

Le Traitement Automatique des Langues en France à l'ère du Big Data À l'aube d'un révolution technologique

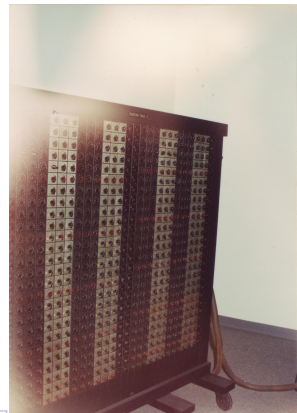
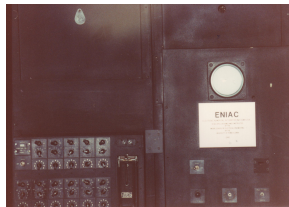
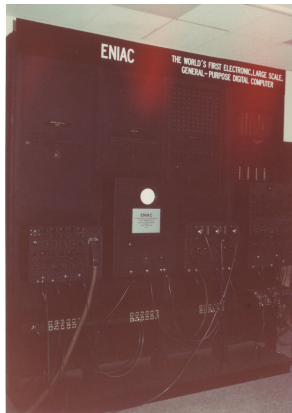
Patrick Paroubek

LIMSI-CNRS
Dépt. CHM - Groupe LIR
Bât. 508 Université Paris Sud, 91403 Orsay Cedex
pap@limsi.fr

ATALA - Association pour le Traitement Automatique des Langues
45, rue d'Ulm
75230 PARIS Cedex 5 - France
<http://www.atala.org>

Naissance de l'informatique

ENIAC - 13 février 1946 - *Moore School of Elec. Engineering U Penn*



La langue naturelle

NOUVELLES REMARQUES de M. DE VAUGELAS SUR LA LANGUE FRANÇOISE, Paris, Guillaume Desprez, 1690.

Peu de gens font difficulté de dire aujourd'hui, *prenez le cas*, en imitant M. de Voiture qui est l'Auteur dont M. de Vaugelas veut parler dans cette remarque. On dit aussi, *posez le cas*: mais je ne voy point qu'on soit si malicieux à-present qu'on l'étoit du tems de M. de Vaugelas, & que ce mot de *cas* dans *prenez le cas* fasse un si mauvais effet sur l'esprit ny des hommes ny des femmes, à-moins que ce ne soit sur celuy de quelques débauchez ou de quelques libertins qui se divertissent de tout, &

Naissance du TAL



En France, l'ATALA (Association pour le Traitement Automatique des Langues) est créée en 1959.

<http://www.atala.org>

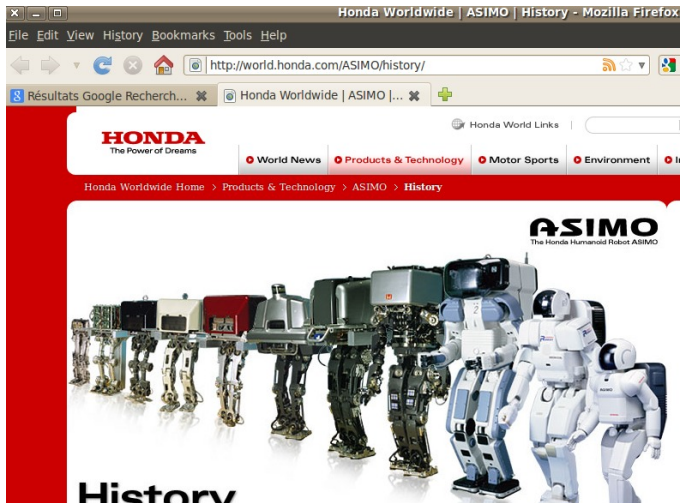


Aux États-Unis, l'ACL (Association for Computational Linguistics) est créée en 1962. <http://www.acl.org>



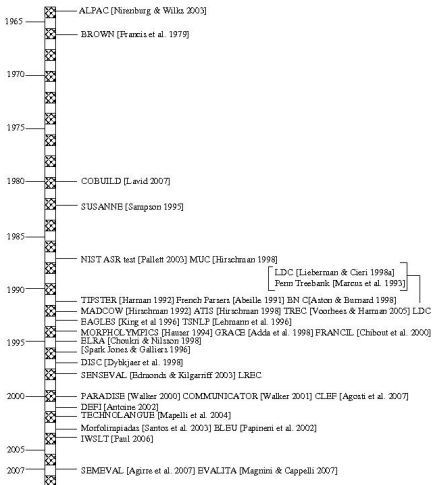
Evolution

Honda Asimo robot



Evolution

Vers plus de corpus et plus d'ingénieries



LA RECHERCHE EN TAL

Si l'analyse et la production de textes sont naturelles pour les êtres humains, il n'en va pas de même pour les ordinateurs : les langues naturelles relèvent à la formalisation complète qui constitue le préalable habituel à un traitement algorithmique. Cette situation motive des recherches de fond sur le traitement automatique des langues, recherches menées dans le public (universités, organismes de recherche) comme dans le privé (centres de recherche de grandes entreprises informatiques) en linguistique et en informatique.

