Classification de sentiments sur le Web 2.0

Vincent Guigue
vincent.guigue@lip6.fr

Laboratoire d'Informatique de Paris 6 (LIP6)

25 octobre 2012

Plan

Opinion Mining & Web 2.0

- **Web 2.0** : basé sur les *User Generated Contents*
- **Opinion Mining** : exploitation du web 2.0

Enjeux économiques : e-reputation

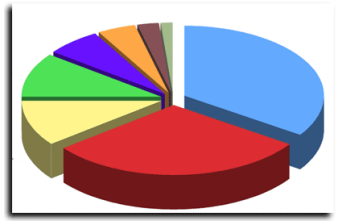
e-reputation.org

L'E-Réputation est l'image véhiculée par une marque (société, personne...) sur tous les types de supports numériques (médias, réseaux sociaux, forums, messagerie instantanée...). Cette image doit être en cohérence avec la stratégie de marque, d'où, les efforts pour veiller, analyser et orienter.

- Maitriser son image
- Asseoir sa crédibilité
- Eviter tout départ de rumeur

⇒ Des outils outils automatiques pour analyser volume et **sentiment** rapidement et à grande échelle

Enjeux économiques : sondages & analyse



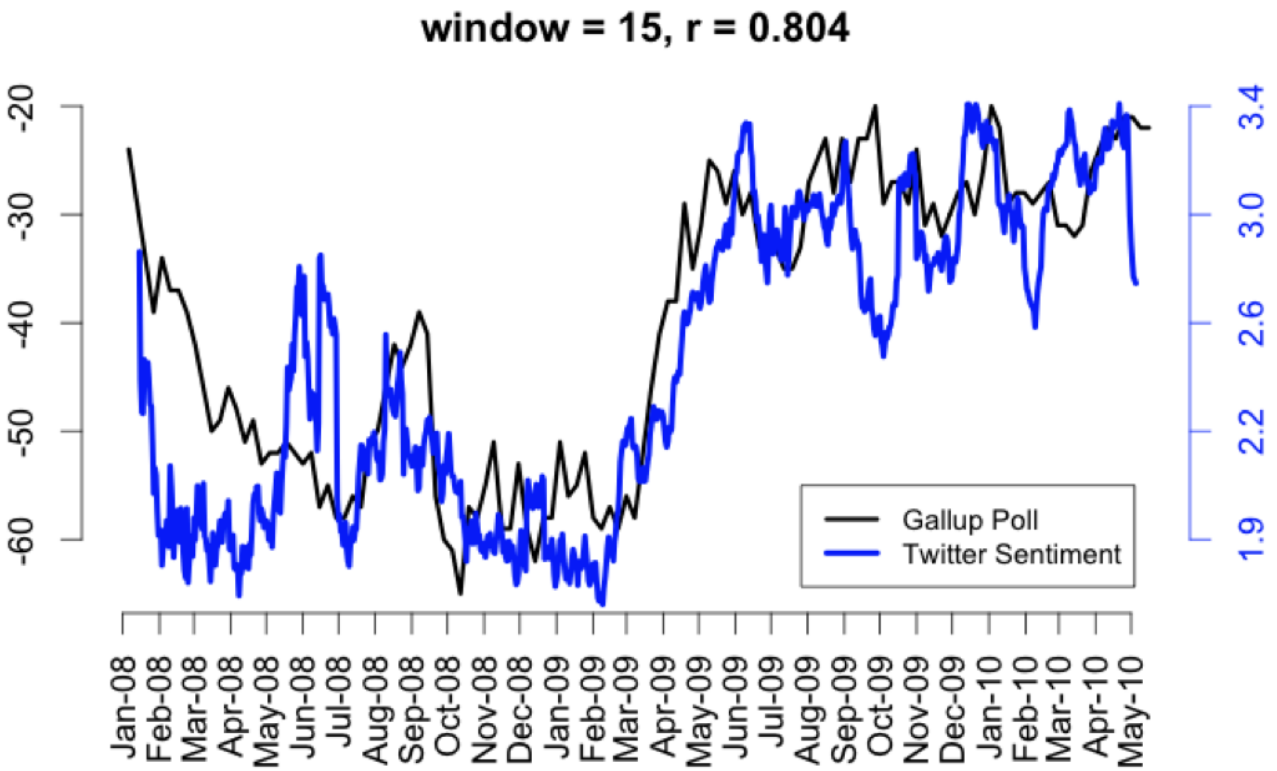
Sur un **sujet donné**, qui est positif, qui est négatif sur :

- un panel de blogs/sites
- twitter
- les sites d'une requête Google

Possibilités de suivi temporel, d'étude des évolutions...

⇒ Des outils outils automatiques pour analyser le **sentiment** sur diverses sources au cours du temps

Enjeux économiques : sondages & analyse



ICS

(Index of Consumer Sentiment), Michigan monthly poll

O'Connor et al., ICWSM 2010

From Tweets to Polls : Linking Text Sentiment to Public Opinion Time Series.

Enjeux économiques : RI

Proposer un moteur de recherche sensible aux sentiments :

- résumer les bons/mauvais points d'un produit
- recherche avancée e.g. *Quelles sont les qualités/défauts du dernier iphone d'après les utilisateurs ?*
- rechercher la comparaison de deux produits
- arguments (contradictoires) dans un débat politique

... Et valoriser les résultats

- choisir les publicités à insérer dans les réponses
- identifier les mauvais points d'un produit pour placer des publicités d'un concurrent

Enjeux économiques : RI

Exemple de résultat d'une requête sur Google Shopping :

Apple iPhone 4 Smartphone 16 GB - AT&T - WCDMA (UMTS) / GSM - Black
$380 online ★★★★½ 618 reviews
June 2010 - Smartphone - Apple - iOS - AT&T - GSM - WCDMA - 5 megapixel camera - 3.5 inch screen - Quad-band - W

Reviews

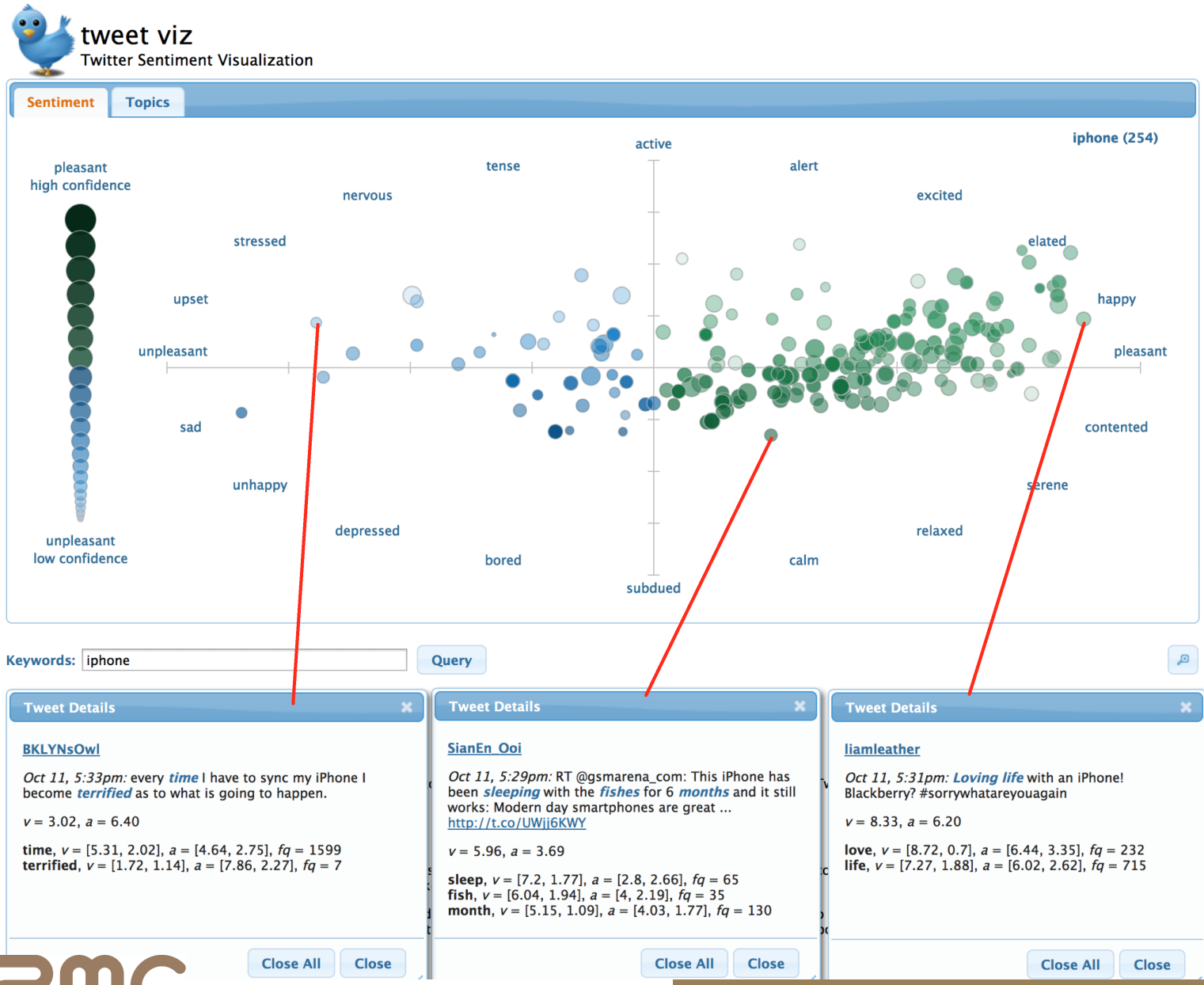
Summary - Based on 618 reviews

1 2 3 4 stars 5 stars

What people are saying

features	"Really enjoy all the features and apps available."
battery	"good but battery still needs improvement."
camera	"Takes pretty good pics."
screen	"Apple's display is awesome."
design	"Overall design, superb."
reception	"Easy to use ... Cons: No 3G networks in my area."
video	"Great video quality."

Enjeux économiques : RI



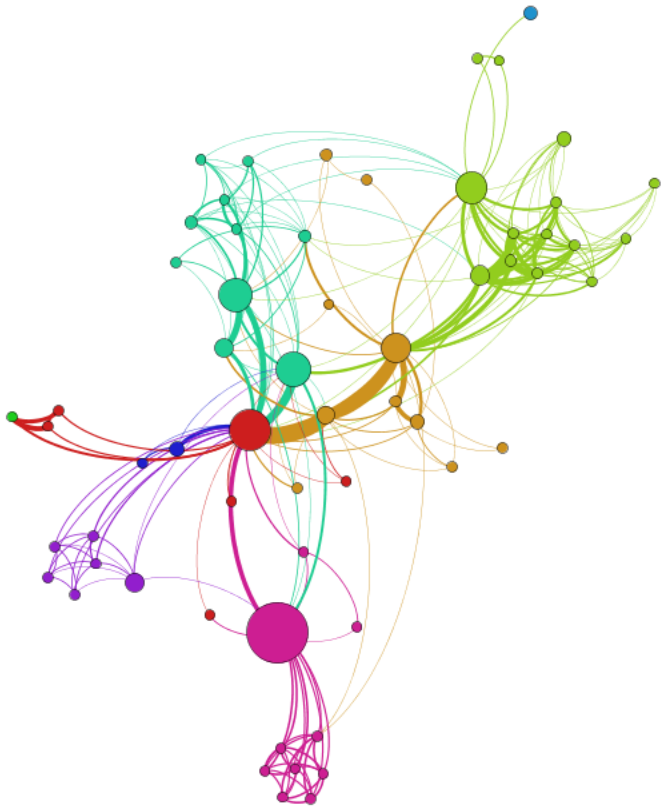
Enjeux économiques : RS & communautés

Toujours, sur un **sujet donné**, effectuer des sondages sur les réseaux sociaux

Mise en évidence de :

- Communautés d'opinion
- Diffusion d'opinion
- Leaders d'opinion

⇒ Segmenter la population pour affiner les analyses, identifier des leaders d'opinion pour manipuler les communautés



Enjeux économiques : détection du spam

Enjeux économiques forts ⇒

- Etape 1 : manipulation d'opinion
- Etape 2 : détecter la manipulation (spam review, faux profils...)

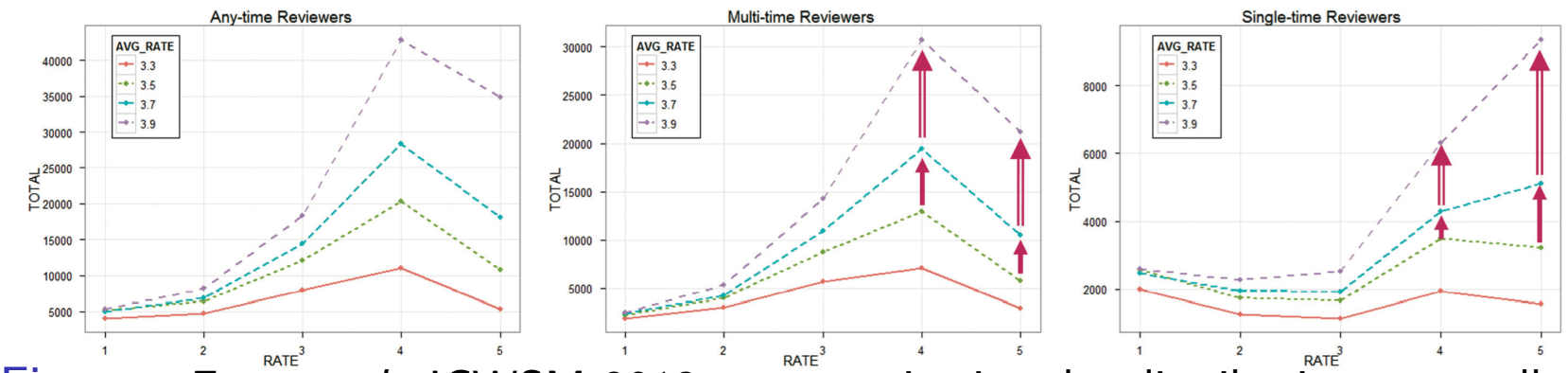


Figure : *Feng et al.*, ICWSM 2012 : caractérisation des distributions naturelles de revues

⇒ Identifier les spams/les revues fiables pour ne pas biaiser les études précédentes

Les données du web 2.0

Blogs

Revues (sites e-commerce, sites spécialisés...)

Réseau Social (e.g. Twitter)

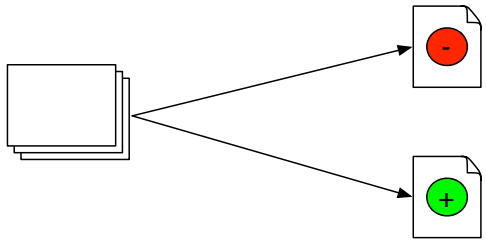
Forums

- Vocabulaires et styles spécifiques
- Accès plus ou moins aisé
- Données différentes $\Leftrightarrow$ approches différentes (avis personnel)

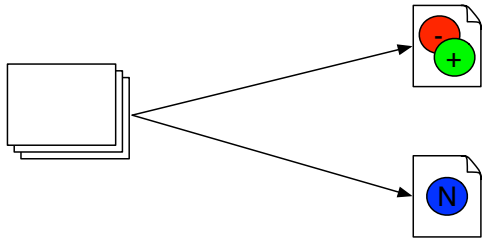
Problématiques Macro :

Exemples de problématiques au niveau des documents :

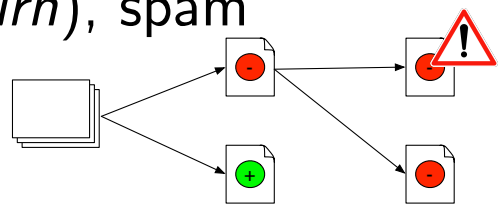
- Détection de polarité



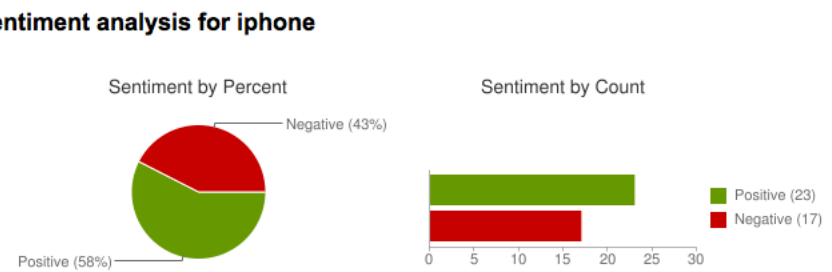
- Classification objectif/subjectif



- Détection de la colère (anticipation du *churn*), spam



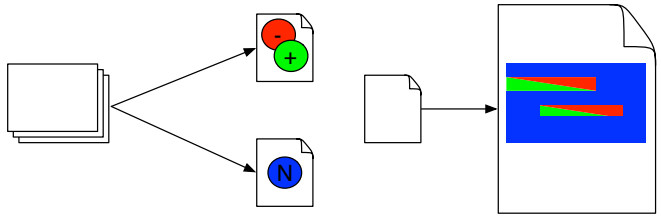
- Sondage, études : quantification



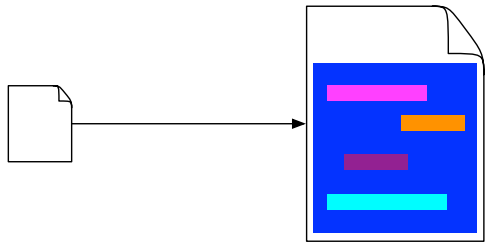
Problématiques Micro :

Exemples de problématiques dans les documents :

- Classification objectif/subjectif



- Détection/extraction/quantification de sentiments complexes



- Recherche d'informations + sentiments, résumés, comparaison :
  e.g. *Quelles sont les qualités/défauts du dernier iphone d'après les utilisateurs ?*
- Entités : qui émet l'opinion ? Quelle est la cible ?

