Advanced Human Machine Interaction
Interaction Data Analysis

Textual Interaction Analysis

Alexandre Pauchet

alexandre.pauchet@insa-rouen.fr - BO.B.RC18

Normandie Université

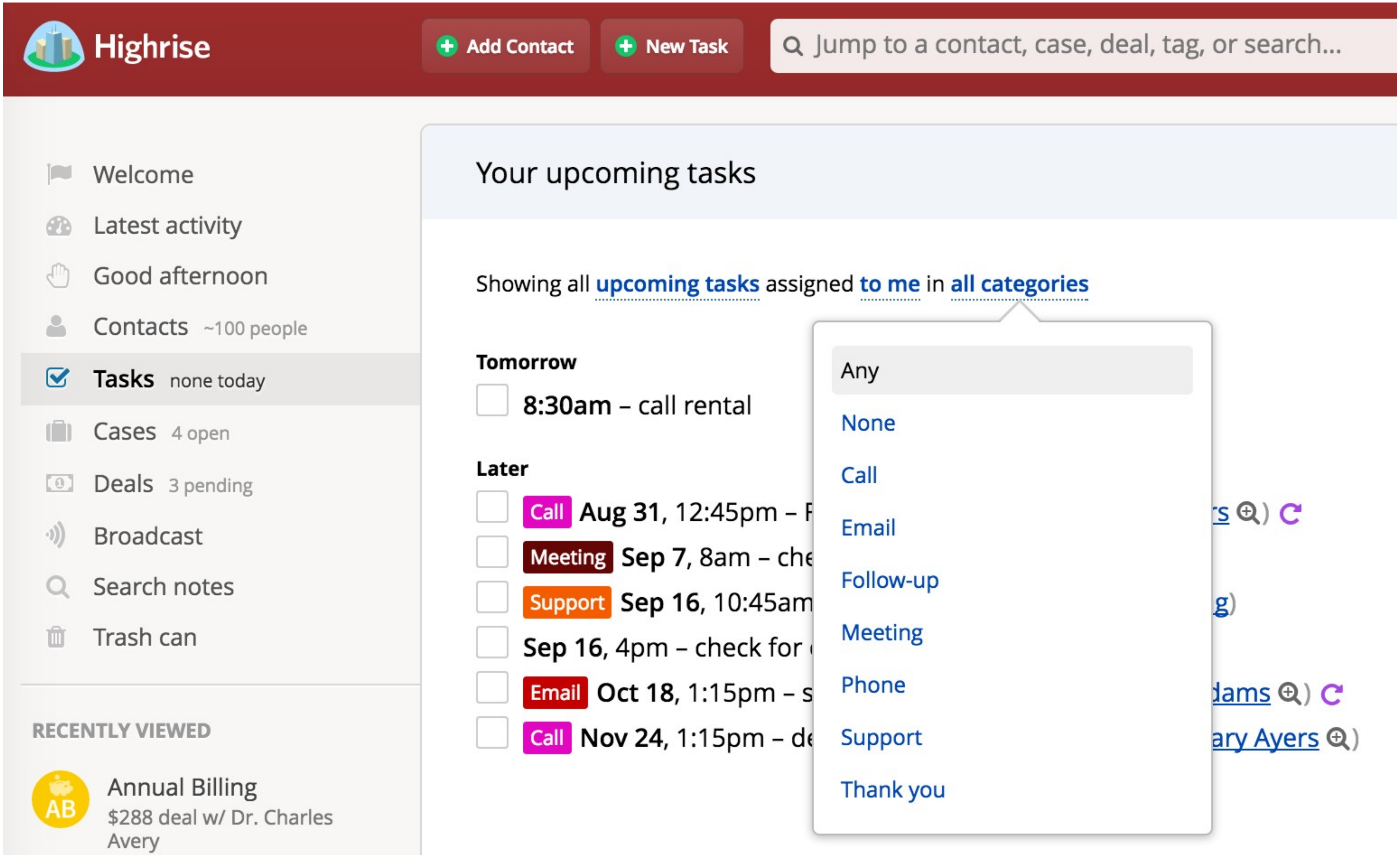
INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES ROUEN NORMANDIE