En linguistique, les chercheurs modélisent (à sa complétude et raisonnement) les propriétés des langues à tous leurs paliers (morphologie, syntaxe, sémantique, pragmatique) en ayant en ligne de mire une formalisation suffisante pour permettre leur implémentation sur machines. Ce travail inclut la rédaction de grammaires formelles et de grands dictionnaires utilisés par des programmes informatiques.

En informatique, il faut mettre au point des algorithmes d'analyse et de génération de langue naturelle, mais aussi des formalismes de représentation des connaissances linguistiques adaptés à des contraintes contradictoires d'expressivité, de calculabilité et de robustesse. Par ailleurs, de la même façon que les êtres humains acquièrent des connaissances linguistiques par l'expérience quotidienne et la lecture, certaines méthodes à base statistique poursuivent le contenu de grandes collections de textes pour engager automatiquement des connaissances qui serviront ensuite à analyser ou produire d'autres textes.

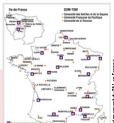
Une part importante de ces travaux se situe à la rencontre de ces disciplines et nécessite une double compétence.

L'ENSEIGNEMENT DU TAL EN FRANCE

L'enseignement universitaire du Traitement Automatique des Langues se répartit uniformément sur le territoire français, aussi bien au niveau licence qu'au niveau Master. Le TAL était une discipline fondamentalement interdisciplinaire entre informatique et linguistique, les IIR de rattachement des formations sont pour moitié des formations en Sciences du Langage, pour l'autre moitié en informatique, toutes adossées à des laboratoires de recherche ayant une compétence forte en TAL.

Selon les formations, une héritance recouvrable à ce jour, le TAL est enseigné comme un composant d'une formation professionnalisante ou comme une finalité en soi de formation professionnelle étendue de recherche. Dans un cas comme dans l'autre, les responsables de formation s'accordent sur le fait que, dans des sciences du langage ou de l'informatique, les étudiants ont un fort taux d'insertion dans l'industrie, notamment parce que les enseignements recouvrent au gré des avancées de la discipline.

La qualité des enseignements et des méthodes permet de répondre aux besoins du monde professionnel tout en assurant un renouvellement et un accordement de la communauté scientifique universitaire en traitement automatique des langues. La demande de stagiaires et de professionnels ne décroît pas malgré la crise actuelle, et justifie pleinement la persévérance des formations en traitement automatique des langues.



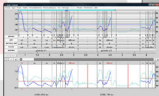
L'enseignement du TAL en France

Une application de TAL



LE TAL EN QUELQUES CHIFFRES

- Le TAL en France c'est :
 - 27 formations universitaires (source RFP) ;
 - 8 métiers typés : argotologue, chef de projet, consultant avant-vente, linguiste informaticien, ingénieur assurance qualité, terminologue, vendeur pédagogique, ingénieur support (source TECHNOLANGUE) ;
 - 5 secteurs d'activité : moteurs de recherche intelligent, gestion de contenu, applications vocales et multimedias, e-learning et traduction automatique (source TECHNOLANGUE) ;
 - Une place dans le Hit des leaders Européen sur le marché du TAL (source Bureau van Dijk Information Management 2006) ;
 - de 82 à 150 sociétés (source APL 2008 et ministère de la Culture 2007) ;
 - 100 professionnels de haut niveau en ingénierie linguistique appartenant tant au secteur de la recherche publique qu'au secteur industriel (ministère de la Culture 2007).