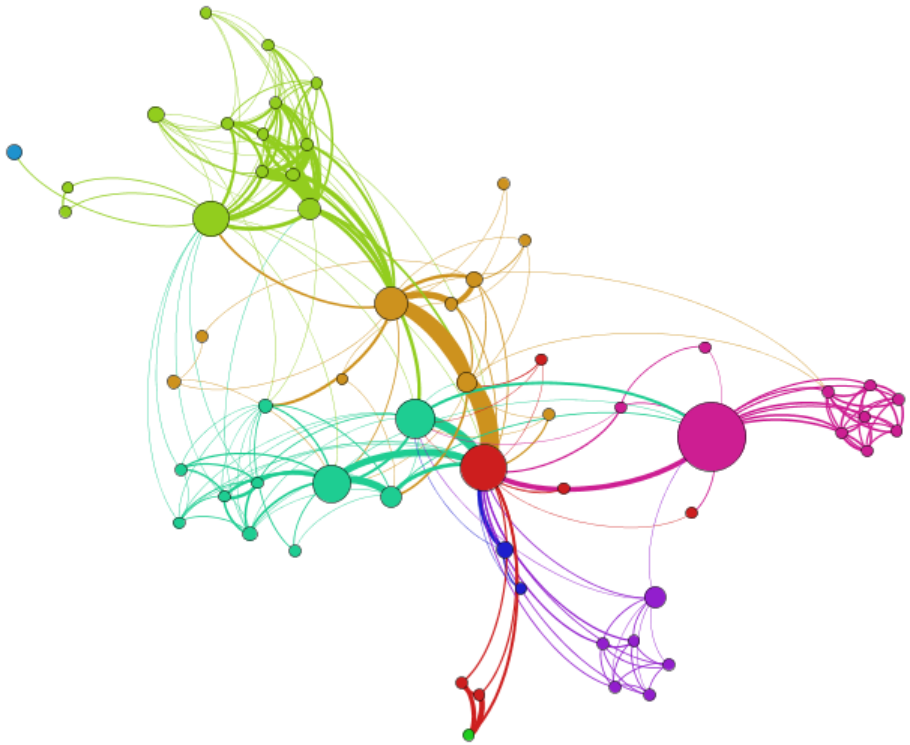
Problématiques Sociales :

Clustering diviser une population en fonction de ses affinités sur un sujet

Diffusion affiner la prédiction de sentiment en fonction des voisins

Contamination prédire qui sera touché par une information au temps t, quel est le patient 0, quels sont les facteurs de contamination ?

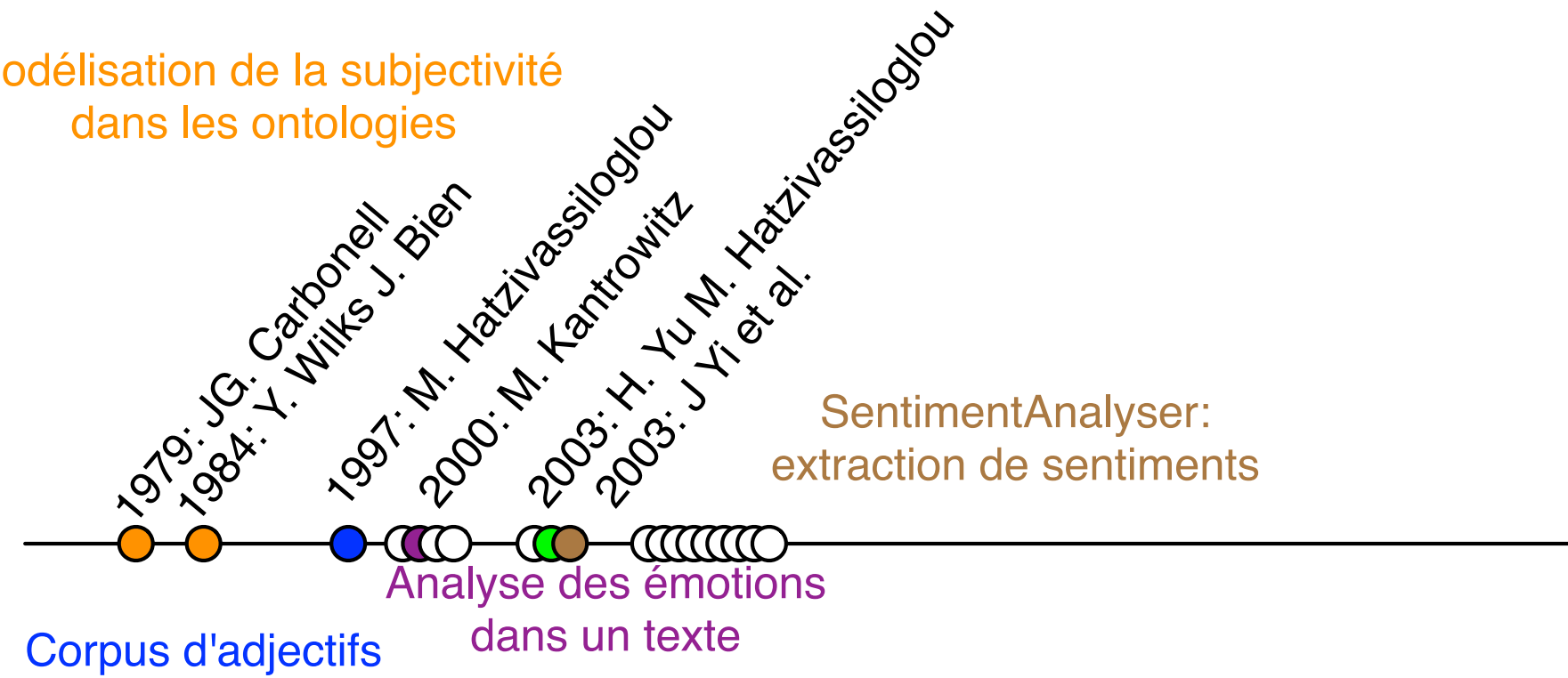
Manipulation identifier les influenceurs par analyse temporelle de la diffusion des messages (techniques de comptage)



Historique NLP (1)

Répondre à une requête contenant des émotions

Modélisation de la subjectivité dans les ontologies



Pang & Lee : *The year 2001 or so seems to mark the beginning of widespread awareness of the research problems and opportunities that sentiment analysis...*

Historique NLP (2)

Axes clés des solutions NLP

- Analyse morpho syntaxique des phrases
- Recherche de *patterns*, utilisation d'automates
- Utilisation de lexiques

Glissement progressif vers des méthodes d'apprentissage à partir de corpus étiquetés (en subjectivité, en opinion...). Le *Pointwise Mutual Information* (PMI) de Turney fait le lien avec le ML dès 2002.

P.D. Turney, ACL 2002

Thumbs up or thumbs down ? : semantic orientation applied to unsupervised classification of review

Historique ML

Web 2.0 : les utilisateur écrivent des commentaires, mais ils mettent aussi des étoiles ! ⇒ technique d'apprentissage supervisée

1999 Wiebe et al. : Création corpus + naïve bayes

2001 Das and Chen : Feature NLP (e.g. négation) + Apprentissage Statistique : ça marche !

2002 Pang et al. : Feature simple (BOW) + SVM (ou *Naive Bayes*) pour la classification de sentiment, ça marche (très) bien

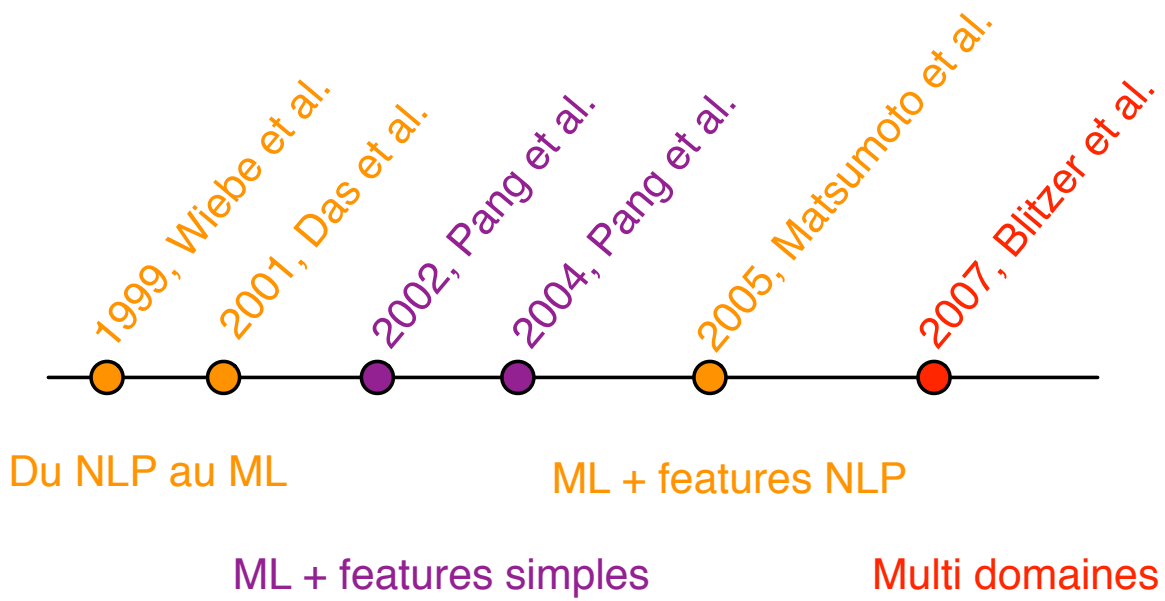
2004 Pang et al. : ... Mais pas si bien quand on passe au niveau phrase

2005 Matsumoto et al. (par exemple) : il faut des descripteurs issus du NLP pour attraper plus d'information

Vers un mélange des techniques NLP et apprentissage statistique (ML)

Historique ML Multi-domaines

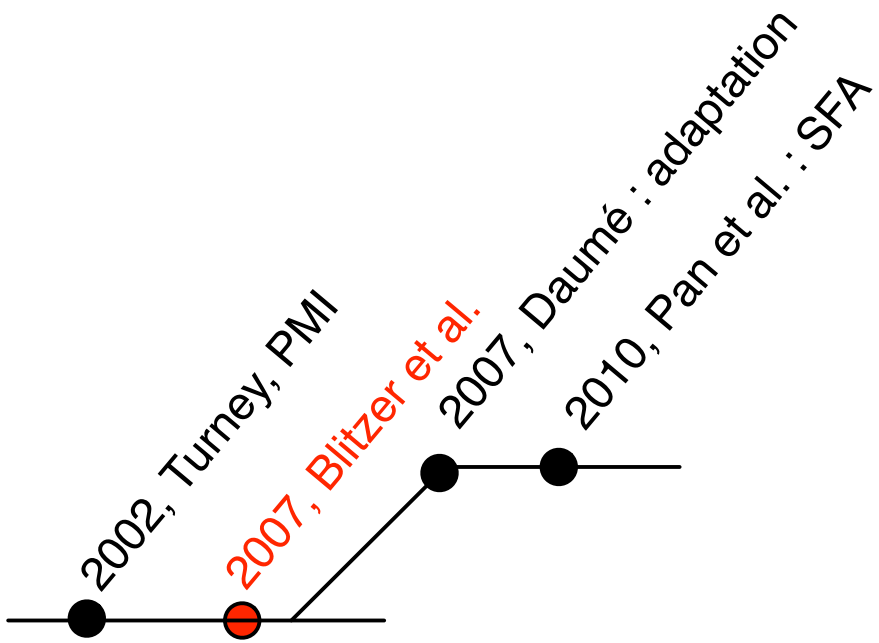
Les données supervisées ne correspondent en général pas au domaine visé... Besoin de généralisation et/ou d'adaptation au domaine cible.



Historique ML Multi-domaines

Les données supervisées ne correspondent en général pas au domaine visé... Besoin de généralisation et/ou d'adaptation au domaine cible.

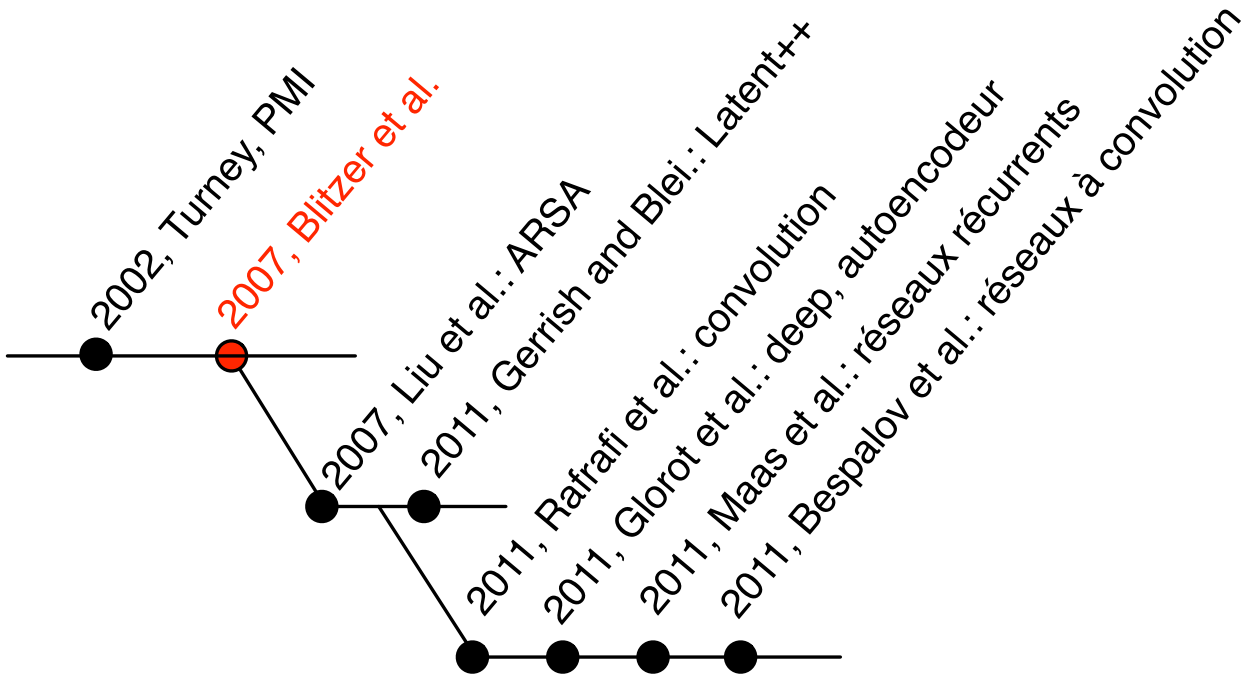
Modèle de transfert explicite : requiert (un peu) de données cibles étiquetées



Historique ML Multi-domaines

Les données supervisées ne correspondent en général pas au domaine visé... Besoin de généralisation et/ou d'adaptation au domaine cible.

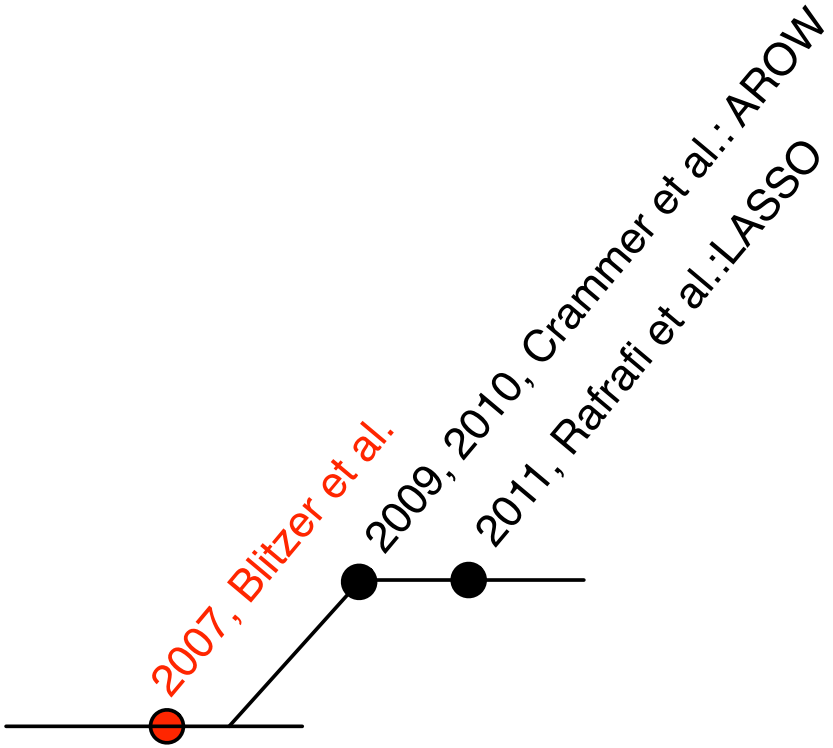
Construction d'un espace continu, analyse en variables latentes



Historique ML Multi-domaines

Les données supervisées ne correspondent en général pas au domaine visé... Besoin de généralisation et/ou d'adaptation au domaine cible.

Généralisation $\Leftrightarrow$ Régularisation



Historique Graphe, Social Network

- Une communauté à part orientée sur l'analyse de graphe (et le comptage...)
- Grosse BD + collecte, usage de machines sentiments existantes
- Diffusion & épidémiologie : modèles + apprentissage
- Un milieu hétérogène

2003 Kempe et al. Diffusion, épidémiologie & influence (cascades)

2006 Leskovec et al., *The Dynamics of Viral Marketing*

2010 Asur and Huberman, *Predicting the Future With Social Media*

2010 Leskovec et al. *Predicting Positive and Negative Links in Online Social Networks*

2011 Wang et al. Etiquetage de # HashTag

Plan

Le problème vu par Bing Liu

- Les *opinions* nous influencent au quotidien
- Nos croyances/perceptions sont conditionnées par la manière dont les autres voient le monde
- Pour prendre une décision, nous cherchons l'avis des autres

Dans le passé

- Individus : opinions des amis, de la famille
- Organisations : sondages, consultants...

Maintenant

- Expériences & opinions personnelles : blogs, tweets...
- Commentaires sur des articles, des produits

- Intérêts pour les particuliers et les entreprises

[Bing Liu] Un problème vraiment motivant

- Beaucoup d'applications
  - Un des sujets les plus abordés dans la littérature depuis 10 ans
  - Des dizaines d'entreprises nouvelles autour de ces problématiques
- Touche tous les aspects du NLP
- C'est difficile
- Nombreuses tâches mais quelques points clés

- Prendre en compte la structure du langage
- Identifier les acteurs et les relations

[Bing Liu] Les deux aspects du problème

Qu'est ce qu'une opinion (=sentiment) ?

- Définition structurée :
  - Cible(s) : entités (+certaines caractéristiques, aspects)
  - Sentiments : positif/négatif (neutre)
  - Emetteur : entité (personne)
  - A une échelle donnée : document/phrase
  - (Eventuellement à un temps donné)

Le résumé d'opinions

- L'opinion est subjective, on ne cherche pas l'opinion d'une personne (sauf VIP $\rightarrow$ RI classique)
- Besoin de plusieurs opinions $\Leftrightarrow$ besoin d'outils de résumé

[Bing Liu] Les deux aspects du problème

Qu'est ce qu'une opinion (=sentiment) ?

$\Longrightarrow$ Une opinion est un quintuplet $(e_j, a_{jk}, so_{ijkl}, h_i, t_l)$

- e_j entité cible
- a_{jk} caractéristique/sous-partie de l'entité e_j.
- so_{ijkl} sentiment de l'emetteur h_i sur la caractéristique a_{jk} de l'entité e_j au temps t_l. so_{ijkl} is +ve, −ve, or neu, ou un score.
- h_i emetteur de l'opinion.
- t_l datation de l'opinion.

[Bing Liu] Sur un exemple

Id : Abc123 on 5-1-2008

"I bought an iPhone a few days ago. It is such a nice phone. The touch screen is really cool. The voice quality is clear too. It is much better than my old Blackberry, which was a terrible phone and so difficult to type with its tiny keys. However, my mother was mad with me as I did not tell her before I bought the phone. She also thought the phone was too expensive, ..."