litis

Scope

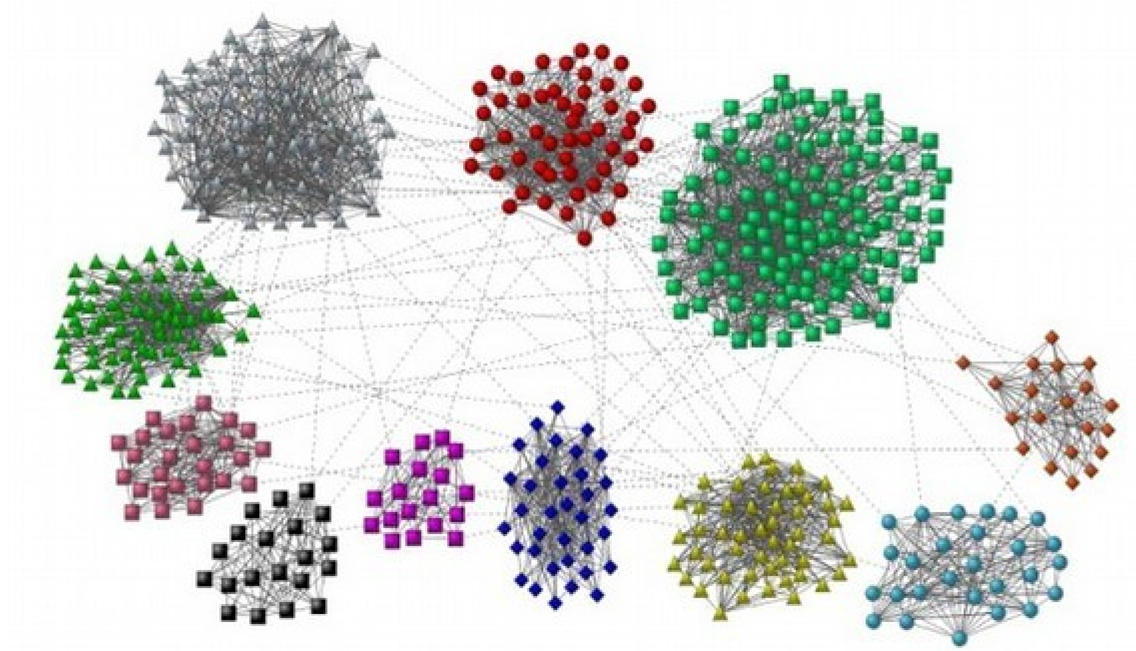
- **Type of textual interactions:**
  - Human-(mediated)-Human interactions: forums, mails, social network messages, blogs, chat, dialogues in online games, …
  - Human-machine interactions: (multi-modal) dialogue
- **Application goals:** filtering / labeling / sorting, opinion mining / sentiment analysis / affective computing, community detection, user assistance, companioning, serious games / virtual environments for learning, ...
- **Scientific problems:** supervised machine learning, unsupervised machine learning / data-mining, natural language processing (NLP) / natural language understanding (NLU) / natural language generation (NLG)

Example

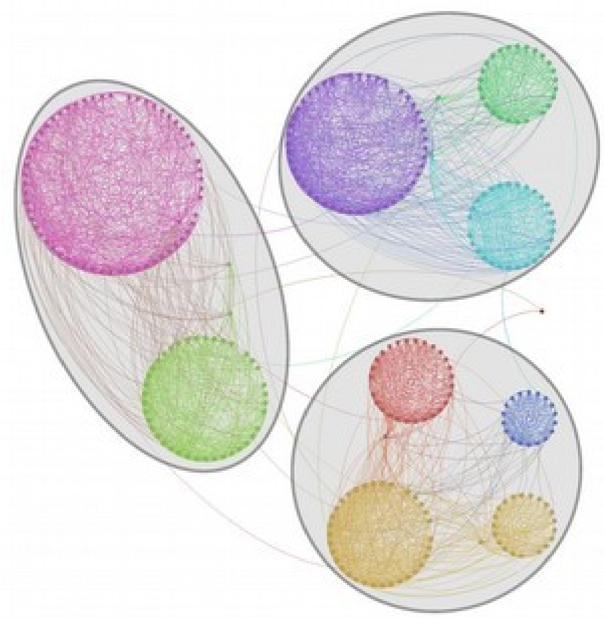


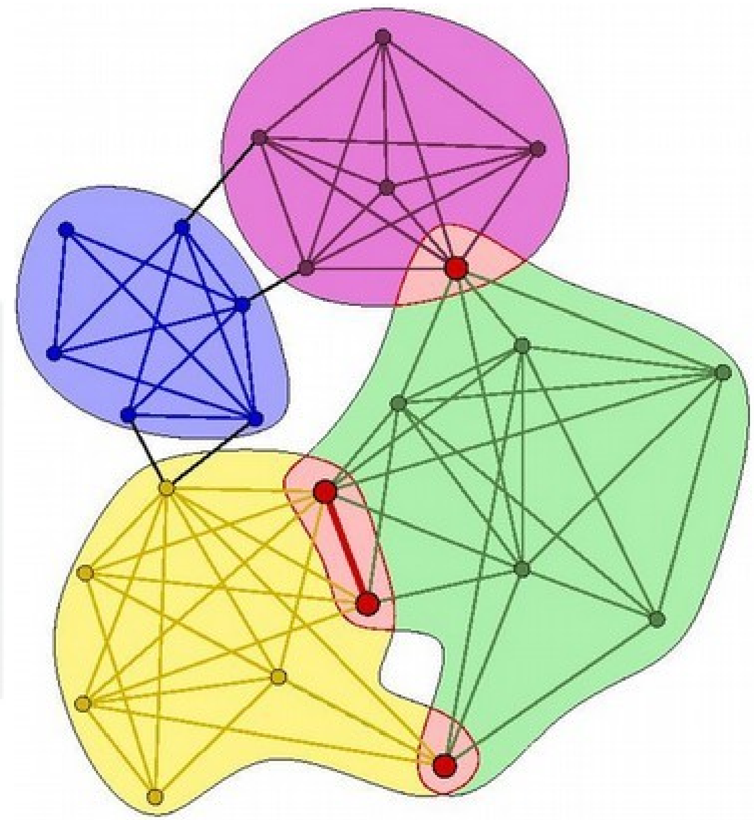
Automatic labeling of emails (supervised learning)