Le langage Anstet

Traduction automatique
Communication homme-machine
Extraction d'information
Analyse de données
Représentation des connaissances



ATALA
Association
pour le Traitement
Automatique
des Langues
www.atala.org

Traitement Automatique

des Langues

Le TAL en France

- une trentaine de formations Universitaires
- **8 métiers types** : linguiste informaticien, terminologue, ergolinguiste, chef de projet, consultant avant-vente, ingénieur assurance qualité, veilleur stratégique, ingénieur support,
- **5 secteurs d'activité** : moteurs de recherche intelligents, analyse et gestion de contenu, application vocales et multimodales, e-learning et traduction automatique ,
- une place dans le trio des leaders Européens du marché du TAL ,
- une centaine de sociétés,
- plusieurs centaines de professionnels de haut niveau en ingénierie linguistique, recherche publique et industrielle

Carte des formations TAL en F... +

thomaslebarbe.free.fr/REPTIL2/formation.php?ville=0

formation tAL

ATALA Association pour le Traitement Automatique des Langues

[ATALA](#) [Adhésion](#) [Plan du site](#)

ATTENTION - Site en cours de reconstruction.
Consulter l'[ancien site REPTIL de l'ATALA](#) - Attention, certaines informations sur ce site sont obsolètes.

Rubriques

REPTIL
Le projet REPTIL
Carte interactive des formations
Tableau Récapitulatif
Appel à participation

Carte des Formations TAL en France

proposée par l'ATALA

Liste complète des formations

| | | | |
|----------------------|----------|---|--|
| voir | Avignon | Université d'Avignon | Master Sciences et Technologies |
| voir | Besançon | Université de Franche-Comté, Centre Lucien Tesnière | Master en TAL // Master Erasmus Mundus in NLP |
| voir | Bordeaux | Université de Bordeaux | Master Linguistique |
| voir | Bordeaux | Université de Bordeaux I | Master Informatique |
| voir | Caen | Université de Caen, Département Informatique | Master Informatique |
| voir | Grenoble | Université Stendhal - Grenoble 3 | Master Industries de la Langue |
| voir | Grenoble | Université Stendhal - Grenoble 3 | Parcours Informatique, Lettres, Langues et Langage |
| voir | Lille | Université de Lille III | Master TAL en Sciences du Langage |

Le LIMSI-CNRS

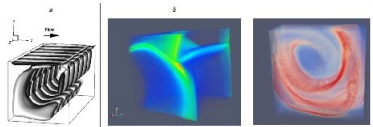
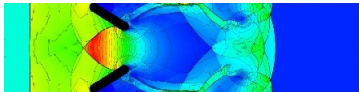
- Laboratoire pour la Mécanique et les Sciences de l'Ingénieur
- Unité propre du CNRS
- localisation : campus Paris-Saclay
- effectif : env. 200 permanents + 120 stagiaires



Le LIMSI-CNRS

2 départements :

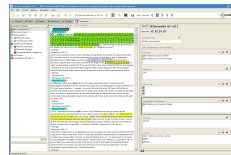
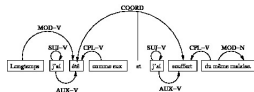
- Dépt. Mécanique-Énergétique (3 groupes)
 - Aérodynamique Instationnaire : turbulence et contrôle
 - Convection et Rotation : instabilités et turbulence
 - Transferts Solide-Fluide



Le LIMSI-CNRS

2 départements :

- **Dépt. Communication Homme-Machine** (6 groupes)
 - Audio & Acoustique
 - Architectures et Modèles pour l'Interaction
 - Réalité Virtuelle & Augmentée
 - Cognition, Perception et Usages
 - Traitement du Langage Parlé
 - **Information, Langue Ecrite et Signée**



Le groupe ILES

Effectif : env. 20 permanents, 15 Doctorants, 15 CDD

- Corpus & représentations
- Multilinguisme & paraphrases
- Extraction d'information précise & systèmes de Questions-Réponses
- Langues des signes

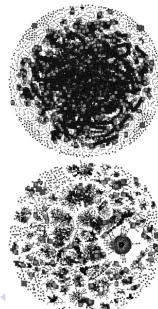
Langage

Pas de norme

- évolution permanente
- Académie Française (1635)
- langue = consensus émergent au sein d'une population



Expérience d'émergence de lexique d'actions (thèse J.



Le rapport du Mac Kinsey Global Institute

<http://www.economie.gouv.fr/files/rapport-mckinsey-company.pdf>,
paru en 2012, positionne le Big Data et le Big Analytics parmi
les 10 technologies stratégiques pouvant créer un impact fort
au sein des entreprises pour les 3 prochaines années à venir.

