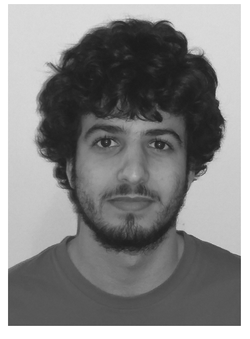
COMUE L'UNMA - École Doctorale STIM
Sciences et Technologies de l'Information et Mathématiques

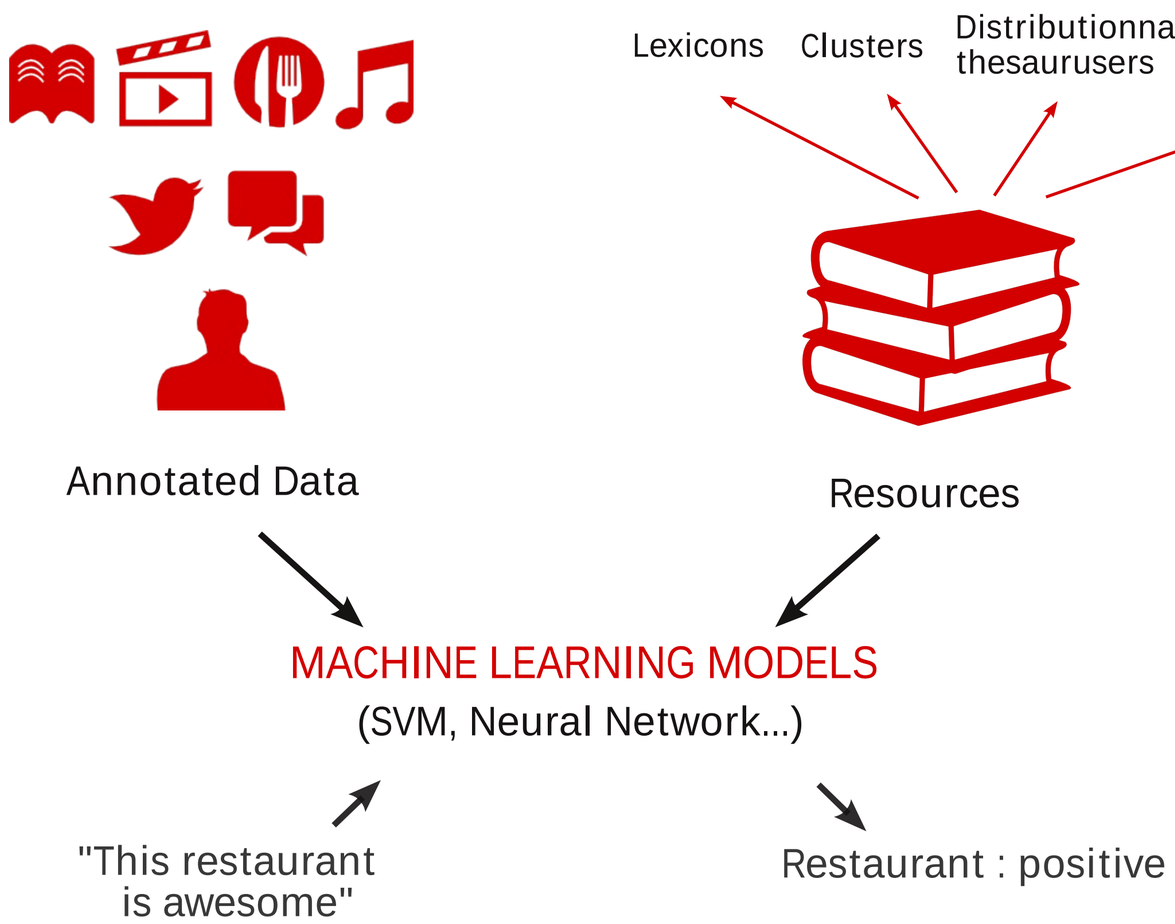
Grégoire JADI
gregoire.jadi@univ-nantes.fr
Spécialité : Informatique
Laboratoire : LINA
Équipe : TALN



Evaluating Lexical Similarity to Build Sentiment Analysis

Directeur de thèse : Béatrice Daille
Co-encadrants : Laura Monceaux, Vincent Claveau

The Context >> An opinion system analysis



Opinion lexicon can be built manually or automatically

Distributionnal thesaurus and vector space are unsupervised ressources used either directly or to build and enhance opinion lexicons

HOW TO EVALUATE RESSOURCES ?

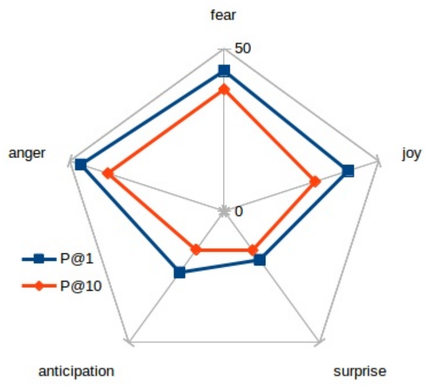
- The classical way :
  Train a machine learning model and evaluate its performance
- Our way :
  Evaluate resources against manually built lexicons

The method

Try to reproduce several lexicons with state of the art word representations and evaluate the results with several metrics : Pearson's r, Spearman's p, Kendall's r, MAP, R-prec, P@1, P@5, P@10, P@50, P@100.

1. Generate a reference list from reference lexicons
2. Try to reproduce the list with the word representations (L2 distance)
3. Compare the predicted list with the reference list

Example : NRC Emotion Lexicon (14182 words)



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

Reference lexicons used

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

The results

Positive but weak correlation between reference list and generated list

Works better on some dimension than others

Lexicons do not fully agree between themselves

Ongoing work

Use analogy solving capabilities of word representations to adresse the issue encountered with distance measure

UNIVERSITÉ DE NANTES
lina
CominLabs