Example



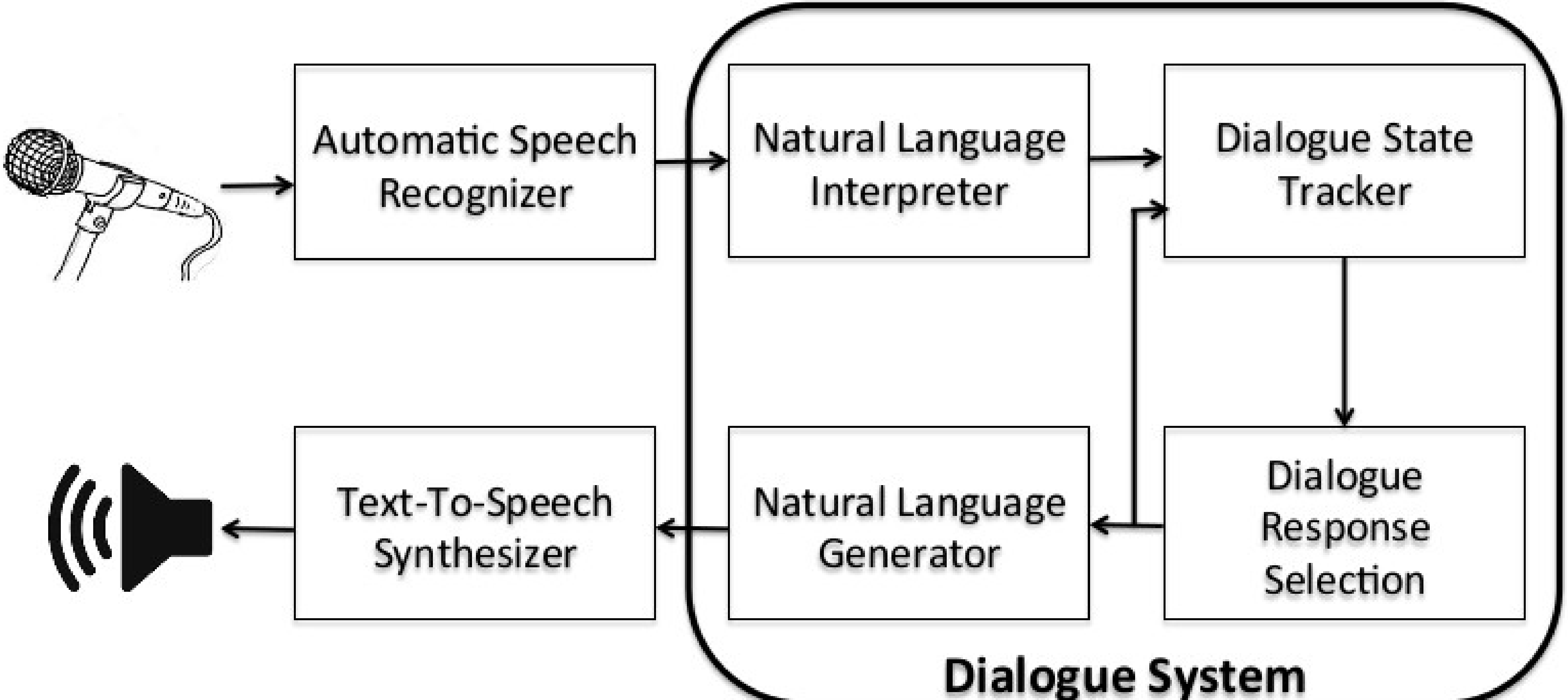
Community detection in online social networks (unsupervised learning)





Example

Dialogical assistant (supervised learning, NLP, NLU, NLG)



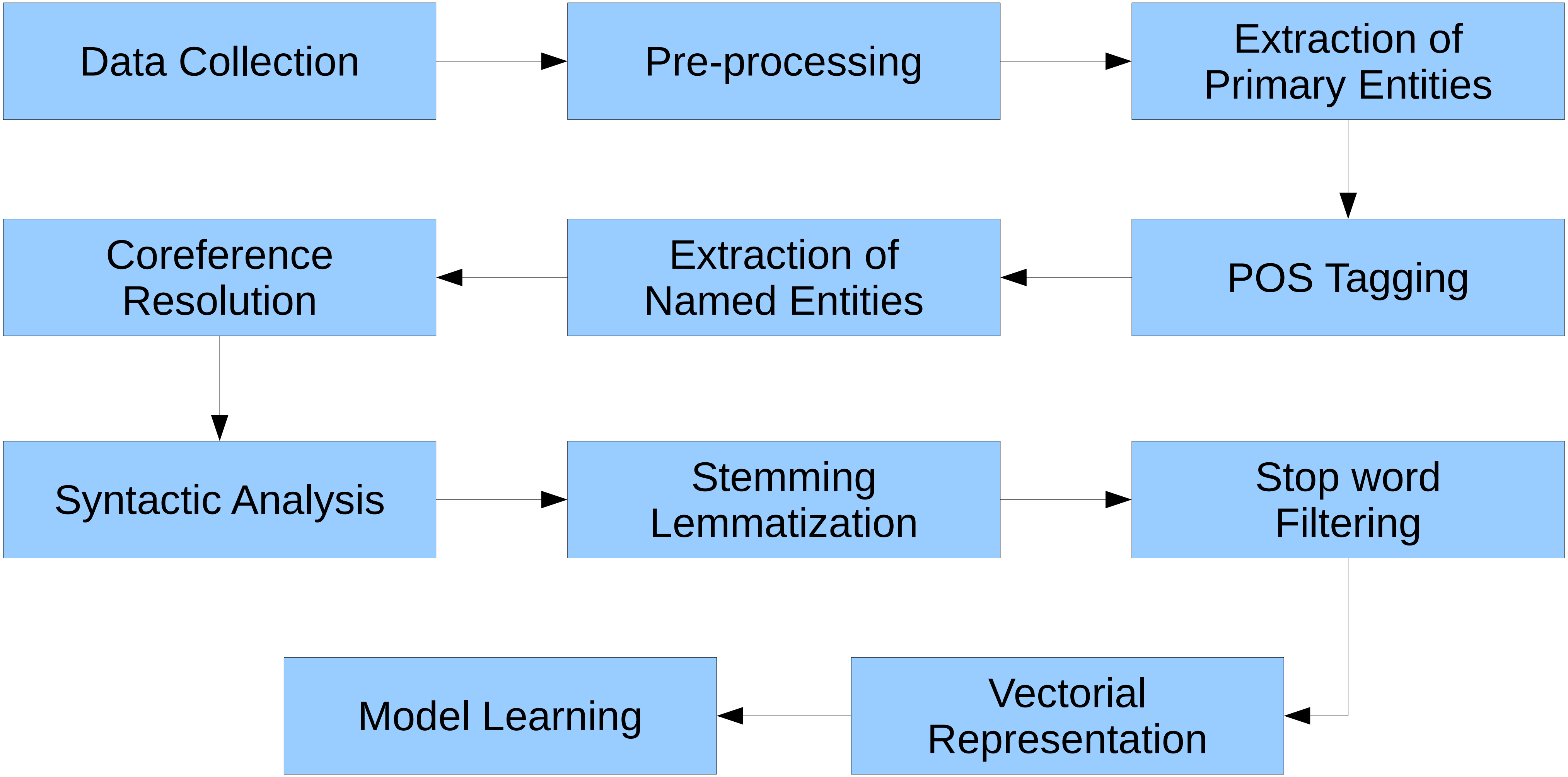
I. V. Serban, R. Lowe, P. Henderson, L. Charlin, J. Pineau, "*A Survey of Available Corpora for Building Data-Driven Dialogue Systems*", https://arxiv.org/pdf/1512.05742.pdf