En 2012, Twitter a publié les statistiques suivantes :

- 465 millions de comptes Twitter,
- 175 millions de Tweets/jour,
- des pointes de 2500 Tweets/seconde,
- +grand nombre de « followers » : 19 millions / Lady Gaga,
- 64% des accès à Twitter se font par Internet,
- 69% des utilisateurs suivent leurs amis,
- 110 nouveaux comptes/seconde,
- près d'un million de nouveaux comptes/jour.
- des un pic record de 24,1 million de tweets envoyés durant la nuit du Super Bowl du 03 février 2013
- l'anglais +utilisé, 39% des messages en octobre 2011 soit 70 millions de tweets quotidiens,
- MAIS +60% des tweets sont rédigés dans d'autres langues.

How can we determine that this is a positive opinion?



"An indescribably funny, altogether remarkable movie
from the creators of 'Being John Malkovich.'"

Peter Travers, ROLLING STONE

Nicolas Cage
Meryl Streep
Chris Cooper



Directed by Spike Jonze
Screenplay by Charlie Kaufman and Donald Kaufman

Adaptation.

COLUMBIA PICTURES PRESENTS IN ASSOCIATION WITH INTERMEDIA FILMS A MAGNET/CLINICA ESTETICO PRODUCTION STARRING NICOLAS CAGE MERYL STREEP CHRIS COOPER "ADAPTATION"
EXECUTIVE PRODUCERS CASEY STORM PRODUCED BY CARTER BURWELL EDITOR ERIC ZIMBRUNNEN EXECUTIVE PRODUCERS BY MICHAEL BARRETT PRODUCED BY LANCE ACORD WRITTEN BY CHARLIE KAUFMAN AND DONALD KAUFMAN DIRECTED BY SPIKE JONZE
SCREENPLAY BY CHARLIE KAUFMAN AND DONALD KAUFMAN PRODUCED BY EDWARD SAXON VINCENT LANDAY JONATHAN DEMME
COLUMBIA PICTURES
Coming Soon
sony.com/Adaptation

“An indescribably **funny**, altogether **remarkable** movie from the creators of 'Being John Malkovich'”

- We can determine positive polarity by finding positive words
- To know which words are positive – we use **affective lexicons** (ex: ANEW)

ANEW (Affective Norms of English Words) is a list of 1034 English words with assigned scores of polarity, intensity, and control

| Description | Valence (sd) | | Arousal | | Dominance | |
|-------------|--------------|------|---------|------|-----------|------|
| abduction | 2.76 | 2.06 | 5.53 | 2.43 | 3.49 | 2.38 |
| abortion | 3.5 | 2.3 | 5.39 | 2.8 | 4.59 | 2.54 |
| absurd | 4.26 | 1.82 | 4.36 | 2.2 | 4.73 | 1.72 |
| abundance | 6.59 | 2.01 | 5.51 | 2.63 | 5.8 | 2.16 |
| abuse | 1.8 | 1.23 | 6.83 | 2.7 | 3.69 | 2.94 |
| acceptance | 7.98 | 1.42 | 5.4 | 2.7 | 6.64 | 1.91 |
| accident | 2.05 | 1.19 | 6.26 | 2.87 | 3.76 | 2.22 |
| ace | 6.88 | 1.93 | 5.5 | 2.66 | 6.39 | 2.31 |
| ache | 2.46 | 1.52 | 5 | 2.45 | 3.54 | 1.73 |
| achievement | 7.89 | 1.38 | 5.53 | 2.81 | 6.56 | 2.35 |
| activate | 5.46 | 0.98 | 4.86 | 2.56 | 5.43 | 1.84 |
| addict | 2.48 | 2.08 | 5.66 | 2.26 | 3.72 | 2.54 |

Valence = polarity takes a value between 1 and 9
Higher value => more positive term



Twitter is the most popular microblogging platform for exchanging short text messages



Kimbo_The_Buppy Kimberly Leak



@RonDon96 upenn is my favorite ivy. and you would've definitely got in :)