Identification des emetteurs, cibles, sentiments

Exemples de quintuplets :

- (iPhone, GENERAL, +, Abc123, 5-1-2008)
- (iPhone, touch_screen, +, Abc123, 5-1-2008)

[Bing Liu] Base pour les autres tâches

Tâche simple : prise en compte de certains morceaux du quintuplet
Tache complexe : résumé, ...

Apple iPhone 4 Smartphone 16 GB - AT&T - WCDMA (UMTS) / GSM - Black
$380 online ★★★★½ 618 reviews
June 2010 - Smartphone - Apple - iOS - AT&T - GSM - WCDMA - 5 megapixel camera - 3.5 inch screen - Qua

Reviews

Summary - Based on 618 reviews

1 | 2 | 3 | 4 stars | 5 stars

What people are saying

features	"Really enjoy all the features and apps available."
battery	"good but battery still needs improvement."
camera	"Takes pretty good pics."
screen	"Apple's display is awesome."
design	"Overall design, superb."
reception	"Easy to use ... Cons: No 3G networks in my area."
video	"Great video quality."

[Bing Liu] Deux types d'opinion

- Opinion *simple*
  - Opinion directe

    The touch screen is really cool.

  - Opinion indirecte

    After taking the drug, my pain has gone.

- Opinion comparative

  iPhone is better than Blackberry.

[Bing Liu] Autres émotions : détection, usage

Pas de consensus sur le groupe des émotions basiques.

- Emotions basiques : les autres émotions peuvent être exprimées comme une combinaison pondérée des premières
- Proposition [Parrott, 2001]
  - amour, joie, surprise, colère, tristesse, crainte
- Les sentiments/opinions sont particulièrement liés à certaines émotions : e.g., joie, colère

Il faut distinguer émotions et sentiments

W. Parrott, 2001, Psychology Press
Emotions in Social Psychology

[Bing Liu] Sentiments vs Emotions

- Evaluation rationnelle (contient une opinion mais pas d'émotion)

  The voice of this phone is clear

- Evaluation émotionnelle

  I love this phone

- Emotion sans sentiment particulier

  I am so surprised to see you

Propositions :

- Sentiment = Opinon
- Sentiment $\neq$ Subjectif $\neq$ Emotion
- Sentiment $\notin$ Subjectif (opinion factuelle)
- Emotion $\subset$ Subjectif
- Sentiment $\notin$ Emotion

Plusieurs problèmes

Une opinion est un quintuplet $(e_j, a_{jk}, so_{ijkl}, h_i, t_l)$
où :

- e_j entité cible : Name Entity Extraction
- a_{jk} caractéristique/sous-partie de l'entité e_j : Information Extraction
- so_{ijkl} sentiment de l'emetteur h_i sur la caractéristique a_{jk} de l'entité e_j au temps t_l. so_{ijkl} is +ve, −ve, or neu, ou un score : Sentiment classification
- h_i emetteur de l'opinion : Information Extraction
- t_l datation de l'opinion : Information Extraction

$\Rightarrow$ Résolution des co-références
$\Rightarrow$ Résolution des synonymies

Classification de sentiments

Lexiques de polarité (essentiellement) :

- General Inquirer, Stone et al. 1966 (4000 mots)
- MPQA, Wilson et al., 2003 (7000 mots, +/-, intensités)
- Liu lexicon, Hu and Liu, 2004 (7000 mots)
- Linguistic Inquiry and Word Count, Pennebaker et al. 2007 (2300 mots, 70 classes)
- SentiWordNet, Esuli et al. 2008, (>10000 termes, +/-, intensités)

	Opinion Lexicon	General Inquirer	SentiWordNet	LIWC
MPQA	33/5402 **(0.6%)**	49/2867 **(2%)**	1127/4214 **(27%)**	12/363 **(3%)**
Opinion Lexicon		32/2411 **(1%)**	1004/3994 **(25%)**	9/403 **(2%)**
General Inquirer			520/2306 **(23%)**	1/204 **(0.5%)**
SentiWordNet				174/694 **(25%)**
LIWC				

Christopher Potts, Sentiment Tutorial, 2011

Règles de décision

- Proposition de *templates*
- Apprentissage de patterns (par extraction sur critère statistique)
- Cascade de règles

SYNTACTIC FORM	EXAMPLE PATTERN
<subj> **passive-verb**	<subj> was satisfied
<subj> **active-verb**	<subj> complained
<subj> **active-verb dobj**	<subj> dealt blow
<subj> **verb infinitive**	<subj> appear to be
<subj> **aux noun**	<subj> has position
active-verb <dobj>	endorsed <dobj>
infinitive <dobj>	to condemn <dobj>
verb infinitive <dobj>	get to know <dobj>
noun aux <dobj>	fact is <dobj>
noun prep <np>	opinion on <np>
active-verb prep <np>	agrees with <np>
passive-verb prep <np>	was worried about <np>
infinitive prep <np>	to resort to <np>

Figure 2: Syntactic Templates and Examples of Patterns that were Learned

PATTERN	FREQ	%SUBJ
<subj> was asked	11	100%
<subj> asked	128	63%
<subj> is talk	5	100%
talk of <np>	10	90%
<subj> will talk	28	71%
<subj> put an end	10	90%
<subj> put	187	67%
<subj> is going to be	11	82%
<subj> is going	182	67%
was expected from <np>	5	100%
<subj> was expected	45	42%
<subj> is fact	38	100%
fact is <dobj>	12	100%

Figure 3: Patterns with Interesting Behavior

E. Riloff and J. Wiebe, EMNLP 2003

Learning Extraction Patterns for Subjective Expressions

Règles de décision

- Proposition de *templates*
- Apprentissage de patterns (par extraction sur critère statistique)
- Cascade de règles

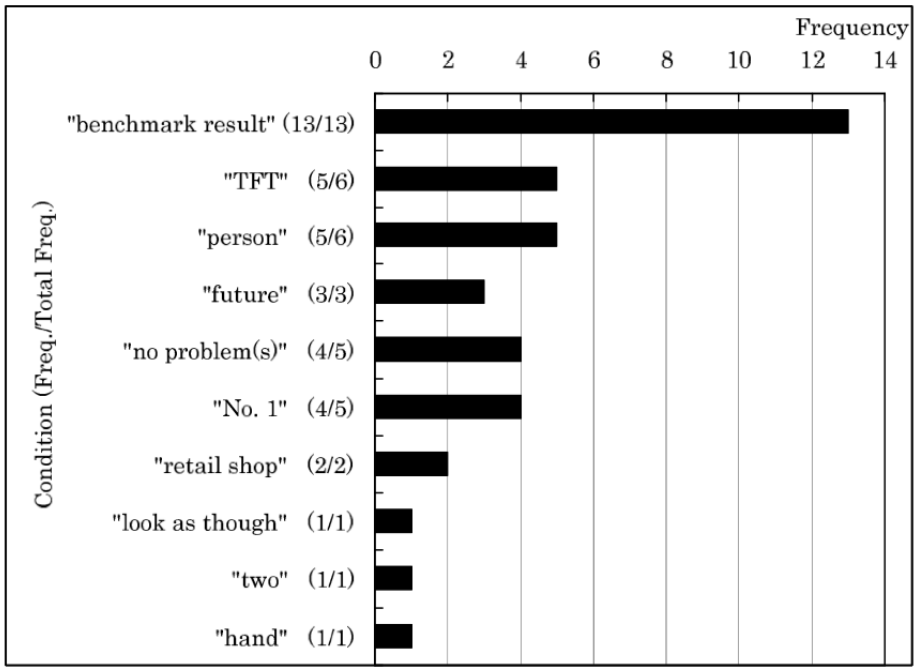


Figure 2: Classification Rules for Cellular Phone A

Morinaga et al., ACM 2002

Mining Product Reputations on the Web

Règles et lexiques

Exemple d'algorithme

1. Procedure **wordOrientation**(*word*, *feature*, *sentence*)
2. if *word* is a Negation word then
3. *orientation* = apply Negation Rules;
4. mark words in *sentence* used by Negation rules
5. elseif *word* is a TOO word then
6. *orientation* = apply TOO Rules;
7. mark words in *sentence* used by TOO rules
8. else
9. *orientation* = orientation of *word* in *opinionWord_list*
10. endif
11. $orientation = \frac{orientation}{dis(word, feature)}$,

Figure 2: Predicting the orientations of opinions on product features

Ding, Liu, Yu, ACM 2004

A holistic lexicon-based approach to opinion mining

1. **Algorithm OpinionOrietation**()
2. for each sentence s_i that contains a set of features do
3. *features* = features contained in s_i;
4. for each feature f_j in *features* do
5. *orientation* = 0;
6. if feature f_j is in the "but" clause then
7. *orientation* = apply the "but" clause rule
8. else remove "but" clause from s_i if it exists;
9. for each unmarked opinion word *ow* in s_i do
10. // *ow* can be a TOO word or Negation word as well
11. *orientation* += **wordOrientation**(*ow*, f_j, s_i);
12. endfor
13. endif
14. if *orientation* > 0 then
15. f_j's orientation in s_i = 1
16. else if *orientation* < 0 then
17. f_j's orientation in s_i = -1
18. else
19. f_j's orientation in s_i = 0
20. endif
21. endif
22. if f_j is an adjective then
23. (f_j).*orientation* += f_j's orientation in s_i;
24. else let o_{ij} is the nearest adjective word to f_j, in s_i;
25. (f_j, o_{ij}).*orientation* += f_j's orientation in s_i;
26. endif
27. endfor
28. endfor;

POS, Patterns, lexiques...

Caractérisation des sentiments sur une caractéristique d'un produit

- Extraction de caractéristiques (POS, Entity extract...)
- Prise en compte des comparaisons
- Utilisation de lexique(s) de sentiments
- A base de patterns de sentiments
- ...

Bonnes performances (!)

Yi, Nasukawa, Bunescu, Niblack, IEEE ICDM 2003

Sentiment Analyzer : Extracting Sentiments about a Given Topic using Natural Language Processing Techniques

Exemple de résultats

HP-Subj		HP-Subj w/Patterns	
Recall	*Precision*	*Recall*	*Precision*
32.9	91.3	40.1	90.2

Table 1: Bootstrapping the Learned Patterns into the High-Precision Sentence Classifier

seems to be <dobj>	I am pleased that there now **seems to be** broad political consensus …
underlined <dobj>	Jiang's subdued tone . . . **underlined** his desire to avoid disputes …
pretext of <np>	On the **pretext of** the US opposition . . .
atmosphere of <np>	Terrorism thrives in an **atmosphere of** hate . . .
<subj> reflect	These are fine words, but they do not **reflect** the reality . . .
to satisfy <dobj>	The pictures resemble an attempt **to satisfy** a primitive thirst for revenge . . .
way with <np>	…to ever let China use force to have its **way with** …
bring about <np>	"Everything must be done by everyone to **bring about** de-escalation," Mr Chirac added.
expense of <np>	at the **expense of** the world's security and stability
voiced <dobj>	Khatami …**voiced** Iran's displeasure.
turn into <np>	…the surging epidemic could **turn into** "a national security threat," he said.

Table 2: Examples of Learned Patterns Used by HP-Subj and Sample Matching Sentences

E. Riloff and J. Wiebe, EMNLP 2003

Learning Extraction Patterns for Subjective Expressions

Exemple de résultats

- Un taux de reconnaissance entre 60 et 75% sur la classification de polarité (sans intégration ML)
- Des analyses fines : au niveau phrase, avec extraction d'entités...
- Des approches généralistes (non supervisées) transposables immédiatement à n'importe quel domaine

Apprentissage de lexique

- Avantages :
  - intégration d'informations spécifiques au domaine
  - plus de mots $\Longrightarrow$ plus robuste
- Algorithme intuitif
  - Initialiser le processus avec des mots évidents ('good', 'poor')
  - Trouver des mots similaires :
    - Utiliser des conjonctions "and" et "but"
    - Utiliser les cooccurrences de mots dans un document
    - Utiliser les synonymes/antonymes de WordNet

Hatzivassiloglou and McKeown 1997

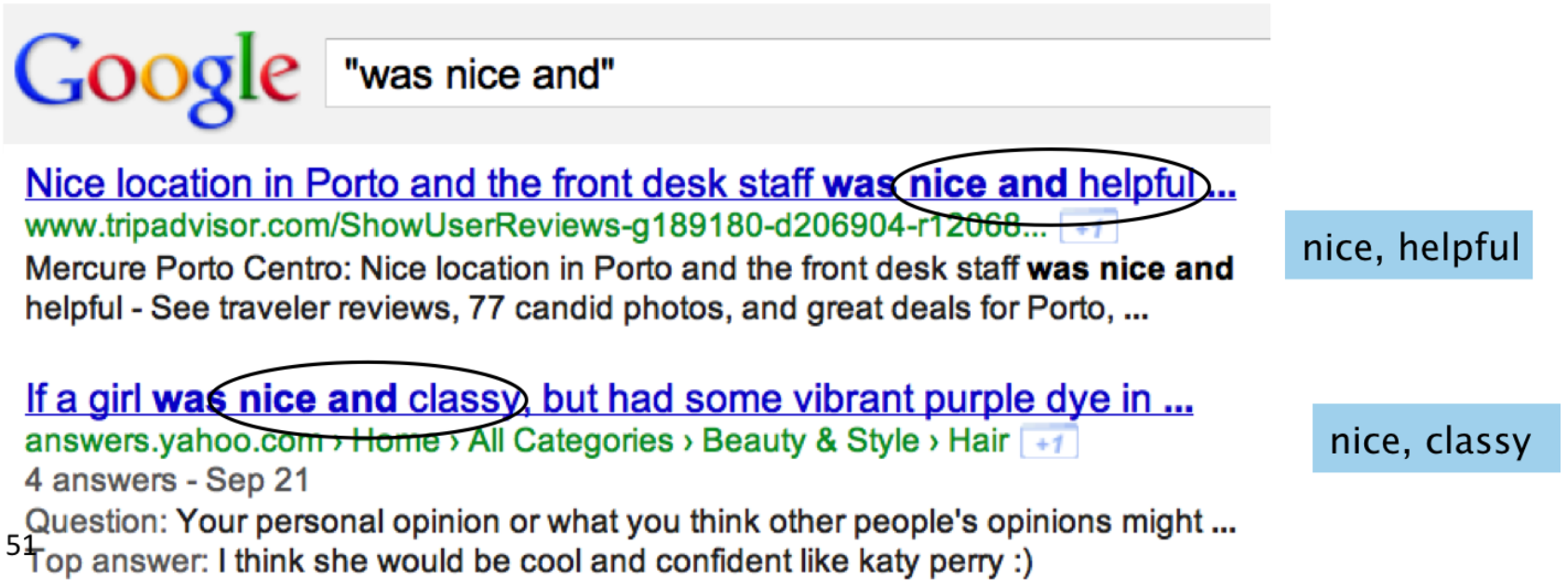
- Les adjectifs liés par **and** ont la même polarité
  - Fair **and** legitimate, corrupt **and** brutal
  - fair **and** brutal, corrupt **and** legitimate
- Les adjectifs liés par **but** non
  - fair **but** brutal
- Initialisation : 1336 adjectifs ($\approx$ 50/50 positifs/négatifs)

Hatzivassiloglou McKeown 1997

Predicting the Semantic Orientation of Adjectives

Hatzivassiloglou and McKeown 1997

Expension :

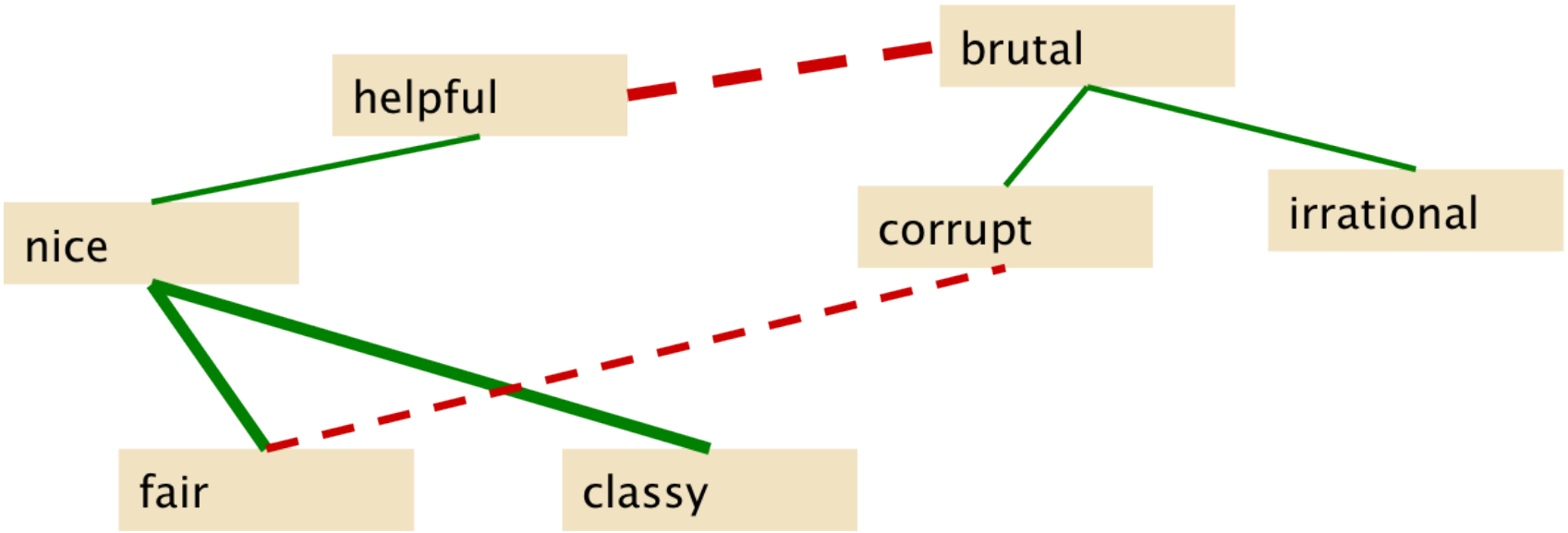


Utilisation de ressources externes (comme wikipedia maintenant)...

Hatzivassiloglou McKeown 1997

Predicting the Semantic Orientation of Adjectives

Hatzivassiloglou and McKeown 1997



+ clustering

Hatzivassiloglou McKeown 1997

Predicting the Semantic Orientation of Adjectives

Hatzivassiloglou and McKeown 1997

Résultats :

- Positive
  bold decisive disturbing generous good honest important large mature patient peaceful positive proud sound stimulating straightforward strange talented vigorous witty. . .
- Negative
  ambiguous cautious cynical evasive harmful hypocritical inefficient insecure irrational irresponsible minor outspoken pleasant reckless risky selfish tedious unsupported vulnerable wasteful. . .

Hatzivassiloglou McKeown 1997

Predicting the Semantic Orientation of Adjectives

Apprentissage de lexique : Hu et Liu 2004

En version raffinée : plus de traitement, plus de documents (seed) mais moins de ressources externes.

