboratoire : LINA
Équipe : TALN
Directeur de thèse : Béatrice Daille
Co-encadrants : Laura Monceaux,
Vincent Claveau

Evaluating Lexical Similarity to Build Sentiment Analysis

The Context

Annotated Data

Resources

- Lexicons
- Clusters
- Distributionnal thesaurus
- Vector Space Models

MACHINE LEARNING MODEL

(SVM, Neural Network...)

The Focus

LEXICAL SIMILARITY FOR OPINION RESOURCES

Distance measures between words in vector space models are commonly used to create or enhance opinion resources

"The food was awesome but the server was a bit grumpy"

Restaurant : positive
Food : positive
Server : negative

The Question

HOW TO EVALUATE RESOURCES ?

- The classical way :
Train a machine learning model and evaluate its performance

- Our way :
Evaluate resources against manually built lexicons

PROCESS

1. Generate ordered lists of words from a lexicon

2. Try to reproduce the ordered lists with a distance measure in a vector space model from a query word

3. Compare the predicted list with the reference list using several metrics (Pearson's r, Kendall's τ, Spearman's ρ, MAP, R-Prec, P@rank)

Ongoing Work

BEYOND WORD DISTANCES

Use analogy solving capabilities of word representations to adress the issue encountered with distance measures

The Results

WORD DISTANCES ARE NOT VERY EFFICIENT

Positive but weak correlation between reference lists and generated lists

Works better on some dimensions than others

Lexicons do not fully agree between themselves

The Method

EXAMPLE

NRC Emotion Lexicon (14182 words)

LEXICONS USED

- ANEW :
3 real valued dimensions
- SentiWordNet :
3 relad valued dimensions
- NRC Emotion Lexicon :
8 binary dimensions

	Abandon	Happy	Weekly
Anger	✗	✗	✗
Anticipation	✗	✓	✗
Disgust	✗	✗	✗
Fear	✓	✗	✗
Joy	✗	✓	✗
Sadness	✓	✗	✗
Surprise	✗	✗	✗
Trust	✗	✓	✗