11 hours ago



rusyniak23 David Rusyniak

wanna visit upenn, but it so far and i wont get in :(

9 Oct

1. Use emoticons :) to collect positive and negative messages
2. Estimate words polarity based on occurrences in sets

Table VI
AN EXAMPLE OF POSITIVE
WORD LIST

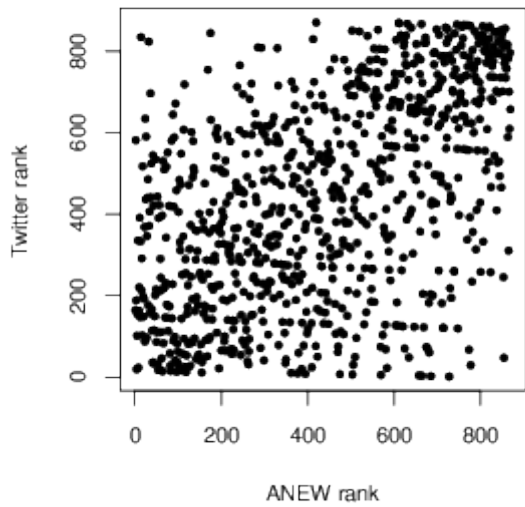
| Word | ANEW | Est. |
|--------------|-------------|-------------|
| sexy | 8.02 | 8.18 |
| adorable | 7.81 | 7.67 |
| plaisir | 8.28 | 7.64 |
| salut | 5.92 | 7.32 |
| magnifique | 7.6 | 7.08 |
| anniversaire | 7.84 | 6.68 |
| gentil | 7.59 | 6.45 |
| bien | 7.47 | 6.2 |
| simple | 7.1 | 6.18 |
| beau | 6.55 | 5.88 |

Table VII
AN EXAMPLE OF NEGATIVE
WORD LIST

| word | ANEW | Est. |
|-------------|-------------|-------------|
| triste | 1.61 | 1.24 |
| terrible | 1.93 | 1.53 |
| malade | 1.9 | 1.73 |
| dommage | 2.49 | 1.79 |
| mort | 1.61 | 2.82 |
| mal | 2.46 | 2.93 |
| dur | 4.74 | 3 |
| peur | 2.76 | 3.26 |
| seul | 2.41 | 3.61 |
| partie | 5.11 | 4.05 |

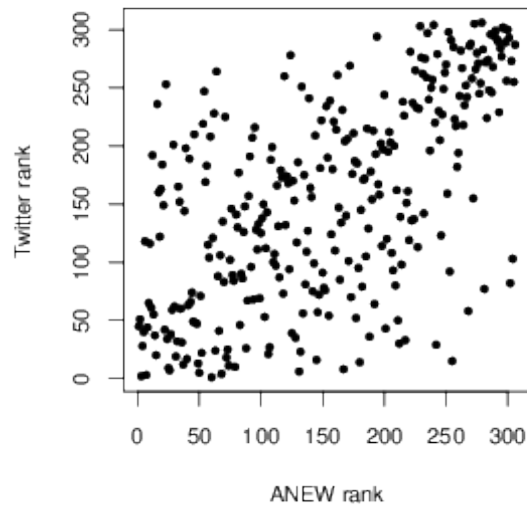
Correlation with ANEW

$\tau = 0.36$



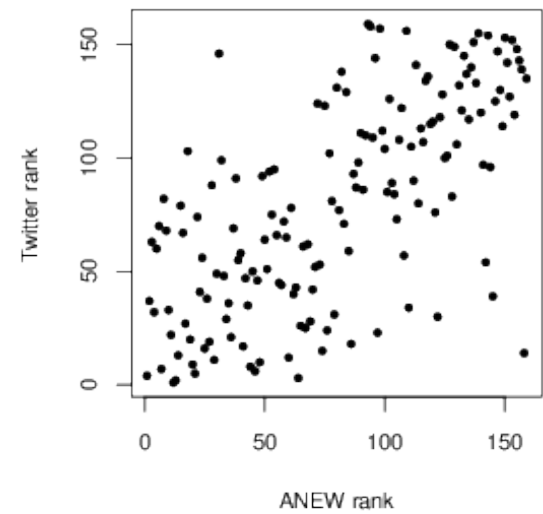
All words

$\tau = 0.45$



Occurring at least
in 100 messages

$\tau = 0.55$



Adjectives

Polarity classification in video game reviews

| Model | Ave. acc | Ave. prec | Prec_{pos} | Prec_{neg} |
|--------------|-----------------|------------------|---------------------------|---------------------------|
| Unigram | 73.86 | 69.57 | 90.57 | 48.57 |
| Bigram | 72.73 | 69.11 | 86.79 | 51.43 |
| Trigram | 64.77 | 60.08 | 83.02 | 37.14 |
| Twitter | 71.59 | 68.16 | 84.90 | 51.40 |

Our naive method worked "almost as good" as a standard N-gram based approach. However, our method does not require training data.