- Filtrage fréquentiel + POS tagging
- Pour chaque mot fréquent
  - Extraire les adjectifs
  - Annoter les adjectifs en fonction du tag document

Hu and Liu, AAAI NCAI 2004

Mining opinion features in customer reviews

PMI [Turney, 2002], slide Dan Jurafsky

1. Séparer le corpus en *phrases* (groupes de mots respectant des patterns POS)
2. Apprendre la polarité des phrases
3. Agréger les scores de phrases (moyenne) pour trouver un doc.

Turney, ACL 2002

Thumbs Up or Thumbs Down ? Semantic Orientation Applied to Unsupervised Classification of Reviews

PMI [Turney, 2002], slide Dan Jurafsky

1. Séparer le corpus en *phrases* (groupes de mots respectant des patterns POS)
2. Apprendre la polarité des phrases
3. Agréger les scores de phrases (moyenne) pour trouver un doc.

Apprendre la polarité :

phrases + co-occurent avec excellent

phrases − co-occurent avec poor

First Word	Second Word	Third Word (not extracted)
JJ	NN or NNS	anything
RB, RBR, RBS	JJ	Not NN nor NNS
JJ	JJ	Not NN or NNS
NN or NNS	JJ	Nor NN nor NNS
RB, RBR, or RBS	VB, VBD, VBN, VBG	anything

Turney, ACL 2002

Thumbs Up or Thumbs Down ? Semantic Orientation Applied to Unsupervised Classification of Reviews

PMI [Turney, 2002], slide Dan Jurafsky

1. Séparer le corpus en *phrases* (groupes de mots respectant des patterns POS)
2. Apprendre la polarité des phrases
3. Agréger les scores de phrases (moyenne) pour trouver un doc.

Apprendre la polarité :

phrases + co-occurent avec excellent

phrases − co-occurent avec poor

First Word	Second Word	Third Word (not extracted)
JJ	NN or NNS	anything
RB, RBR, RBS	JJ	Not NN nor NNS
JJ	JJ	Not NN or NNS
NN or NNS	JJ	Nor NN nor NNS
RB, RBR, or RBS	VB, VBD, VBN, VBG	anything

Comment mesurer la co-occurence ?

Turney, ACL 2002

Thumbs Up or Thumbs Down ? Semantic Orientation Applied to Unsupervised Classification of Reviews

PMI [Turney, 2002], slide Dan Jurafsky

Mutual Information :

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left(\frac{p(x,y)}{p(x)\,p(y)} \right),$$

Interprétation : mesurer la dépendance mutuelle entre deux variables aléatoire, une forme de similarité entre X et Y.

Pointwise Mutual Information :

$$PMI(X,Y) = \log \left(\frac{p(x,y)}{p(x)\,p(y)} \right)$$

Interprétation : est ce que les évènements x et y co-occurent plus que s'ils étaient indépendants ? (i.e. $PMI = 0$ en cas d'indépendance)

PMI [Turney, 2002], slide Dan Jurafsky

Estimation des probes par requête Altavista :

- $P(word)$ estimé par : $\text{hits}(word)/N$
- $P(word_1, word_2)$ par : $\text{hits}(word_1\ NEAR\ word_2)/N^2$

Polarité dune phrase

$$Pol(s) = PMI(s,''\ excellent'') - PMI(s,''\ poor'')$$

$$Pol(s) = \log\left(\frac{\text{hits}(s\ NEAR\ ''excellent'')\text{hits}(''poor'')}{\text{hits}(s\ NEAR\ ''poor'')\text{hits}(''excellent'')}\right)$$

Utilisation de ressources *pré-wikipedia*, précurseur des analyses en variables latentes sur wikipédia.

PMI [Turney, 2002] : Résultats

Revues positives

Phrase	POS tags	Polarity
online service	JJ NN	2.8
online experience	JJ NN	2.3
direct deposit	JJ NN	1.3
local branch	JJ NN	0.42
...		
low fees	JJ NNS	0.33
true service	JJ NN	-0.73
other bank	JJ NN	-0.85
inconveniently located	JJ NN	-1.5
Average		0.32

Revues négatives :

Phrase	POS tags	Polarity
direct deposits	JJ NNS	5.8
online web	JJ NN	1.9
very handy	RB JJ	1.4
...		
virtual monopoly	JJ NN	-2.0
lesser evil	RBR JJ	-2.3
other problems	JJ NNS	-2.8
low funds	JJ NNS	-6.8
unethical practices	JJ NNS	-8.5
Average		-1.2

Ressources externes : possibilité de trouver des groupes négatifs dans les revues positives !

PMI [Turney, 2002] : Résultats

- 410 reviews from Epinions
  - 170 (41%) negative
  - 240 (59%) positive
  - 106,580 phrases
- Majority class baseline : 59%
- Turney algorithm : 74%
- Only 66% on movie reviews (average is not a good solution...)

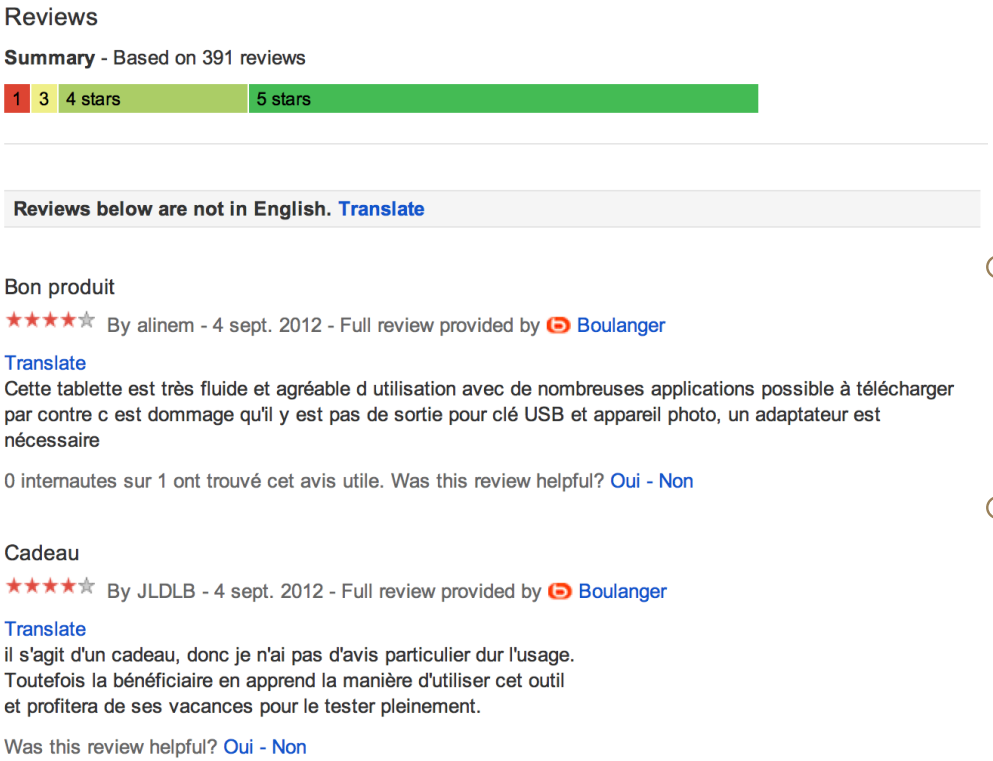
Key points :

- Phrases rather than words
- Learns domain-specific information
- Fast & require no labeled dataset

Domain of Review	Accuracy
Automobiles	84.00 %
Honda Accord	83.78 %
Volkswagen Jetta	84.21 %
Banks	80.00 %
Bank of America	78.33 %
Washington Mutual	81.67 %
Movies	65.83 %
The Matrix	66.67 %
Pearl Harbor	65.00 %
Travel Destinations	70.53 %
Cancun	64.41 %
Puerto Vallarta	80.56 %
All	74.39 %

Plan

Données étiquetées



Reviews

Summary - Based on 391 reviews

1 | 3 | 4 stars | 5 stars

Reviews below are not in English. Translate

Bon produit

★★★★☆ By alinem - 4 sept. 2012 - Full review provided by Boulanger

Translate

Cette tablette est très fluide et agréable d utilisation avec de nombreuses applications possible à télécharger par contre c est dommage qu'il y est pas de sortie pour clé USB et appareil photo, un adaptateur est nécessaire

0 internautes sur 1 ont trouvé cet avis utile. Was this review helpful? Oui - Non

Cadeau

★★★★☆ By JLDLB - 4 sept. 2012 - Full review provided by Boulanger

Translate

il s'agit d'un cadeau, donc je n'ai pas d'avis particulier dur l'usage.
Toutefois la bénéficiaire en apprend la manière d'utiliser cet outil
et profitera de ses vacances pour le tester pleinement.

Was this review helpful? Oui - Non

- Les *User Generated Contents* sont non seulement valorisables directement...
- Mais ils sont régulièrement étiquetés !

Possibilité de faire de l'apprentissage supervisé à moindre coût

Corpus usuels

Coté NLP, les corpus sont de l'ordre de la centaine de documents (par thème et par catégorie).

2002 Pang and Lee, Movie reviews : 2000 documents

2007 Blitzer et al. Amazon : 4x 2000 documents

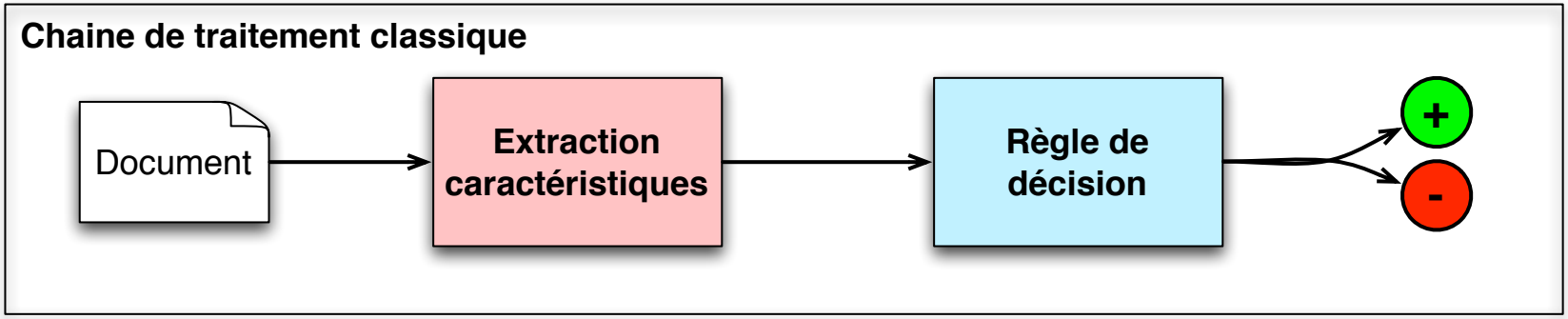
2009 Whitehead : 10x [200, 2000] documents

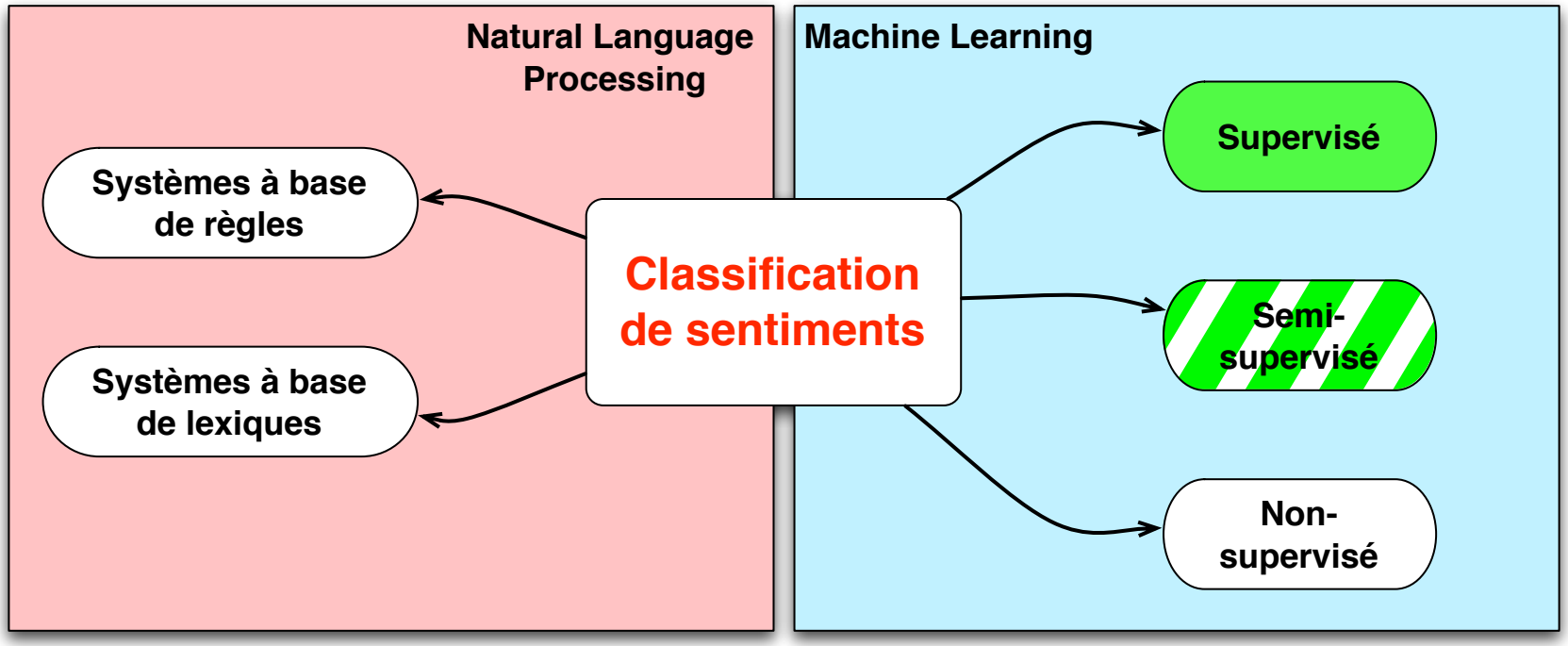
... Multiplication des corpus et augmentation de la taille

- Big Amazon : 20x [5k, 50k] documents
- Big IMDB : 2x 25k documents
- Trip advisor : 2x 50k documents
- ...

La littérature est centrée (principalement) sur la classification en polarité

Chaine de traitement (hors séquences)





Descripteurs [Représentation vectorielle]

- Unigrammes + négation [Das et al. 2001]
  Add `NOT_` to every word between negation and following punctuation :
  `didn't like this movie , but I`
  ⇒ `didn't NOT_like NOT_this NOT_movie but I`
- Unigrammes, U+Bigrammes, sel. POS (adjectifs), codage position des mots... [Pang and Lee, 2002]
- Trigrammes, sous-séquences, sel. POS [Dave et al. 2003]
- Sous-séquences, arbre syntaxique [Matsumoto et al. 2005]

Les conclusions/comparaisons doivent comparer les algorithmes ET les descripteurs : pas facile dans une littérature dense.

Descripteurs [Représentation vectorielle]

- Unigrammes + négation [Das et al. 2001]
- Unigrammes, U+Bigrammes, sel. POS (adjectifs), codage position des mots... [Pang and Lee, 2002]
- Trigrammes, sous-séquences, sel. POS [Dave et al. 2003]

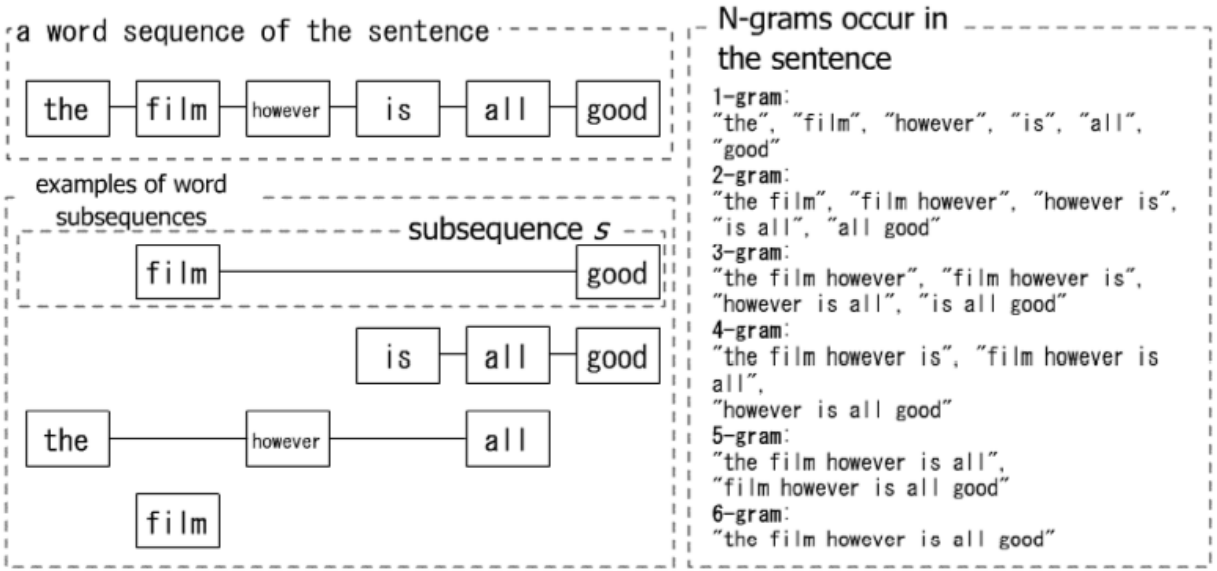


Fig. 2. A word sequence of a sentence "*The film however is all good*" and exampl subsequences

- Sous-séquences, arbre syntaxique [Matsumoto et al. 2005]

usions/comparaisons doivent comparer les algorithmes ET

Descripteurs [Représentation vectorielle]

- Unigrammes + négation [Das et al. 2001]
- Unigrammes, U+Bigrammes, sel. POS (adjectifs), codage position des mots... [Pang and Lee, 2002]
- Trigrammes, sous-séquences, sel. POS [Dave et al. 2003]
- Sous-séquences, arbre syntaxique [Matsumoto et al. 2005]

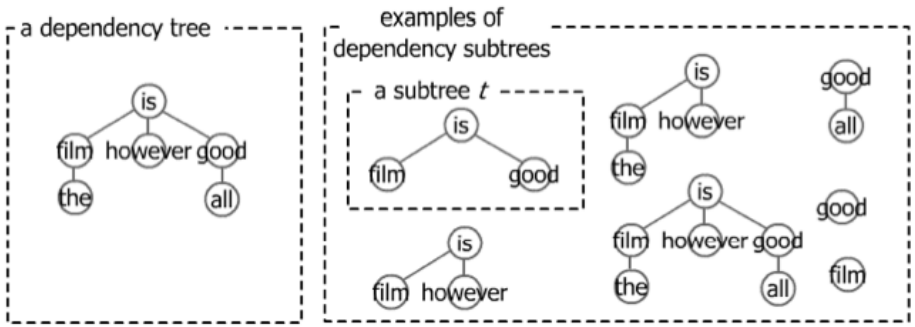


Fig. 3. A dependency tree of a sentence "*The film however is all good*" and examples of subtrees

Les conclusions/comparaisons doivent comparer les algorithmes ET les descripteurs : pas facile dans une littérature dense.