Exercise

- Formalize the problem to automatically group similar old messages every year, according to similar topics

Classification of (short) texts

Process (1/5)

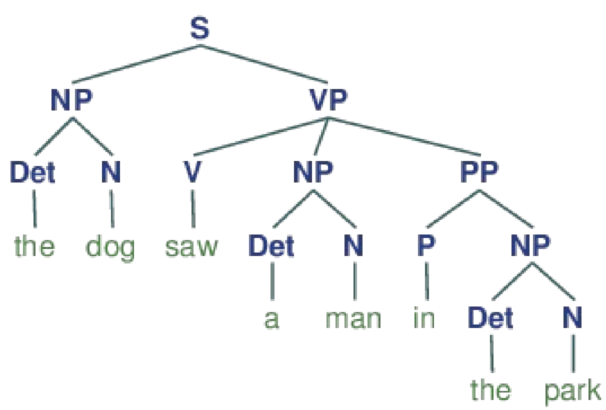

Data Collection
Pre-processing
Extraction of Primary Entities
POS Tagging
Extraction of Named Entities
Coreference Resolution
Syntactic Analysis
Stemming Lemmatization
Stop word Filtering
Vectorial Representation
Model Learning

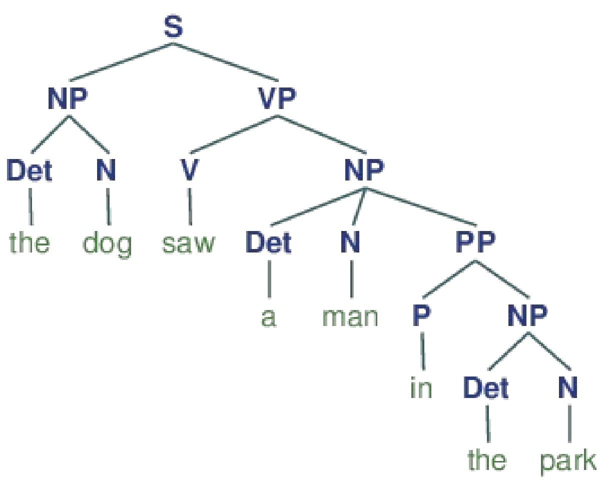
Process (2/5)

- **Textual data collection**: identification of sources, collection of texts from sources and text extraction / filtering
  - Transcription? Collection by keywords?
  - “Grammar”: emoticons, slangs, onomatopoeia, ...
- **Pre-processing of textual data**: encoding standardization, filtering of special characters, "translation" of SMS language, ...
- **Extraction of primary entities**: words, nominal and verbal expressions, …
  - Segmentation: tokenizer + lexicon → language dependent
  - Words ≠ concepts: synonyms and homonyms
- **POS tagging** (part-of-speech): grammatical characterization of text components with a lexical category and a function.
  - Treetagger: http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
  - Stanford POS: https://nlp.stanford.edu/software/tagger.shtml

Process (3/5)

- **Extraction of named entities**: names of persons and characters, places, organizations, dates, …
  - Difficulties: metonymy *(*ex: *Little Red Riding Hood, The White House will be making an announcement),* polysemy
  - Stanford NER: https://nlp.stanford.edu/software/CRF-NER.shtml
- **Coreference resolution**: finding all expressions that refer to the same entity in a text (he / she / it, his / her, …)
  - Stanford Coref Annotator: https://stanfordnlp.github.io/CoreNLP/coref.html
- **Syntactic analysis**: negation, "quantification" of adverbs, …
  - Syntax trees
  - Ex:



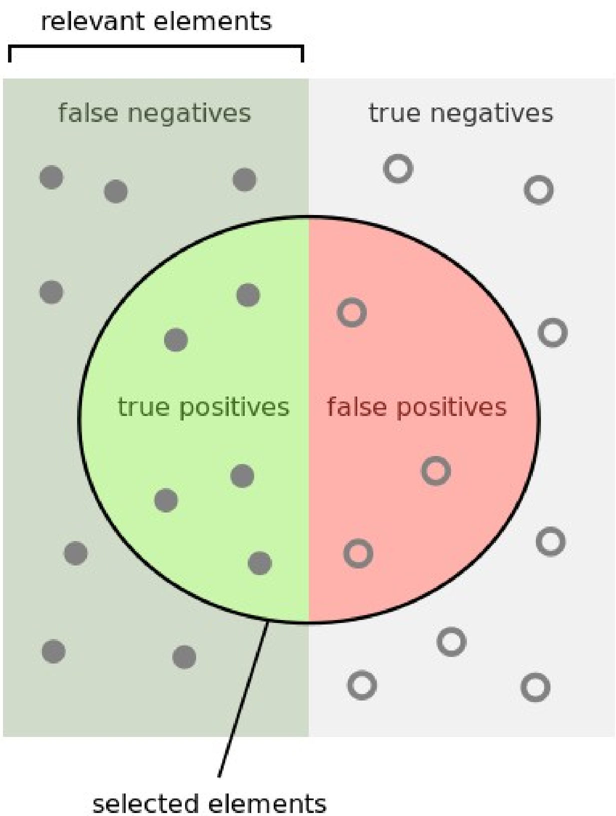


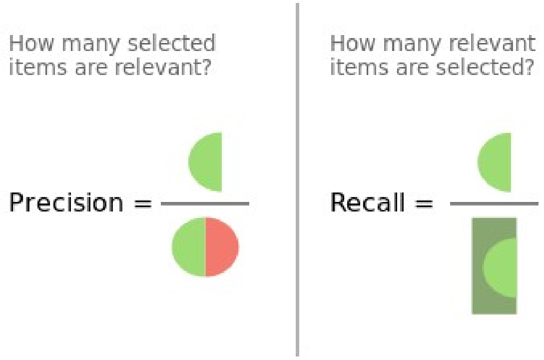
  - Tool: NLTK tree module (http://www.nltk.org/)

Process (4/5)

- **Stemming or Lemmatization:**
  - Stemming: replace each word by its root; → English
  - Lemmatization: replace each word by its canonical form; → French
- **Stop words filtering**: prepositions, conjunctions, articles, auxiliary verbs, ...
- **Vectorial representation:**
  - *Principle*: each text (document, request, utterance) is represented by a large and sparse vector; bag of words approach.
  - *Possible dictionaries*: set of words of the corpora, smaller or larger external set of words, n-grams, …
  - Comparison of vectors using the cosine distance
  - *Improved representations*: co-occurrence matrix, tf*idf, LSA, word embeddings (Word2Vec), language model (BERT-like model), ...
- **Learn the models**

Process (5/5)





- **Model evaluation**:
  - Precision

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

  - Recall

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

  - F-measure $F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$

- **Remarks**:
  - Unbalanced classes!
  - Evaluation by class

https://en.wikipedia.org/wiki/Precision_and_recall

TF*IDF

- **Objective**: weight the terms according to their "importance" in the selection process; particularly useful for information retrieval:
  - *Term Frequency*: the importance of a term for a document is proportional to the number of occurrences of the term in the document,
  - *Inverse document frequency*: the importance of a term for all documents is inversely proportional to the number of documents in which it appears (terms appearing in few documents are more discriminating that terms in many documents)
- **Definition**: $P_{ij} = \frac{n_{ij}}{\|d_j\|} * \log\left(\frac{n}{n_i}\right)$

Exercise

1)Human-machine interfaces for computer applications

2)User opinion of computer system response time

3)User interface management system

4)System engineering to improve the computer response time

5)Complex systems made of humans, computers and agents

6)A dialogical agent is interacting with a human user on computer

- **Filter the texts**
- **Construct a BoW vectorial representation**
- **Construct the co-occurrence matrix**
- **Use a TF*IDF vectorial representation**

Latent Semantic Analysis (LSA)

- **Objective**: identification of "concepts", corresponding to correlations between terms, to represent the documents in a collection
  - Removal of "noise" from data;
  - Replacement of lemmas by corresponding concepts.
- **Approach**: singular value decomposition (SVD) applied to a document-term matrix (can be weighted using a TF*IDF):
  - Matrix: 1 line/document and 1 column/term
  - SVD: $M = U \cdot S \cdot V^T$
  - Rank reduce singular value decomposition: considering only the k largest values and associated vectors

$$M = U_k \cdot S_k \cdot V_k^T$$

Word embeddings (Word2Vec)

- **Objective**: vectorial representation of words from large text corpora, that incorporate semantic and syntactic features:
  - The projection space is constructed using a (very large) “independent” corpus of texts
  - Skip-gram model: find representations to predict the best possible context of words; let w_i … w_T be a sequence of words and k be a context width

$$Max \frac{1}{T} \sum_{t=1}^{T} \sum_{j=-k}^{j=k} \log\left(p\left(w_{t+j}/w_t\right)\right)$$