Text representation using dependency tree
subgraphs



Traditional approaches use **bag-of-words** text representation model and an **SVM** classifier

Given two sets of texts: **positive** and **negative**

Represent each text using **bag-of-words** model

Construct a **feature vector** for each text

Train an **SVM** classifier (or any other ML)

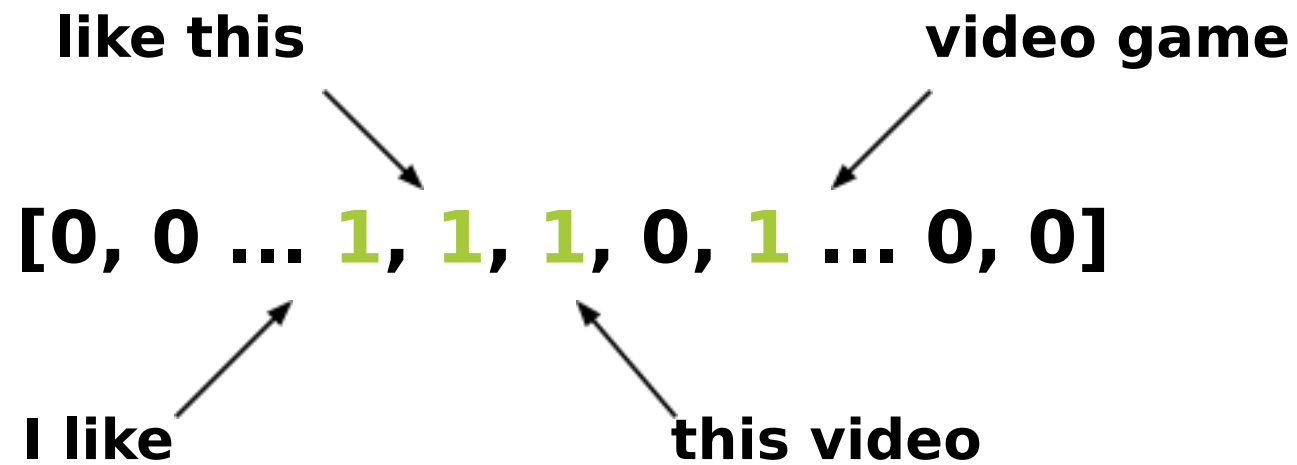
Classify an **unknown** text

Bag-of-words model



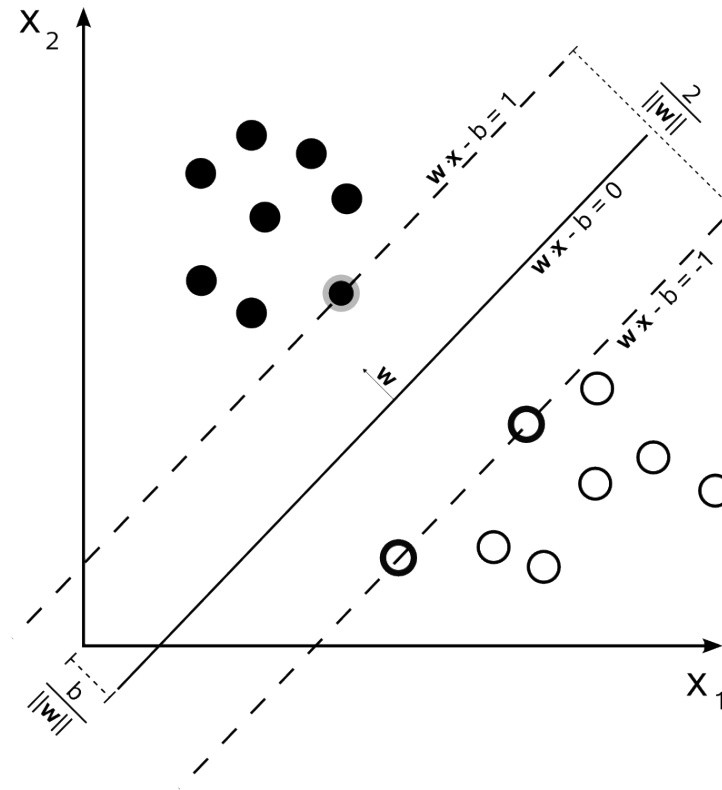
"I like this video game"

Feature vector



"I like this video game"