Descripteurs & pre-processing

2008 Pang and Lee, Lemmatisation : l'intérêt varie selon

- Les expériences, la langue
- La taille de base d'apprentissage

2006 Kennedy and Inkpen, Quantifieurs émotionnels : codage en sacs de mots, modification des poids

2005 Wilson et al., EMNLP, Discussion sur la complexité/subtilité de prendre en compte la négation

- Potts : proposition de détection des *smileys*

```
[<>]?                           # optional hat/brow
[:;=8]                          # eyes
[\-o\*\']?                      # optional nose
[\)\]\(\[dDpP/\:\}\{@\|\\]      # mouth
|                               #### reverse orientation
[\)\]\(\[dDpP/\:\}\{@\|\\]      # mouth
[\-o\*\']?                      # optional nose
[:;=8]                          # eyes
[<>]?                           # optional hat/brow
```

Représentation en (très) grande dimension

Extraction brute : pas de sélection POS
Filtre Fréquentiel : ≥ 2 itérations dans la base

Corpus	Tailles des corpus (N)	Long. moy. des revues	Vocabulaire (V)		
			Unigr.	U+Bigr.	U+B+SSéq.
Books	2000	240	10536	45750	78664
Dvd	2000	235	10392	48955	89313
Electronics	2000	154	5611	30101	49994
Kitchen	2000	133	5314	26156	40773
Movie Rev.	2000	745	26420	148765	308564

Table : Descriptions de cinq corpus classiques (IMDB, Amazon).

Représentation en (très) grande dimension

- Coût du dictionnaire (mémoire et temps de calcul)
  - Filtrage POS, ajout de connaissances...
- Problème d'optimisation
  - Problèmes mal posés : quelques milliers de documents, dizaines (ou centaines) de milliers de caractéristiques
  - Risque (très) fort de sur-apprentissage
  - Variabilité dans les résultats des algorithmes stochastiques
- Problème d'ingénierie
  - Base trop grosse ⇒ développement plus ardu ⇒ risque de bug/biais

Codage de l'information

Un consensus depuis [Pang et al. 2002] :

- tf-idf $\Rightarrow$ baisse de performance (et de stabilité)
- codage présentiel : 0/1 sans prise en compte de la fréquence

Quelques propositions alternatives :

- Δ-tf : différence de fréquences entre les classes $+$ et $-$
  $\Leftrightarrow$ insérer une sorte de naïve Bayes au niveau du codage...

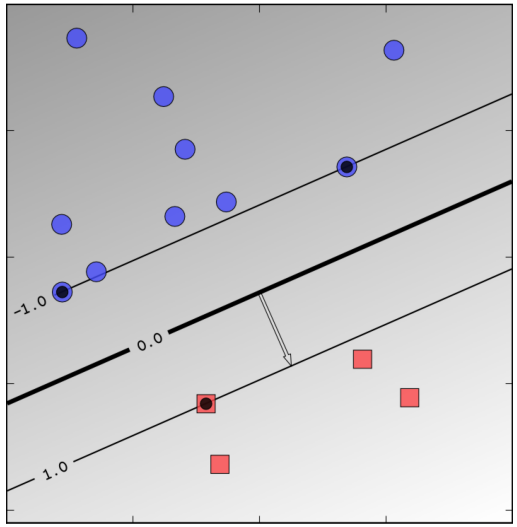
Modèles : SVM, MaxEnt, Naive Bayes

- Une fois les données $X = \{\mathbf{x}_i\}_{i=1,\ldots,n}$ codées en vecteurs $\mathbf{x} = \{x_j\}_{j=1,\ldots,d}$
- Trouver $f(\mathbf{x})$

SVM linéaires, codage des étiquettes en $y_i \in \{-1, 1\}$:

$$f(\mathbf{x}) = \langle \mathbf{x}, \mathbf{w} \rangle$$

$$w^\star = \arg\min_w \left(\sum_i (1 - y_i f(\mathbf{x}_i))_+ + \lambda \|w\|^2 \right)$$



Modèles : SVM, MaxEnt, Naive Bayes

- Une fois les données $X = \{\mathbf{x}_i\}_{i=1,\ldots,n}$ codées en vecteurs $\mathbf{x} = \{x_j\}_{j=1,\ldots,d}$
- Trouver $f(\mathbf{x})$

Naive Bayes (optimisation d'un loi multinomiale)
Hypothèse d'indépendance des mots : $p(\mathbf{x}|C) = \prod_j p(w_j|C)$
Optimisation de la log-vraisemblance de la base documentaire :

$$L(\{\mathbf{x}_i \in C\}, \Theta) = \sum_{i=1}^{n} \sum_{j=1}^{d} x_i^j \log P(w_j|\Theta)$$

$$P(w_j|\Theta) = \frac{\sum_{\mathbf{x}_i \in X_C} x_i^j}{\sum_{\mathbf{x}_i \in X_C} \sum_{k=1}^{d} x_i^k} = \frac{nw_j}{nw}, \quad \text{stabilisé : } P(w_j|\Theta) = \frac{nw_j + \alpha}{nw + \alpha d}$$

Modèles : SVM, MaxEnt, Naive Bayes

- Une fois les données $X = \{\mathbf{x}_i\}_{i=1,\ldots,n}$ codées en vecteurs $\mathbf{x} = \{x_j\}_{j=1,\ldots,d}$
- Trouver $f(\mathbf{x})$

Maximum d'entropie / régression logistique

Modèles linéaires probabilistes : $p(\mathbf{x}_i|C) = \frac{1}{1+\exp(-(\mathbf{x}_i\mathbf{w}+b))}$

Apprentissage par maximum de vraisemblance (algo=descente de gradient) :

$$\frac{\partial}{\partial w_j} L = \sum_{i=1}^{N} x_{ij} \left(y_i - \frac{1}{1 + exp(-(\mathbf{x}_i\mathbf{w} + b))} \right)$$

$$\frac{\partial}{\partial b} L = \sum_{i=1}^{N} \left(y_i - \frac{1}{1 + \exp(-(\mathbf{x}_i\mathbf{w} + b))} \right)$$

Implémentations

En terme de ressources, les SVM possèdent un avantage : de nombreuses implémentations très efficaces :

- SVM light + variantes [Joachims]
  `http://svmlight.joachims.org`
- Pegasos : résolution efficace dans le primal [Shalev-Shwartz et al.]
  `http://www.cs.huji.ac.il/~shais/code/index.html`
- Liblinear : descente de coordonnées, le plus intéressant pour moi... [CJ Lin et al.]
  `http://www.csie.ntu.edu.tw/~cjlin/liblinear/`

Lien annexe : Mallet [McCallum], tous les outils pour le processing de corpus...
`http://mallet.cs.umass.edu`

Résultats [Pang et al. 2002]

- Des performances assez proches suivant les algorithmes ([Pang et al. 2002]) : SVM un peu meilleur, Naive Bayes un peu moins bon.
- Pas de différence significative sur les codages *riches*

	Features	# of features	frequency or presence?	NB	ME	SVM
(1)	unigrams	16165	freq.	**78.7**	N/A	72.8
(2)	unigrams	"	pres.	81.0	80.4	**82.9**
(3)	unigrams+bigrams	32330	pres.	80.6	80.8	**82.7**
(4)	bigrams	16165	pres.	77.3	**77.4**	77.1
(5)	unigrams+POS	16695	pres.	81.5	80.4	**81.9**
(6)	adjectives	2633	pres.	77.0	**77.7**	75.1
(7)	top 2633 unigrams	2633	pres.	80.3	81.0	**81.4**
(8)	unigrams+position	22430	pres.	81.0	80.1	**81.6**

Figure 3: Average three-fold cross-validation accuracies, in percent. Boldface: best performance for a given setting (row). Recall that our baseline results ranged from 50% to 69%.

Résultats [Matsumoto et al. 2005]

- Codage *riche* $\Rightarrow \nearrow$ reconnaissance

Table 2. Results for dataset 1

Features	Acurracy(%)		
	test1	test2	test3
Pang et al. [1]	82.9	N/A	N/A
Mullen et al. [10]	**84.6**	N/A	N/A
word unigram (= *uni*)	83.0	83.7	83.0
lemma unigram (= uni_l)	82.8	83.8	83.2
word bigram (= *bi*)	79.6	80.4	80.1
lemma bigram (= bi_l)	80.4	80.9	80.7
$uni + bi$	83.8	84.6	84.0
$uni + bi_l$	83.6	84.2	83.5
$uni_l + bi$	**84.4**	84.8	**84.6**
$uni_l + bi_l$ (= *bow*)	84.0	**84.9**	84.2
$bow + seq$	84.1	85.3	84.9
$bow + seq_l$	84.4	85.7	84.9
$bow + dep$	86.6	87.6	87.5
$bow + dep_l$	**87.3**	**88.3**	**88.0**
$bow + seq + dep$	86.2	87.2	87.2
$bow + seq + dep_l$	87.0	87.5	87.5
$bow + seq_l + dep$	86.5	87.5	87.0

Table 3. Results for dataset 2

Features	Acurracy(%)		
	test1	test2	test3
Pang et al. [2]	**87.1**	N/A	N/A
word unigram (= *uni*)	87.1	88.1	87.0
lemma unigram (= uni_l)	86.4	86.9	85.9
word bigram (= *bi*)	84.2	85.3	85.1
lemma bigram (= bi_l)	84.3	85.2	84.7
$uni + bi$ (= *bow*)	**88.1**	**88.8**	**88.0**
$uni + bi_l$	87.8	88.6	87.8
$uni_l + bi$	87.3	88.2	87.3
$uni_l + bi_l$	87.7	88.3	87.9
$bow + seq$	88.2	89.4	88.3
$bow + seq_l$	88.5	89.8	88.5
$bow + dep$	92.4	**93.7**	92.7
$bow + dep_l$	92.8	**93.7**	92.9
$bow + seq + dep$	92.6	93.5	92.8
$bow + seq + dep_l$	**92.9**	**93.7**	**93.2**
$bow + seq_l + dep$	92.6	93.2	93.0

Analyses qualitative des résultats

- Tri des poids w_j

Table 4. Examples of patterns with their weights in the weight vector

the type of the pattern	weight	pattern
Word unigram (*uni*)	1.0618	*hilarious*
	0.6221	*masterpiece*
	-1.4265	*stern*
	-0.3749	*movie*
	-1.9937	*bad*
Word bigram (*bi*)	8.0561	*pull off*
	0.8565	*one of*
	-0.1150	*little life*
	-0.3330	*bad movie*
Word subsequence (*seq*)	-7.0200	*little-life*
	0.2340	*not-only-also*
	0.3497	*good-film*
	0.3243	*film-good*
	-0.5828	*bad-movie*

Analyses qualitative des résultats

- Tri des poids w_j

Classification de critiques de livres :

Unigrams		Bigrams	
top −	top +	top −	top +
disappointing	summer	not_recommend	a_must
useless	terrific	best_answer	I_sure
valuation	bible	save_your	really_enjoyed
outdated	displayed	too_much	read_from
shallow	editions	skip_this	excellent_book
poorly	refreshing	reference	wow
wasted	concise	disappointing	loved_this
unrealistic	profession	shallow	gift
norm	bike	unless_you	good_reference
hype	shines	way_too	enjoyed_this
incorrect	coping	was_looking	very_pleased
burn	blended	nothing_new	it_helped
boring	humorous	your_money	terrific
york	lighten	very_disappointing	great_!
hated	amazed	first_trip	all_ages

Prise de décision & modèle statistique

Exemple de prise de décision dans le détail avec des bigrammes :

Modèle linéaire et codage présentiel : $f(\mathbf{x}_i) = \sum_{j \in \mathbf{x}_i} w_j$

```
 This book is great
Score : 0.20597761553808144
   this_book : -0,0569
   be : 0,0760
   book_be : 0,0195
   great : 0,2507
   book : -0,0183
   be_great : 0,0057
   this : -0,0707
```

```
 I felt disappointed
Score :-0.07434288371014953
   I : 0,0189
   disappointed : -0,0946
   I_feel : 0,0062
   feel : -0,0048
```

Prise de décision & modèle statistique

Exemple de prise de décision dans le détail avec des bigrammes :

```
 This book is not easy to
read
Score : -0.02990834226477354
   this_book : -0,0569
   read : 0,0866
   book_be : 0,0195
   easy : 0,1688
   this : -0,0707
   be_not : -0,0661
   not : -0,3336
   be : 0,0760
   not_easy : 0,0348
   easy_to : 0,1046
   book : -0,0183
   to : 0,1163
   to_read : -0,0909
```

Quelques exemples de négations

```
 could couldn't should
shouldn't
   should : -0,0245
   could : -0,0272
   shouldn't : 0,0026
   couldn't : 0,0293
```

Prise de décision & modèle statistique

Exemple de prise de décision dans le détail avec des bigrammes :

Imprécision au niveau de la phrase :

I don't like this book

Score : -0.3866063036824649
- this_book : -0,0569
- I_don't : -0,1249
- don't : -0,1555
- I : 0,0189
- don't_like : 0,0391
- book : -0,0183
- like : -0,0706
- this : -0,0707
- like_this : 0,0522

I like this book

Score : -0.06739329361414403
- this_book : -0,0569
- I : 0,0189
- book : -0,0183
- like : -0,0706
- this : -0,0707
- like_this : 0,0522
- I_like : 0,0780

Exemples d'usage des CRF

- Proposition de CRF sur des pattern NLP pour l'identification des cibles/émetteurs d'opinions [Choi et al.].
- CRF sur l'arbre syntaxique de la phrase [Nakagawa et al.]

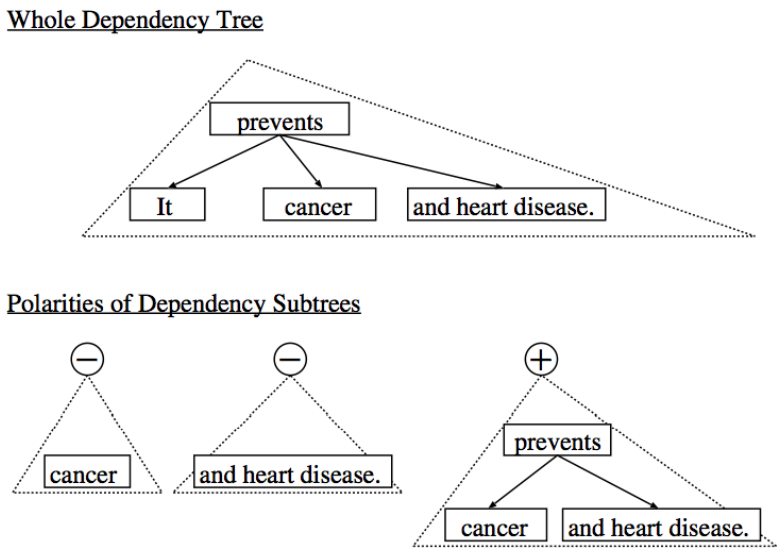


Figure 1: Polarities of Dependency Subtrees

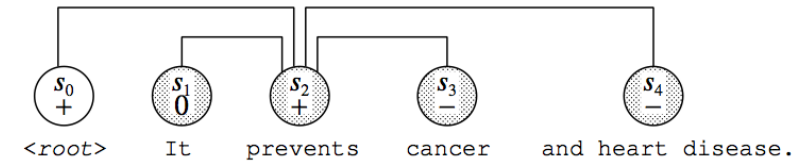


Figure 2: Probabilistic Model based on Dependency Tree

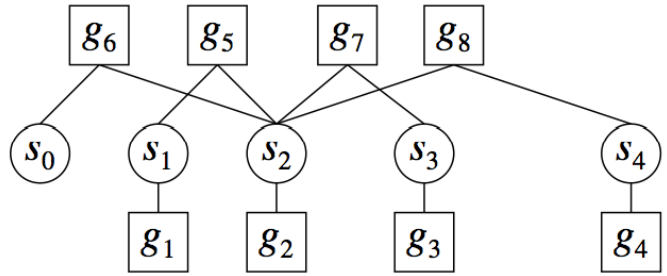


Figure 3: Factor Graph

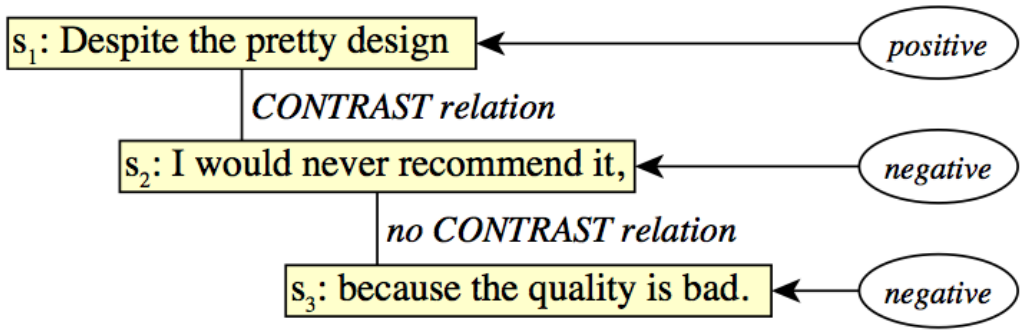
Choi, Cardie, Riloff, Patwardhan, EMNLP 2005
Identifying Sources of Opinions with Conditional Random Fields and Extraction Patterns

Nakagawa, Inui, Kurohashi, ACL 2010
Dependency Tree-based Sentiment Classification using CRFs with Hidden Variables

CRF/Logic Markov Network

Revenir au niveau de la phrase avec des systèmes à état entre les phrases (ou portion de phrases).