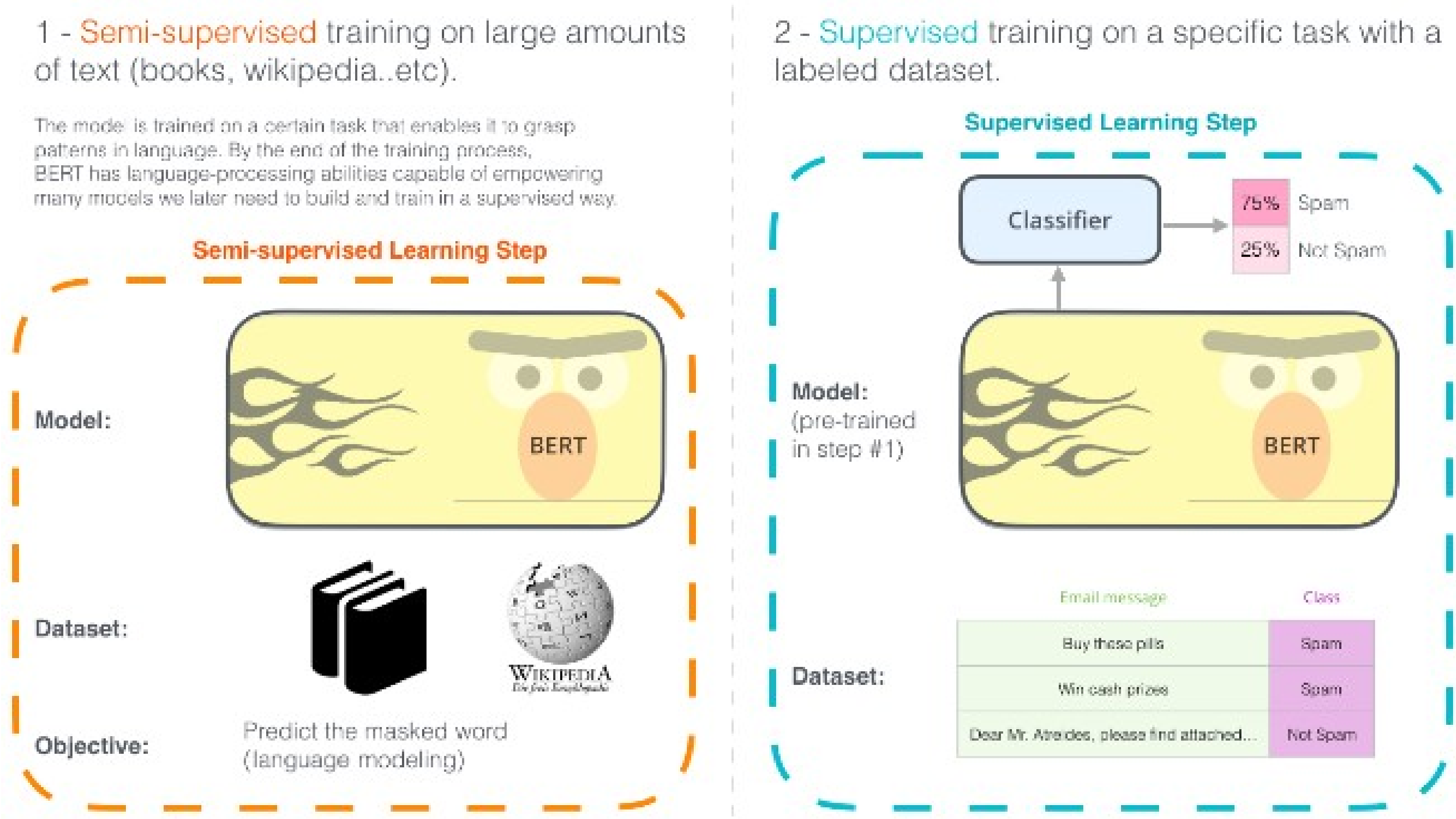
- **Remarks**:
  - With this representation, the words are grouped by similarity context
  - A form of "additivity" is possible:

vec(King) - vec(Man) + vec(Woman) = vec(Queen)

- **Ref**: T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, and J. Dean, “*Distributed representations of words and phrases and their compositionality*”, In Advances in neural information processing systems, pp. 3111–3119, 2013.

Language models

ELMo, ULM-FiT, BERT, ...



https://medium.com/huggingface/introducing-fastbert-a-simple-deep-learning-library-for-bert-models-89ff763ad384

Opinion Mining
Affect detection

Definitions and objectives

- **Opinion Mining / Sentiment Analysis**:
  - Automatic recognition of opinions in texts
  - Group users according to their opinions
- **Social media mining**:
  - Information diffusion
  - Recognition of roles and influence detection
- **Opinion** = {author, date, target, feature, polarity/sentiment};

B. Pang and L. Lee, "*Opinion mining and sentiment analysis*", Foundations and Trends in Information Retrieval, Vol. 2:1-2, pp. 1–135, 2008.

- **Affect detection**:
  - Emotion={Anger, Disgust, Fear, Joy, Sadness, Surprise, Neutral}
  - Affect/sentiment = Valence [-1:1]

Process

- **Classic process**: textual data collection, pre-processing, POS tagging, stemming (en) / lemmatization (fr), stop words filtering, vectorial representation (bag of words + TF*IDF + LSA), supervised classification (SVM, Random Forests, NN)
- **Approaches**:
  - 2-step approach:
    - Objectivity/Subjectivity
    - Positive/Negative
  - 3 classes: Positive/Negative/Neutral
  - Valence (regression)
  - Stance detection for a given target
- **Linguistic resources**: Wordnet affect, Sentiwordnet
- **Inter-Annotator agreement**: low
- **Expected results**: ~75% of precision!