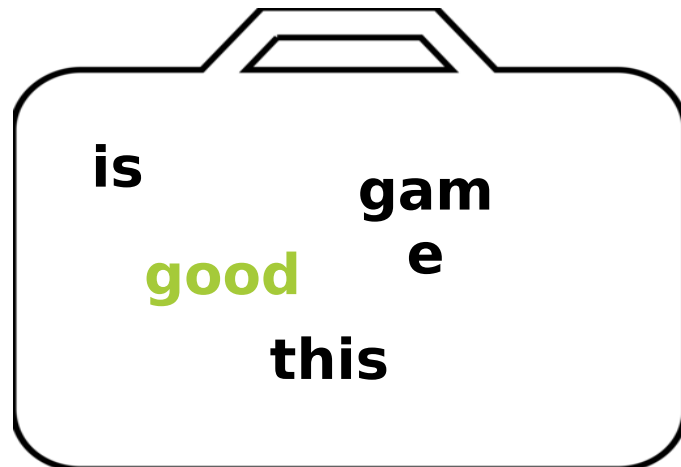
SVM



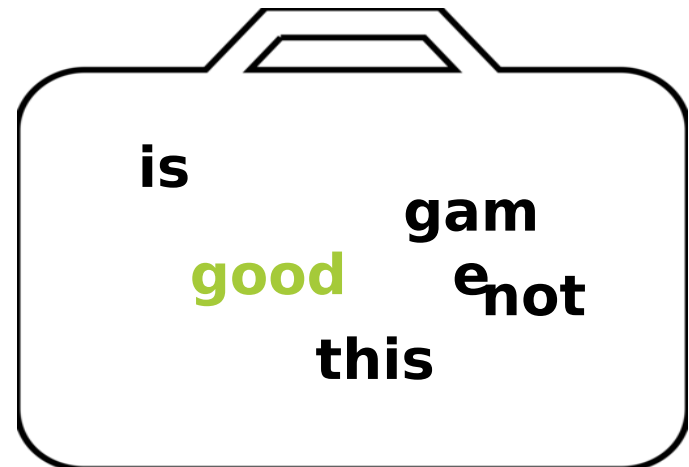
"I like this video game"

Problems of bag-of-words model

This game is good



This game is not good



By considering words as **independent** terms, we lose information from the **links** between words

Problems of bag-of-words model

This game is good



This game is not good



To solve this problem, we can try to use **n-grams** of a higher order

Problems of bag-of-words model

This game is **not good**



This game is **not really that good**



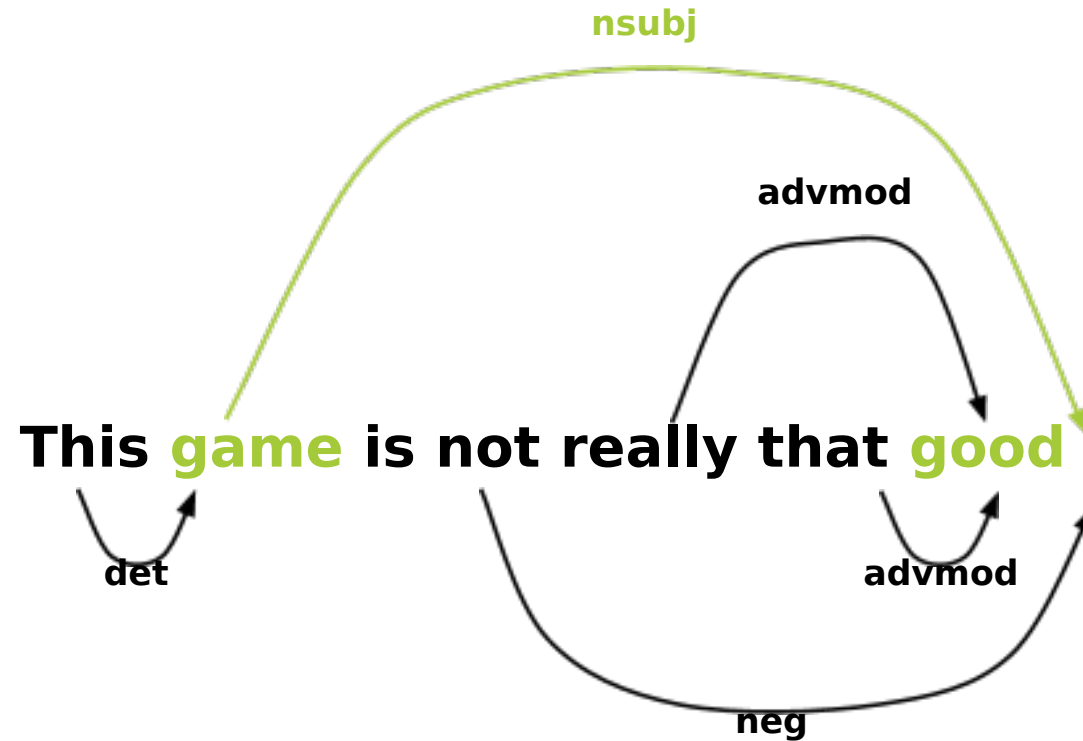
But we still have a problem with **long dependencies**