- CRF enrichis pour la classification de sentiments au niveau phrase → capture de la dynamique du document [Zhao et al.].
- LMN sur des portions de phrases

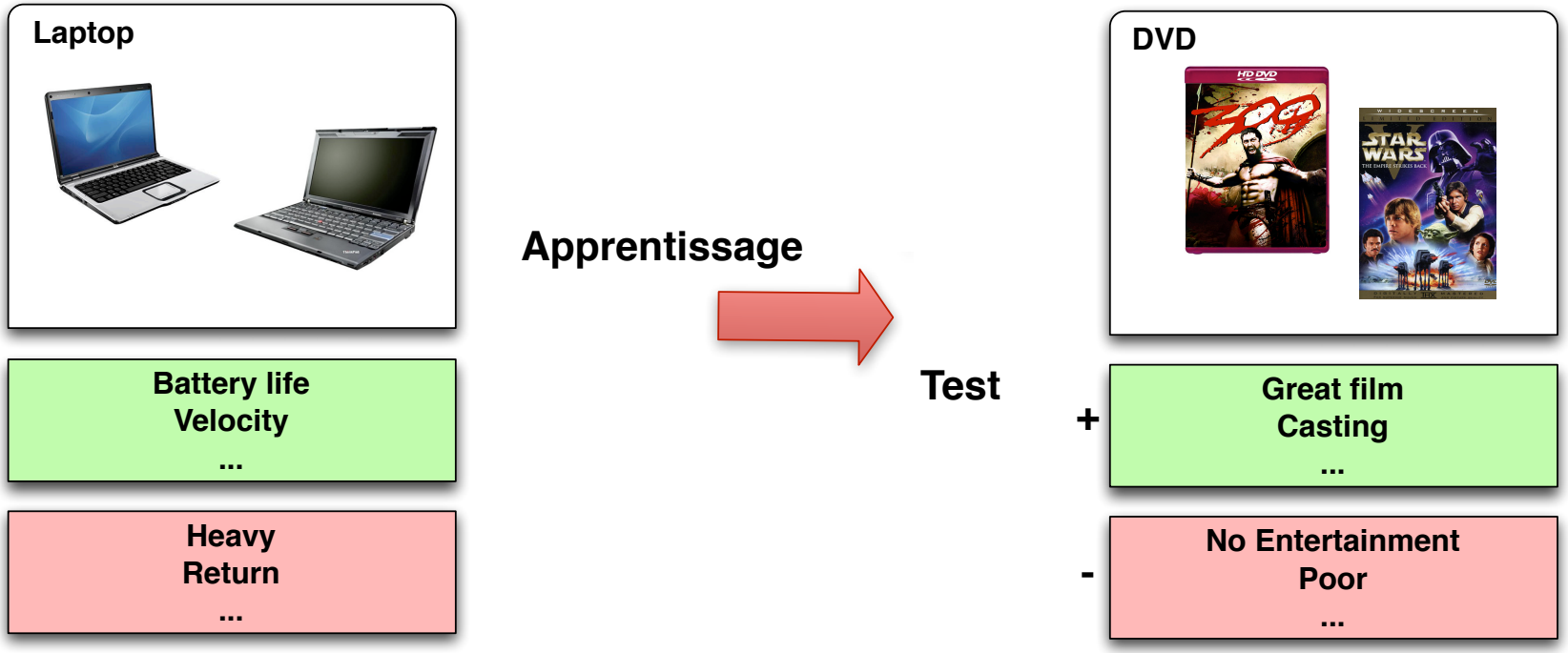


Zhao Liu Wang, EMNLP 2008
Adding Redundant Features for CRFs-based Sentence Sentiment Classification

Zirn, Niepert, Stuckenschmidt, Strube, IJCNLP, 2011
Fine-Grained Sentiment Analysis with Structural Features

Plan

Problématique



Difficultés

- Vocabulaire/expression spécifique au domaine
- Besoin de généralisation

Hypothèse i.i.d.

- Multi-domaines : les données ne sont pas identiquement distribuées
  - Distributions des mots différentes en apprentissage et en test
- Multi-utilisateurs : les données ne sont pas indépendantes
  - Les utilisateurs notent différemment (l'échelle n'est pas universelle)

⇒ Comment améliorer la généralisation ?

Données et protocoles

1. Apprentissage sur une [Blitzer et al.] ou plusieurs sources
2. *Amélioration du modèle*
3. Test sur un autre domaine

Propositions d'amélioration (adaptation/généralisation) :

- Utiliser quelques documents cibles pour :
  - Aligner le vocabulaire
  - Construire un second modèle
- Occam : simplifier le modèle de base
- Construire un espace de concept pour réduire le fossé sémantique

Blitzer Dredze Pereira, ACL 2007

Biographies, Bollywood, Boom-boxes and Blenders : Domain Adaptation for Sentiment Classification

Rapide historique

- Apparition de la problématique (spécifique au ML !) [Aue, 2005]
- Proposition de solutions générales à base de modèles multiples [Daumé, 2007]
- Propositions spécifiques aux sentiments...

Aue, Gamon, RANLP 2005

Customizing sentiment classifiers to new domains : A case study

Daumé, ACL 2007

Frustratingly Easy Domain Adaptation

Algorithme simple [Daumé, ACL 2007]

Hypothèse : on dispose d'une petite base de test

- La base de test est trop petite pour être bien prise en compte dans l'apprentissage (démo empirique)
- Proposition : codage redondant

$$\Phi^s(\mathbf{x}) = \langle \mathbf{x}, \mathbf{x}, 0 \rangle, \quad \Phi^t(\mathbf{x}) = \langle \mathbf{x}, 0, \mathbf{x} \rangle$$

NB : implémentation = recopie des vecteurs

- Régularisation(s) possible(s) : sur $\|\mathbf{w}\|^2$, sur $\|\mathbf{w}^s - \mathbf{w}^t\|^2$

$\Longrightarrow$ Très bonnes performances... Sur des applications non sentiment.

Daumé, ACL 2007

Frustratingly Easy Domain Adaptation

SCL [Blitzer 2006]

Input: labeled source data $\{(\mathbf{x}_t, y_t)_{t=1}^T\}$,
unlabeled data from both domains $\{\mathbf{x}_j\}$

Output: predictor $f : X \rightarrow Y$

1. Choose m pivot features. Create m binary prediction problems, $p_\ell(\mathbf{x})$, $\ell = 1 \ldots m$

2. For $\ell = 1$ to m

$$\hat{\mathbf{w}}_\ell = \underset{\mathbf{w}}{\text{argmin}} \left(\sum_j L(\mathbf{w} \cdot \mathbf{x}_j, p_\ell(\mathbf{x}_j)) + \lambda ||\mathbf{w}||^2 \right)$$

end

3. $W = [\hat{\mathbf{w}}_1 | \ldots | \hat{\mathbf{w}}_m]$, $[U\ D\ V^T] = \text{SVD}(W)$,
$\theta = U^T_{[1:h,:]}$

4. Return f, a predictor trained on $\left\{ \left(\begin{bmatrix} \mathbf{x}_t \\ \theta \mathbf{x}_i \end{bmatrix}, y_t \right)_{t=1}^T \right\}$

Figure 3: SCL Algorithm

- Source étiquetée, Test non supervisé
- Pivot p= mot fréquent
- w = quels autres mots cooccurent avec p ?
- Analyse en valeurs singulières de W = extraction concept
- Apprentissage dans l'espace : origine+concepts

Blitzer, McDonald, Pereira, ACL 2007

Domain Adaptation with Structural Correspondence Learning

Algo. [Blitzer et al. 2007], [Pan et al., 2010]

- Trouver les mots qui ont des comportement similaires pour passer au nouveau domaine
  - SCL-MI : extraire des pivots de sentiments par MI
  - Spectral Feature Alignment : ACP sur une matrice de liens dans un graphe bipartite de mots
- Projeter les documents dans l'espace des concepts pour la classification

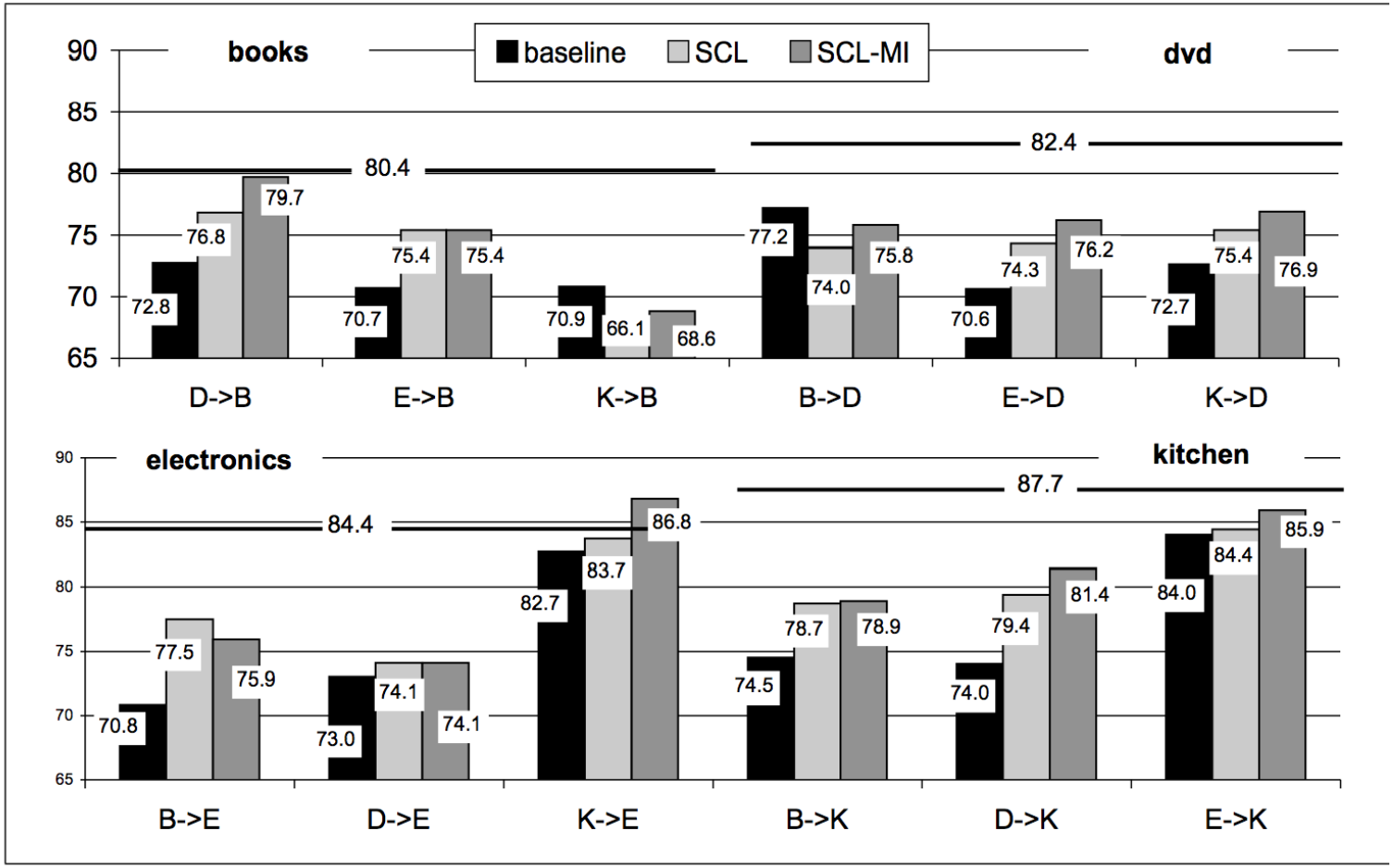
Blitzer, Dredze, Pereira, ACL 2007

Biographies, Bollywood, Boom-boxes and Blenders : Domain Adaptation for Sentiment Classification

Pan, Ni, Sun, Yang, Chen, WWW 2010

Cross-Domain Sentiment Classification via Spectral Feature Alignment

Algo. [Blitzer et al. 2007], [Pan et al., 2010]



Blitzer, Dredze, Pereira, ACL 2007

Biographies, Bollywood, Boom-boxes and Blenders : Domain Adaptation for Sentiment Classification

Li, Sun, Yang, Chen, WWW 2010

[Whitehead, 2009]

Utiliser plusieurs sources pour apprendre un modèle... Non concluant

Table 1: Cross Domain Classification

Testing Dataset \ Training Dataset	camera	camp	doctor	drug	laptop	lawyer	movie	music	radio	rest	tv	Ave. Test
camera	**90**	64	67	57	64	63	51	55	60	61	62	63
camp	71	**85**	69	57	53	68	52	63	62	66	71	65
doctor	59	65	**84**	58	53	72	50	57	63	72	65	63
drug	57	59	59	**72**	53	62	50	54	57	59	56	58
laptop	74	56	56	53	**96**	63	57	50	55	61	52	61
lawyer	57	61	71	55	59	**83**	51	60	60	66	65	63
movie	57	63	56	52	59	59	**81**	55	53	58	59	59
music	58	58	65	61	50	53	50	**88**	61	63	56	60
radio	55	64	62	53	52	62	50	58	**73**	59	58	59
rest	64	67	76	64	53	62	56	64	64	**85**	65	65
tv	61	70	63	53	51	71	52	58	62	63	**82**	62
Ave. Train	64	65	66	58	58	65	55	60	61	65	63	

Whitehead, Yaeger, IEEE 2009

Building a General Purpose Cross-Domain Sentiment Mining Model

[Whitehead, 2009]

Utiliser plusieurs sources pour apprendre un modèle... Non concluant

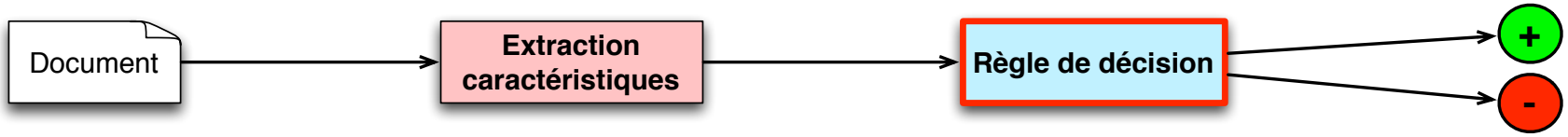
Table 3: Leave-One-Out vs. Single-Domain

Test Set	LOO	Ave. Other	Best Other	Same
camera	62	63	67	90
camp	66	65	71	85
doctor	61	64	72	84
drug	53	58	62	72
laptop	65	61	74	96
lawyer	58	63	71	83
movie	62	59	63	81
music	61	60	65	88
radio	56	59	64	73
restaurant	67	65	76	85
tv	60	62	71	82
Average	61	62	69	83

Whitehead, Yaeger, IEEE 2009

Building a General Purpose Cross-Domain Sentiment Mining Model

Baselines CLASSIQUES



Modèle linéaire et formulation générale unifiée :

$$f(\mathbf{x}) = \langle \mathbf{x}, \mathbf{w} \rangle$$

$$w^\star = \arg\min_w \left(C(X, Y) + \lambda_1 \Omega_{\mathcal{L}_1}(f) + \lambda_2 \Omega_{\mathcal{L}_2}(f) \right)$$

Zou & Hastie. Journal of the Royal Statistical Society, 2005
Regularization and variable selection via the Elastic Net

Dredze, Kulesza, Crammer, MLJ, 2010
Multi-domain learning by confidence-weighted parameter combination

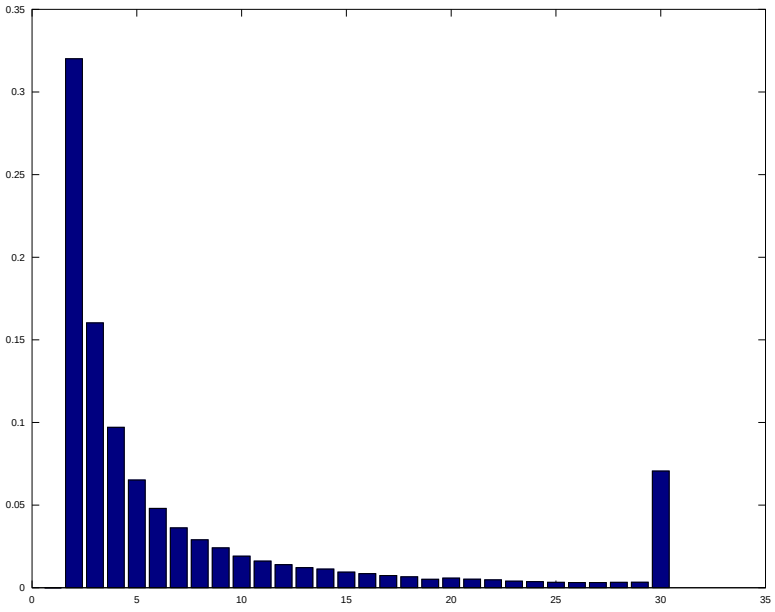
Fonctions de coût : $C(X, Y) = \begin{cases} (1 - yf(\mathbf{x}))_+ & \text{charnière} \\ (f(\mathbf{x}) - y)^2 & \text{moindres carrés} \end{cases}$

Régularisation $\mathcal{L}_1$: $\Omega_{\mathcal{L}_1}(f) = \sum_j |w_j|$, Régularisation $\mathcal{L}_2$: $\Omega_{\mathcal{L}_2}(f) = \sum_j w_j^2$

Unification des modèles linéaires : SVM, $\mathcal{L}_1$-SVM, *Splines*, LASSO...

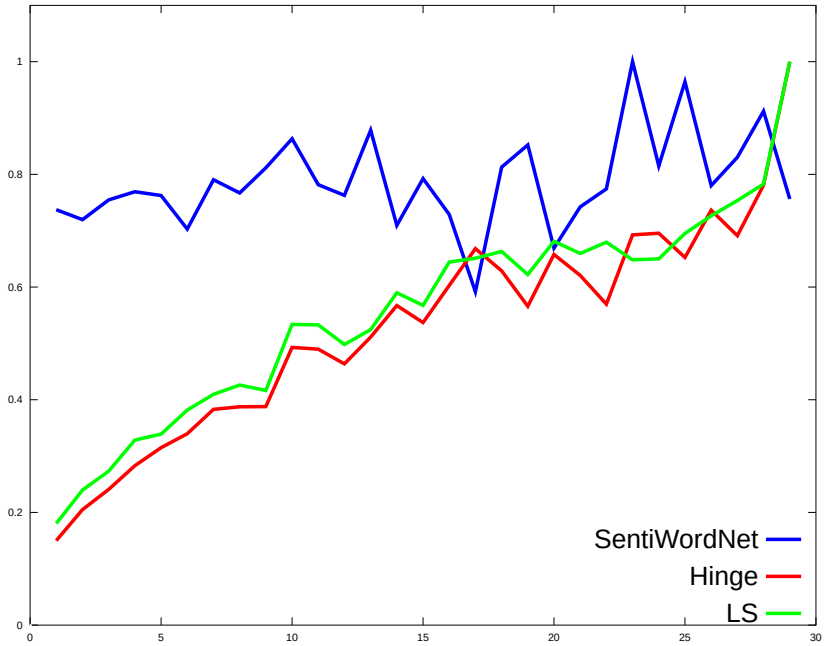
Biais Fréquentiel

Distribution des unigrammes en fonction de leurs fréquences



Distribution classique à longue traîne (*long tail distribution*)

Subjectivité moyenne des termes dans SentiWordNet et dans nos modèles (fréquence des mots en abscisse).



A. Esuli and F. Sebastiani. LREC, 2006

Sentiwordnet : A publicly available lexical resource for opinion mining.