Stance versus Opinion

"***Sharknado 3 may be the best film I've seen yet. #Sharknado3 #America***"

- Target: Sharknado 3 ; polarity: POSITIVE
- Target: Sharknado 3 ; SUBJECTIVE → POSITIVE

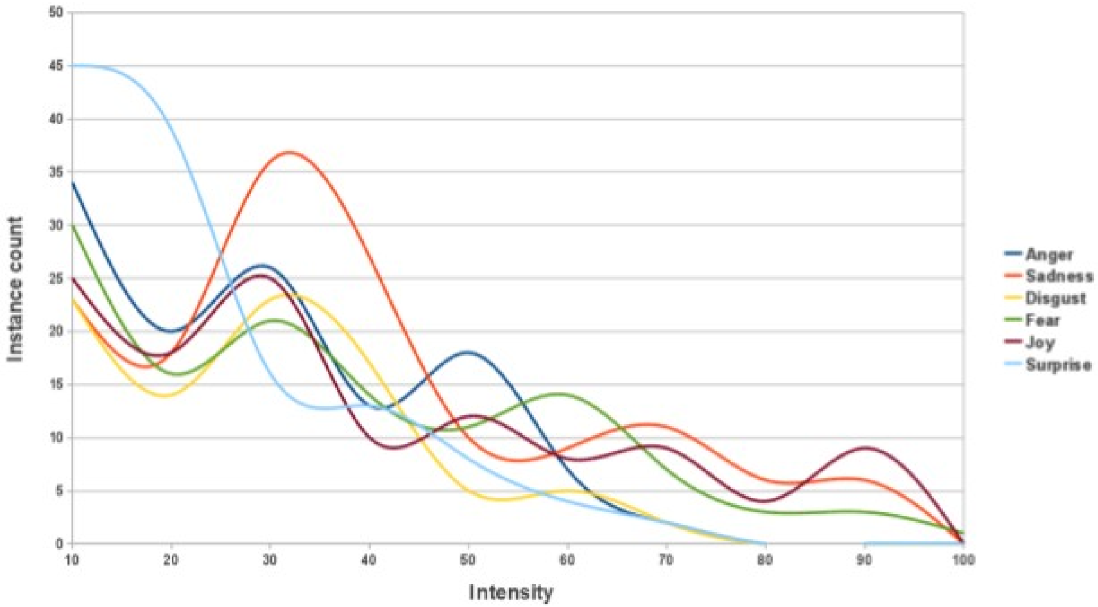
"***@HilaryClinton going to prison. She can help #build-thewall.***"

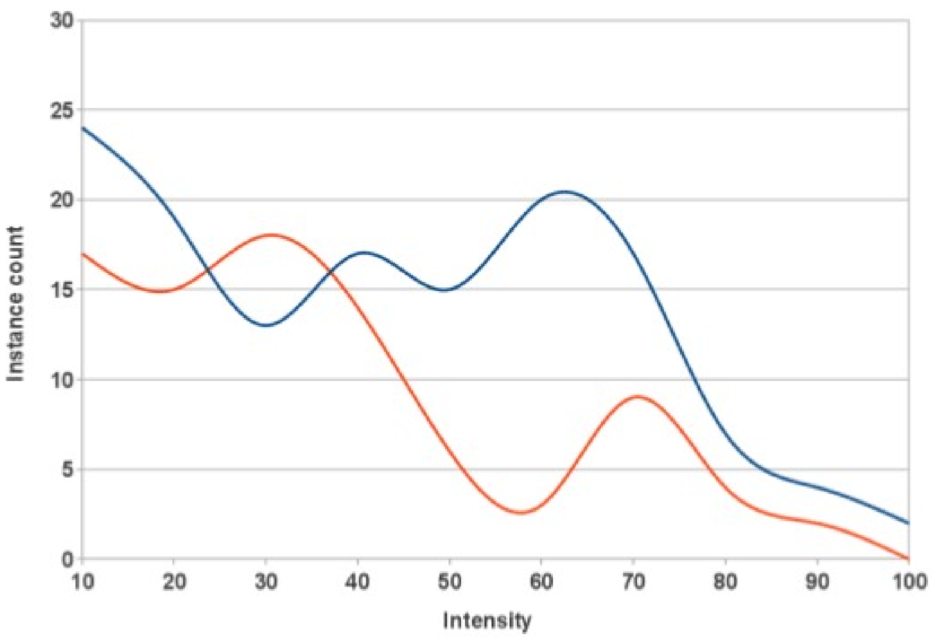
- Target: Hilary Clinton ; polarity: NEGATIVE
- Stance about Donald Trump: POSITIVE

Example: emotion detection

A	D	F	J	Sad.	Sur.	Headline
-	-	-	0.15	0.25	-	Bad reasons to be good
-	-	-	-	-	0.36	Martian Life Could Have Evaded Detection by Viking Landers
-	-	0.68	-	-	-	Hurricane Paul nears Category 3 status
-	-	-	0.75	-	0.57	Three found alive from missing Russian ship - report
0.52	0.64	0.50	-	0.43	-	Police warn of child exploitation online

Anger=**A**, Disgust=**D**, Fear=**F**, Joy=**J**, Sadness=**Sad.**, Surprise=**Sur.**



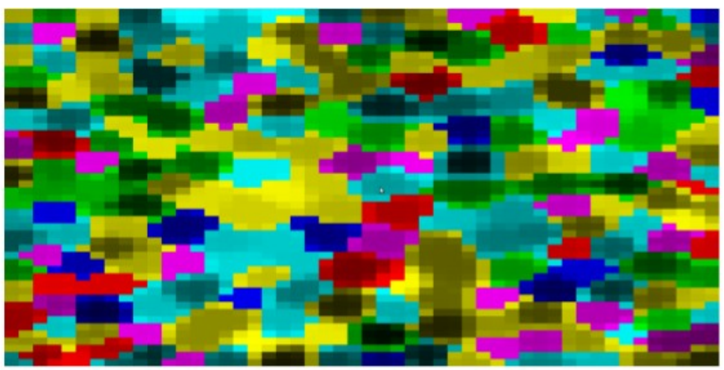


SemEval 2007, task 14 : Strapparava, C. and Mihalcea, R. (2008). Learning to identify emotions in text, In Proceedings of the 2008 ACM symposium on Applied computing, pp. 1556–1560, 2008.