Régularisation fréquentielle

Modification de la régularisation inspirée de K. Crammer :

$$\Omega(f) = \sum_{j=1}^{V} \nu_j \Omega_j(f), \quad \nu_j = \frac{\#\{\mathbf{x}_i | x_{ij} \neq 0\}}{\#\mathbf{X}} \in [0, 1]$$

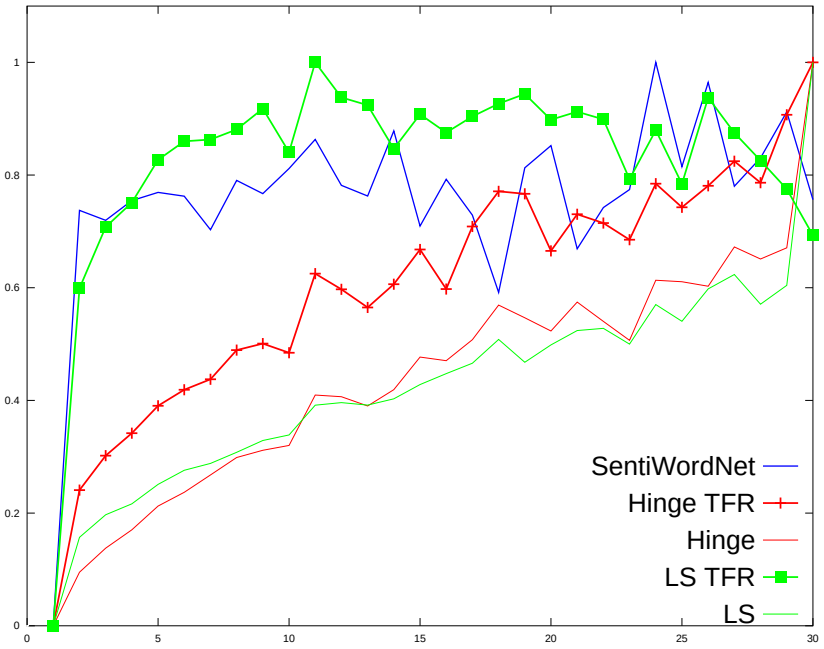
- La nouvelle régularisation est différente pour chaque terme.
- Les termes plus fréquents sont plus pénalisés.
- Afin de rester sur une technique *on-line*, les ν_j sont estimés sur des mini-batchs

K. Crammer et al. NIPS, 2009.
Adaptive Regularization of Weight Vectors

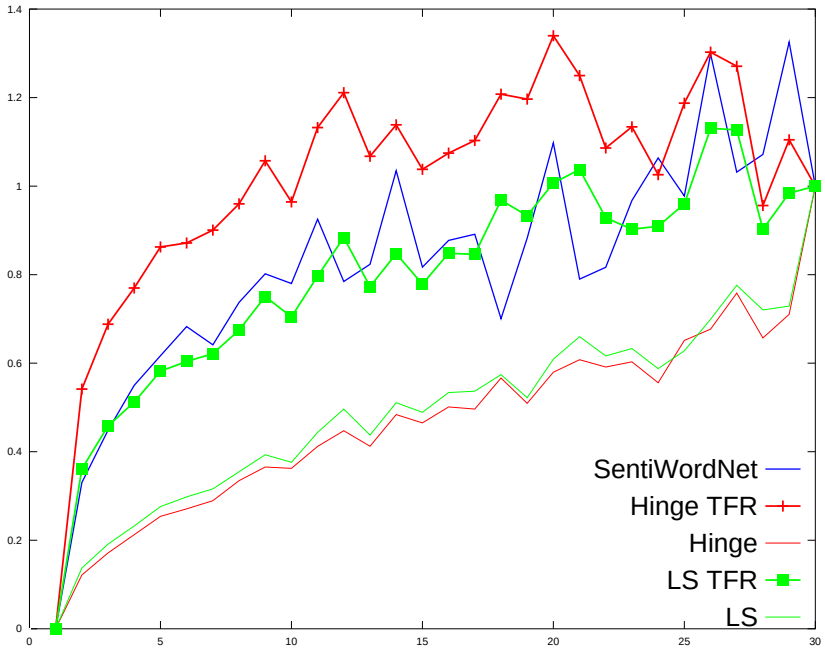
Bilan (1) en subjectivité

Bilan : nous donnons plus de poids aux mots rares.

Unigrammes :



Bigrammes :



Bilan (2) en taux de reconnaissance

	Meilleure RFT			Meilleure base			[1]	[2]	[3]	[4]
Modèle	Spline (Amazon), LASSO (M.R.)			SVM						
Caract.	U	UB	UBS	U	UB	UBS	U	UBS+	UB	UBS
Books	82.35	**85.5**	85.1	80.8	84.1	83.5	80.4	81.4	-	-
Dvd	84.25	**87.1**	85.6	82.7	84.0	84.3	82.4	82.55	-	-
Electr.	84.4	**88.2**	87.3	82.2	85.6	85.5	84.4	84.6	-	-
Kitchen	85.4	**88.3**	88.0	83.5	86.2	86.2	87.7	87.1	-	-
Mov. rev.	88.4	88.8	**92.2**	87.1	88	91.4	-	-	87.1	88.9

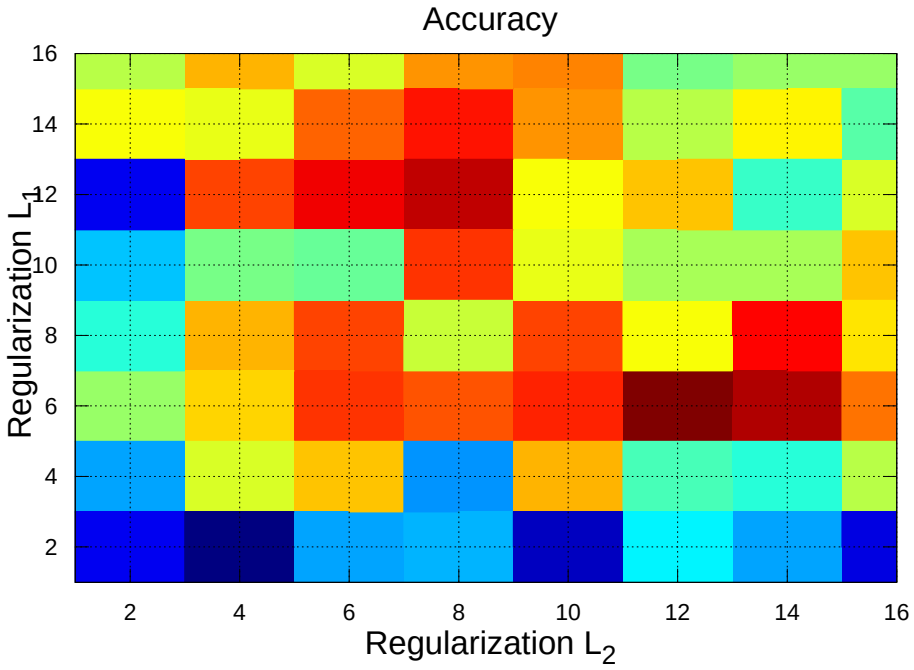
Table : Références : [1] (Blitzer *et al.*, 2007) [2] (Pan *et al.*, 2010) [3] (Pang *et al.*, 2004) [4] (Matsumoto *et al.*, 2005)

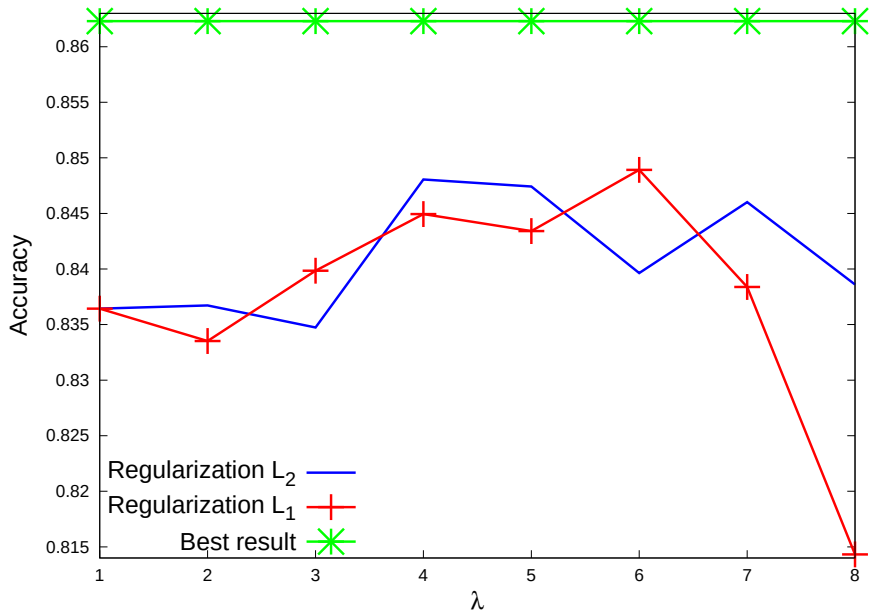
- Régularisation Adapt. $\Rightarrow$ +1.4% à +2.8%
- Performances équivalentes $\forall$ modèles
- UBS = ajout de bruit $\Rightarrow$ généralisation difficile

A. Rafrafi et al, CORIA, 2012.
Représentations et régularisations pour la classification de sentiments

Bilan (3) influence de la régularisation

Intérêt du cadre *Elastic Net* : Taux de reconnaissance en fonction des régularisations TFR+Bigrammes :



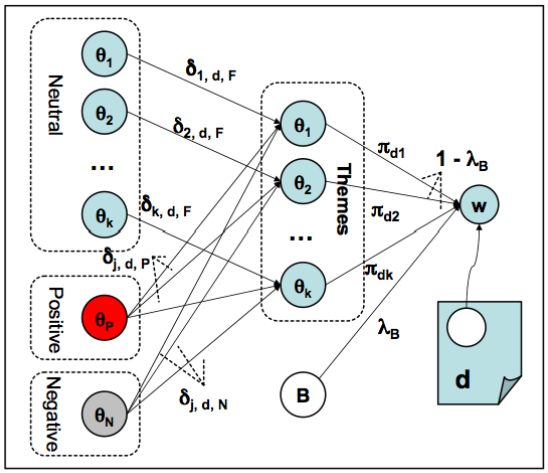


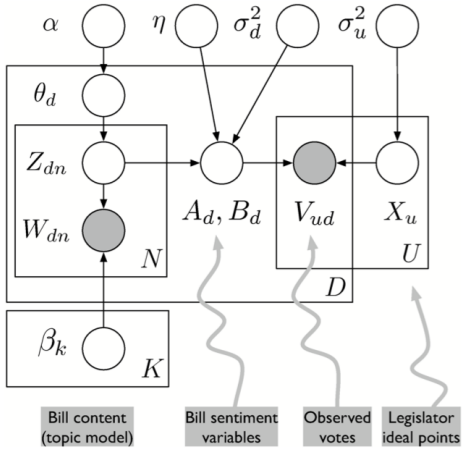
Meilleurs modèles : toujours + de 99% de coefs non nuls

TFR = boosting des mots rares mais sans parcimonie !

Analyse Latente Classique

- Combler le fossé sémantique
- Représenter les domaines et le sentiment





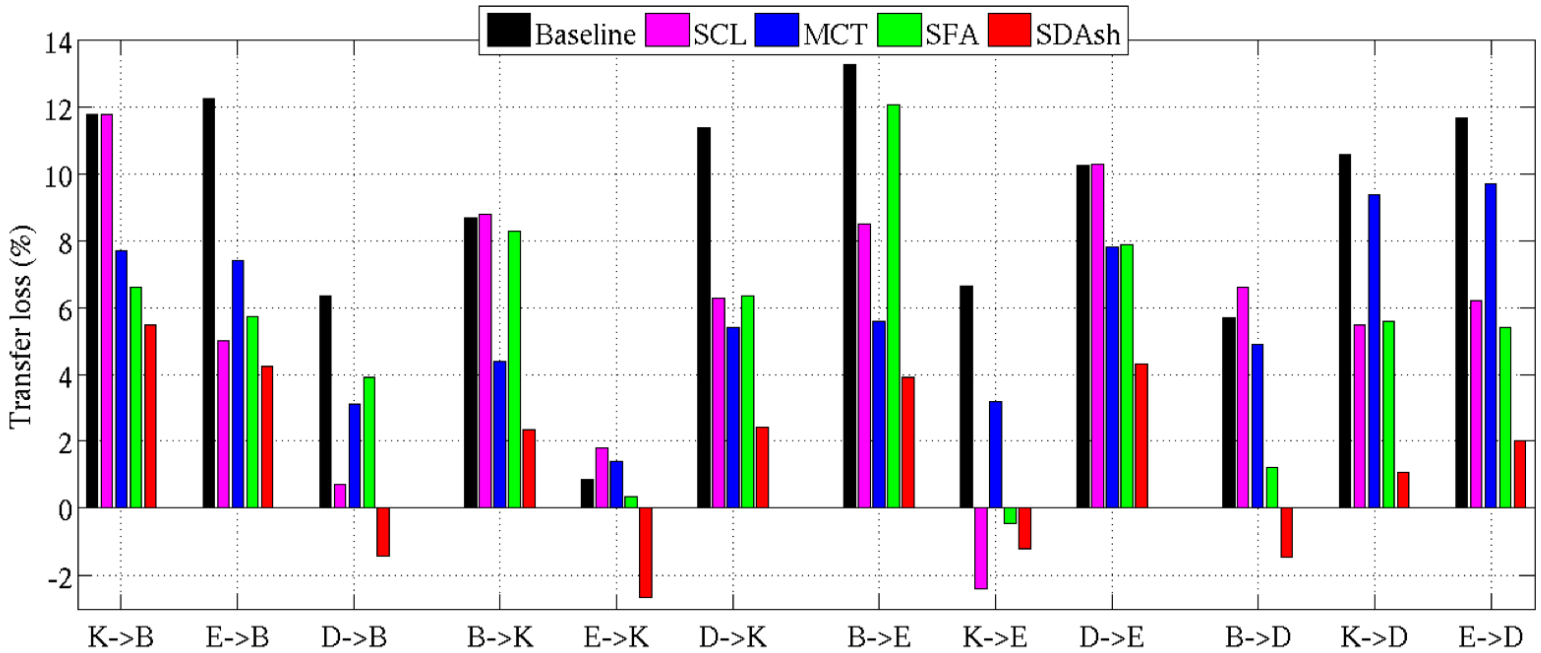
Liu, Huang, An, Yu, SIGIR 2007
ARSA : a sentiment-aware model for predicting sales performance using blogs

Mei, Ling, Wondra, Su, Zhai, WWW 2007
Topic Sentiment Mixture : Modeling Facets and Opinions in Weblogs

Gerrish, Blei, ICML 2011
Predicting Legislative Roll Calls from Text

[Glorot et al. 2011]

- **Passage à l'échelle** via réseau de neurones : autoencodeur
- Prise en compte de données de sentiments non-supervisées
- Dépassement (net) des performances mono-domaine



Glorot, Bordes, Bengio, ICML 2011

Domain Adaptation for Large-Scale Sentiment Classification : A Deep Learning Approach

Réseau de neurones à convolution(s)

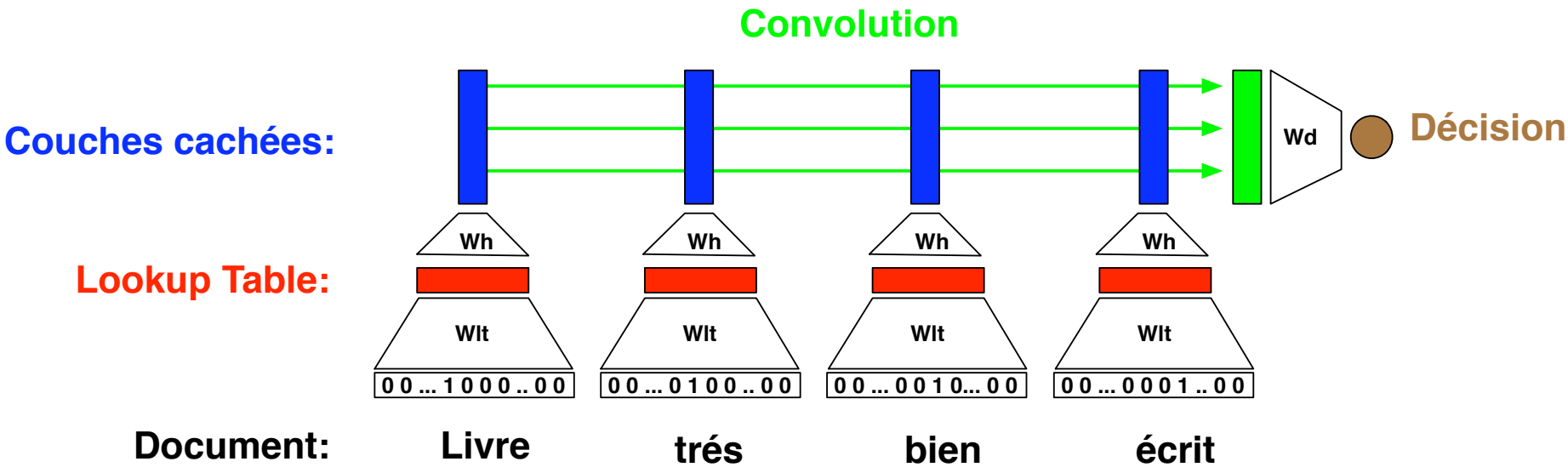
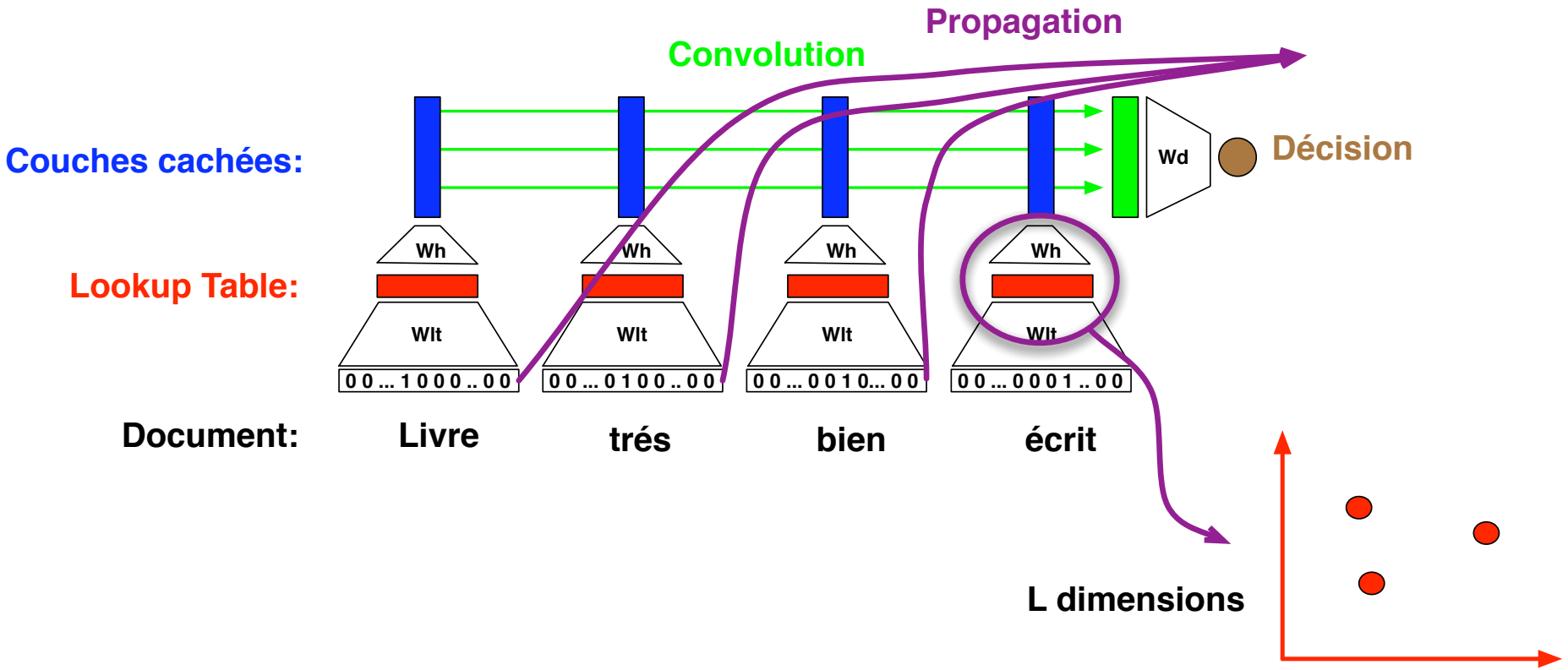


Figure : Réseau de neurones pour la classification de sentiments et la construction d'un espace sémantique.

Collobert & Weston, ICML, 2008.
A Unified Architecture for Natural Language Processing : Deep Neural Networks with Multitask Learning

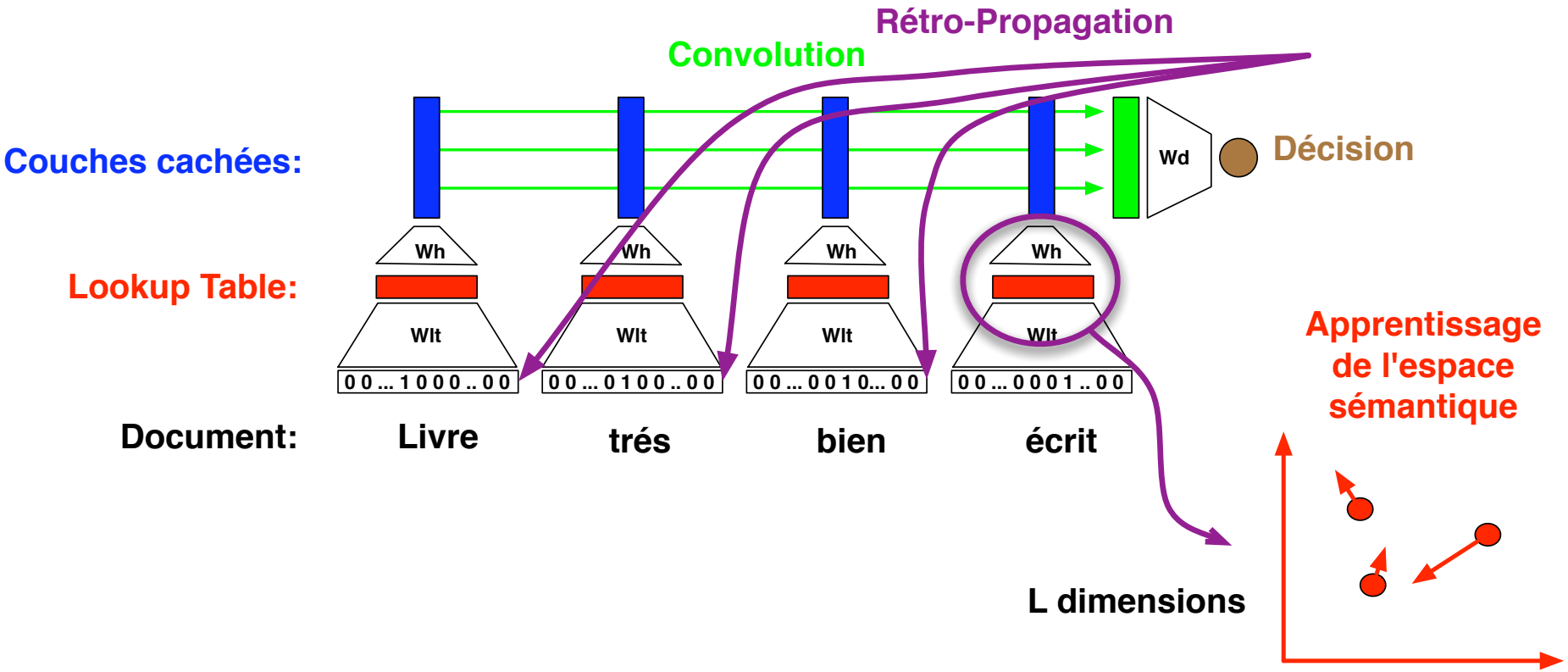
Rafrafi et al. CORIA, 2011
Réseau de neurones profond et SVM pour la classification de sentiments

Flexibilité et fonctionnement du modèle



- Apprentissage multi-tâches / modèle de langue
- Introduction d'informations TALN (POS-Tag, négation...)

Flexibilité et fonctionnement du modèle



- Apprentissage multi-tâches / modèle de langue
- Introduction d'informations TALN (POS-Tag, négation...)

[Ng 2011] [Bespalov 2011]

- Ng 2011 : (ré)introduction de taches non supervisées dans les réseaux à convolution
- Bespalov 2011 : étude de l'influence de la taille du corpus sur les résultats
  - On ne peut pas augmenter la taille du dictionnaire indéfiniment
  - ...Mais on peut trouver des données à l'infini !

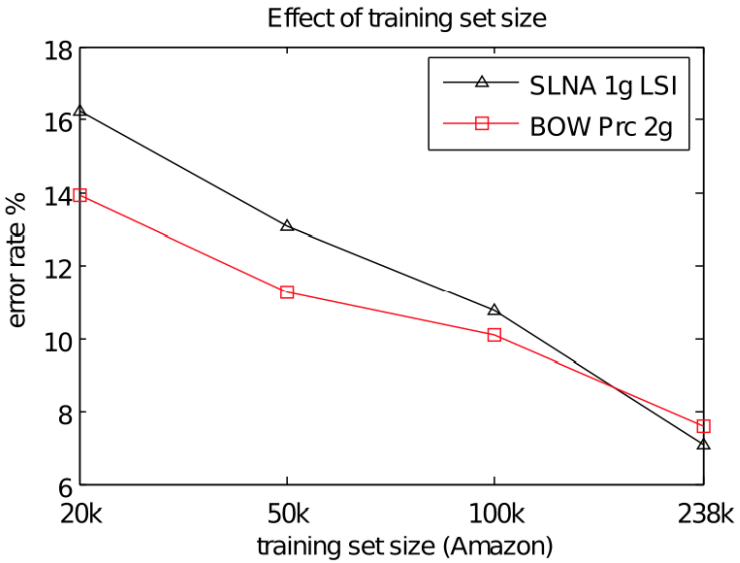


Figure 2: Testing error on Amazon with different training set size.

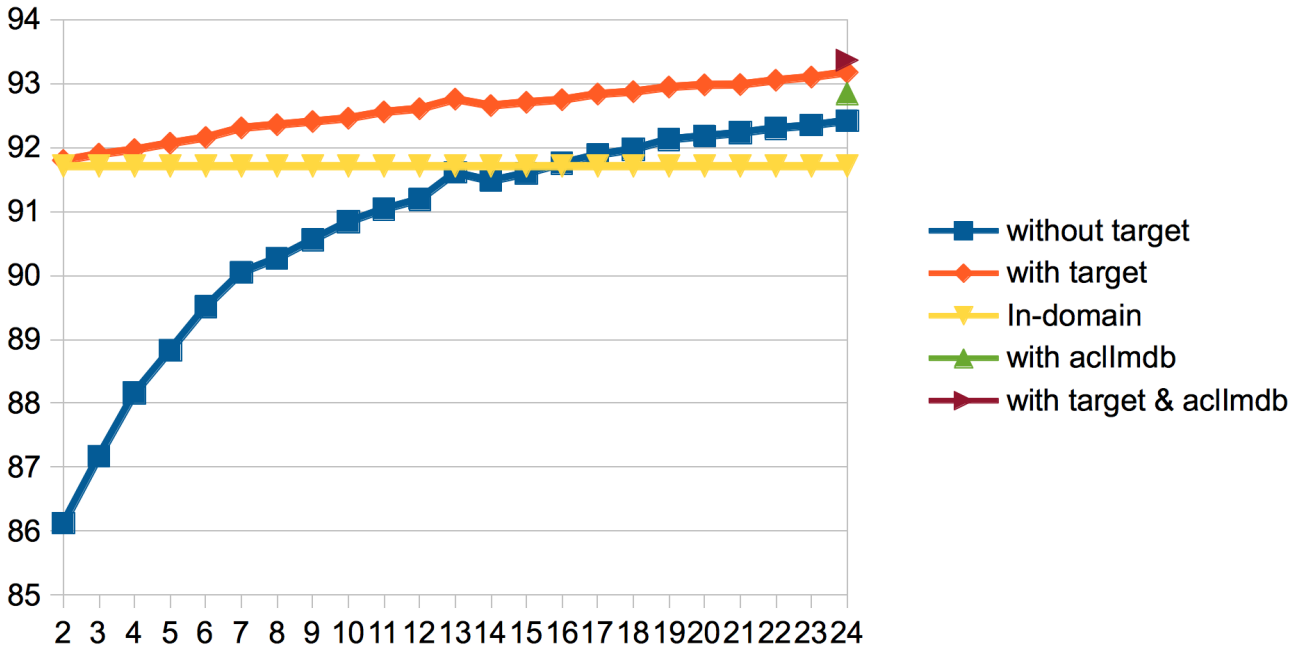
Socher Pennington Huang Ng Manning, EMNLP 2011
Semi-Supervised Recursive Autoencoders for Predicting Sentiment Distributions

Bespalov, Bai, Qi, Shokoufandeh, CIKM 2011

Sentiment Classification Based on Supervised Latent n-gram Analysis

Volume & Multi-domaines [Rafrafi 2013]

Que se passe-t-il si on augmente la taille de l'ensemble d'apprentissage par rapport à la tâche multi-domaines ?



Niveaux de transfert

- De produit à produit (Amazon)
- De produit à d'autres choses (Whitehead, Movie Reviews, Trip advisor...)
- De revue à d'autres formes de contenus (Tweets, blogs...) [Mejova 2012]

Les revues, c'est ce qu'il y a de mieux... Mais surtout pour les revues.

Mejova, Srinivasan, ICWSM 2012

Crossing Media Streams with Sentiment : Domain Adaptation in Blogs, Reviews and Twitter

Plan

Retour au NLP !

- Un problème différent des revues et des blogs (vocabulaire, expression, abréviations...)
- Mais un problème pas si difficile
  - Message court = homogène
  - Règles simples et efficaces (smileys, mots caractéristiques)
  - Quelques corpus qui apparaissent (TREC 2011, ICWSM 2012 + Amazon Mechanical Turk)

Mejova, Srinivasan, ICWSM 2012

Crossing Media Streams with Sentiment : Domain Adaptation in Blogs, Reviews and Twitter

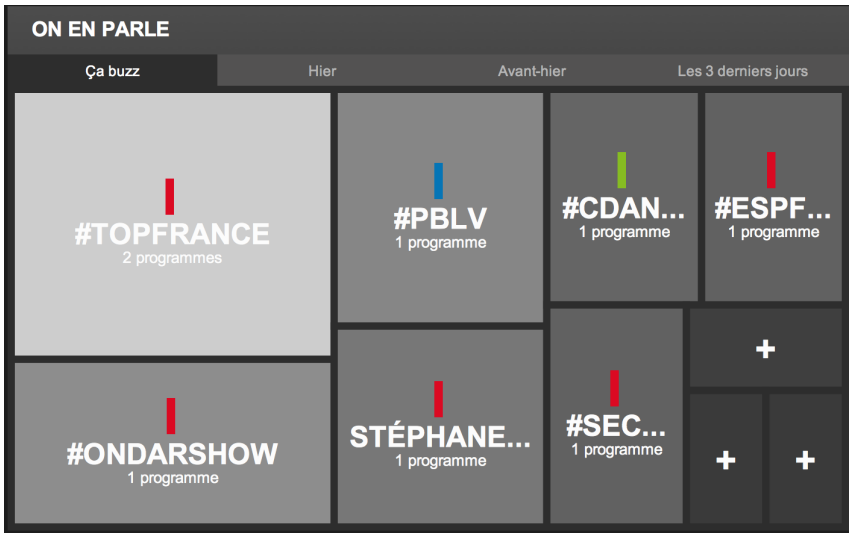
Pak Paroubek, LREC 2010

Twitter as a Corpus for Sentiment Analysis and Opinion Mining

Une tâche de base dépassée

... Mais de nombreuses applications

- Beaucoup d'étude quantitative basées sur des outils existants
  - Prédiction de vote
  - Prédiction de box-offices
  - Prédiction de cours de bourse
  - Suivi de tendances
  - Médiamétrie, churn...
- Peu de littérature récente sur la détection de polarité elle-même



Applications liées à l'analyse de sentiments

- Diffusion (1) : hypothèse de régularité. Dans un graphe, les noeuds connectés (utilisateurs, messages...) expriment la même opinion
  ⇒ amélioration de la détection de polarité, problème d'apprentissage
- Diffusion (2) : qui influence qui ? Une requête, une polarisation : on recherche qui est un influenceur
  ⇒ comptage, recherche de centralité
- Sondage (& prédiction) : comptage + régression sur un couple (polarité, thème)
  ⇒ comptage + régression
- Etude quantitative : comptage seul

De tâche en tâche, la détection de polarité devient marginale et l'apprentissage disparaît...

Etude amusante :

Dans un graphe de Hashtag (obtenu par projection), comment sont réparties les polarités ?

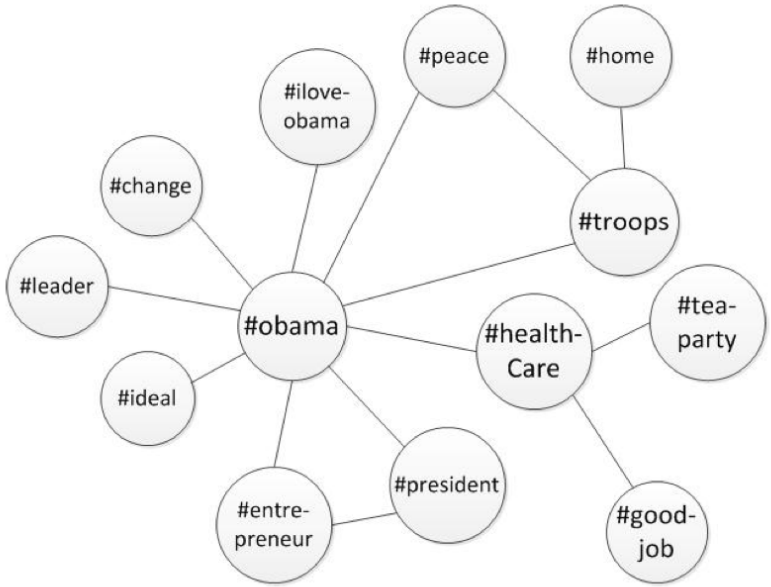


Figure 1: An example of a Hashtag Graph Model

Wang, Wei, Liu, Zhou, Zhang, CIKM 2011

Topic Sentiment Analysis in Twitter : A Graph-based Hashtag Sentiment Classification Approach

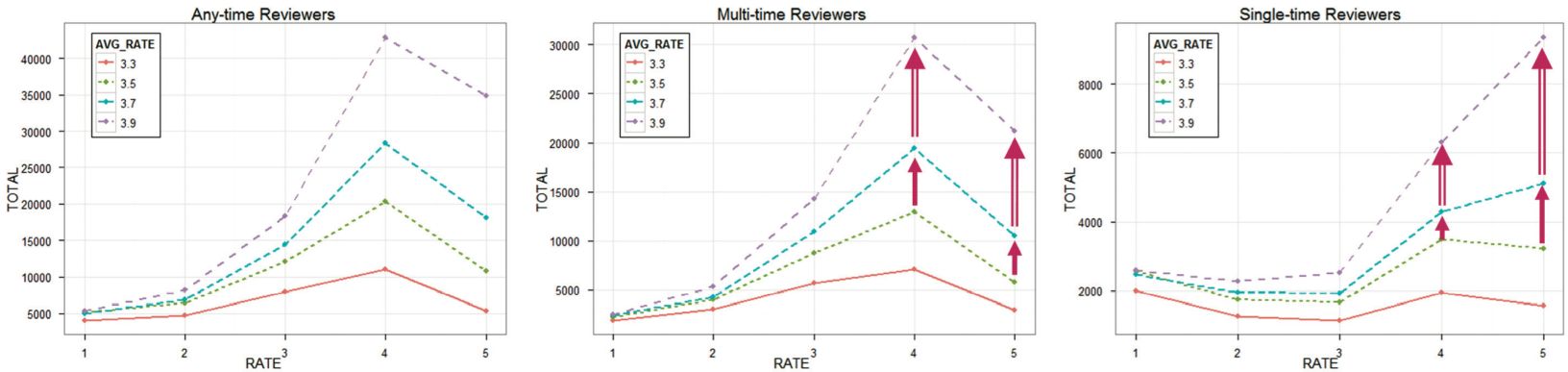
Lien avec la médecine : contamination

- Epidémiologie Qui (ou combien) sont touchés par une information au temps t ?
- Patient 0 Soit une situation donnée, où se trouve la source de l'information ? Et quand a débuté la contamination ?
- Définition d'un leader d'influence : quelqu'un qui contamine beaucoup...

Dans ce domaine il reste de l'apprentissage, pour comprendre les mécanismes de diffusion

Spam(s)

1. Identification du spam **mail** : stade commercial
   - Filtres bayerions, listes noires, maintient profil utilisateur
2. Identification du web spam : intégré dans les moteurs de recherche
   - Fermes de liens (topologie réseau), forme des pages (topologie page)...
3. Identification des review spams [très difficile !] :



Feng et al., ICWSM 2012 : caractérisation des distributions naturelles de revues

Conclusions

- Une littérature abondante
- Beaucoup de ressources disponibles (corpus, implémentations de modèles, lexiques...)
- Beaucoup de problématiques... En NLP, mais assez peu en apprentissage.
- Grand nombre d'applications (et de start-up, et d'ANR...)
- **Un problème difficile, motivant, multi-disciplinaire**

Perspectives

- Améliorer le transfert (il y a trop d'étiquettes pour les négliger !)
- Mieux étudier le passage à l'échelle
- Mieux comprendre les mécanismes de prise de décision dans les modèles statistiques
  - proposer de nouveaux modèles plus efficaces à toutes les granularités
- Et bien sûr, toutes les applications que l'on peut imaginer !