Results

	LSA All emotional			UA			UPAR7		
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
Anger	6.20	**88.33**	11.59	12.74	21.60	16.03	16.67	1.66	3.02
Disgust	1.98	**94.12**	3.88	0.00	0.00	-	0.00	0.00	-
Fear	12.55	**86.44**	21.92	16.23	26.27	20.06	**33.33**	2.54	4.72
Joy	18.60	**90.00**	30.83	40.00	2.22	4.21	**54.54**	6.66	11.87
Sadness	11.69	**87.16**	20.62	25.00	0.91	1.76	**48.97**	22.02	**30.38**
Surprise	7.62	**95.31**	14.11	13.70	16.56	**14.99**	12.12	1.25	2.27

Classifier: self organizing map (SOM)



	Nb. of instances	
No emotion	642	64.85%
Anger	14	1.41%
Disgust	6	0.61%
Fear	65	6.57%
Joy	110	11.11%
Sadness	81	8.18%
Surprise	38	3.84%
Combined	34	3.43%

	Precision	Recall	F1
LSA training	20.50	19.57	20.02
LSA Gutenberg	24.22	23.31	**23.76**
LSA All emotion	9.77	**90.22**	17.63
UA	17.94	11.26	13.84
UPAR7	**27.60**	5.68	9.42

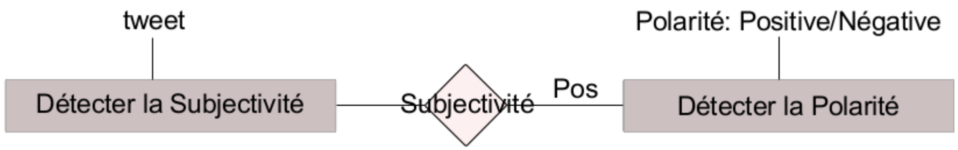
	LSA training			LSA Gutenberg		
	Prec.	Rec.	F1	Prec.	Rec.	F1
Anger	10.00	11.86	10.85	**18.52**	15.38	**16.80**
Disgust	3.33	4.17	3.70	**8.33**	7.69	**8.00**
Fear	19.01	17.76	18.36	28.39	27.67	**28.03**
Joy	36.75	36.75	36.75	40.49	64.62	**49.79**
Sadness	24.14	40.00	30.11	27.08	19.60	22.74
Surprise	**29.73**	6.92	11.23	22.50	4.95	8.11

Example: Opinion mining on tweets

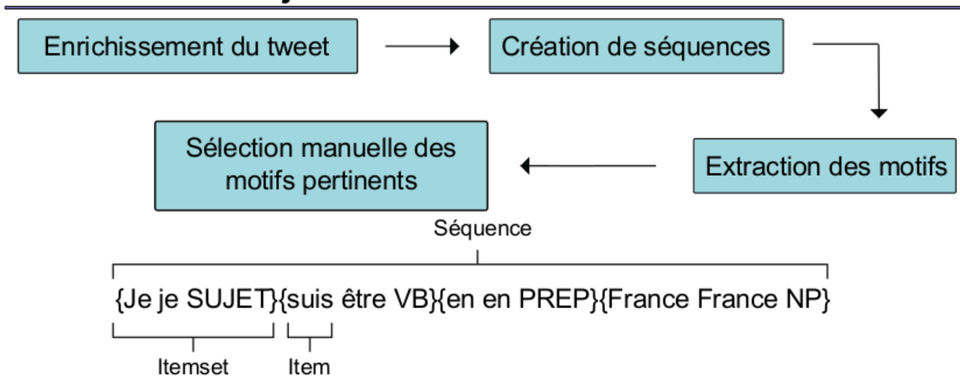
Corpus de tweets

Corpus d'apprentissage :
Chaque tweet est annoté automatiquement suivant les émoticônes qu'il contient : polarité positive (:-p,:-D...) ou négative (:-(, :-s...).
Résultat : 300 000 tweets français et 300 000 tweets anglais.
Corpus de test :
Annotation manuelle par des ingénieurs.
Résultat : 800 tweets anglais et 700 tweets français.

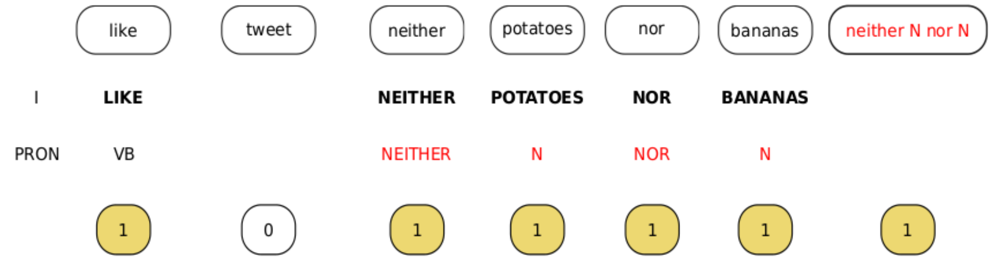
Méthode proposée



Détection de subjectivité



Détection de polarité



Résultats obtenus

Détection de subjectivité (motifs séquentiels fréquents) :

	précision	rappel
français	65%	66%
anglais	64%	62%

Détection de polarité (SVM) :

		précision	rappel
PRESENCE	français	62%	61%
	anglais	74%	73%
POSITION	français	62.6%	57%
	anglais	60%	60%