Our method



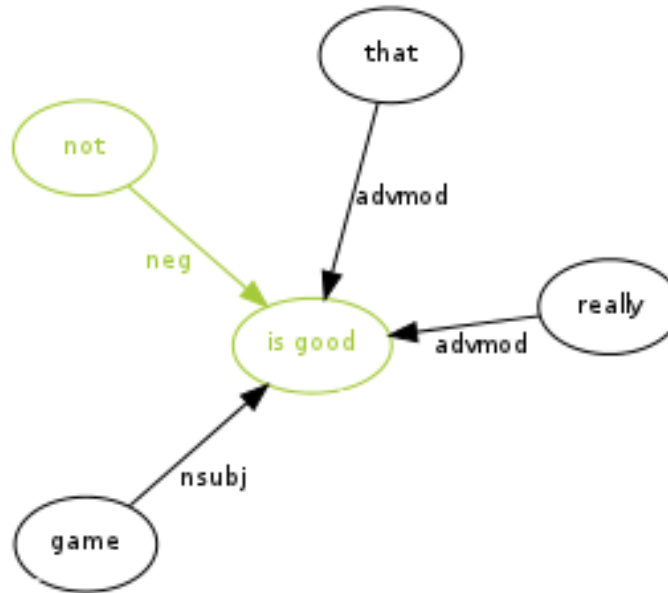
We propose to use sentence **dependency tree**
for text representation

Dependencies



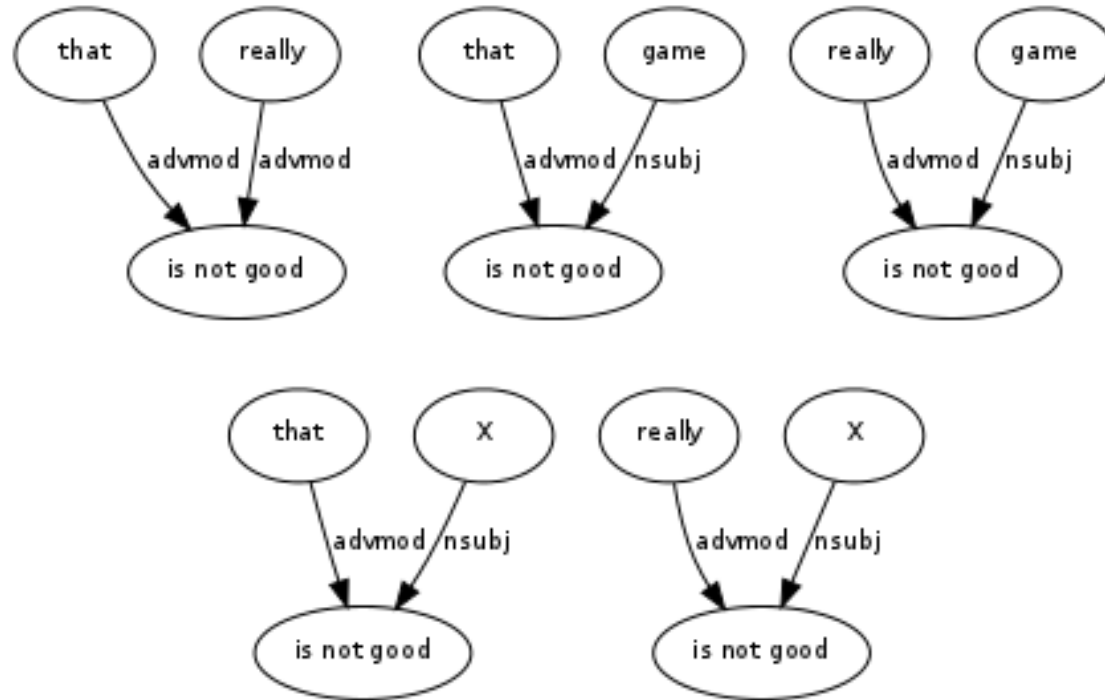
And find the **opinion subject**

Dependency tree



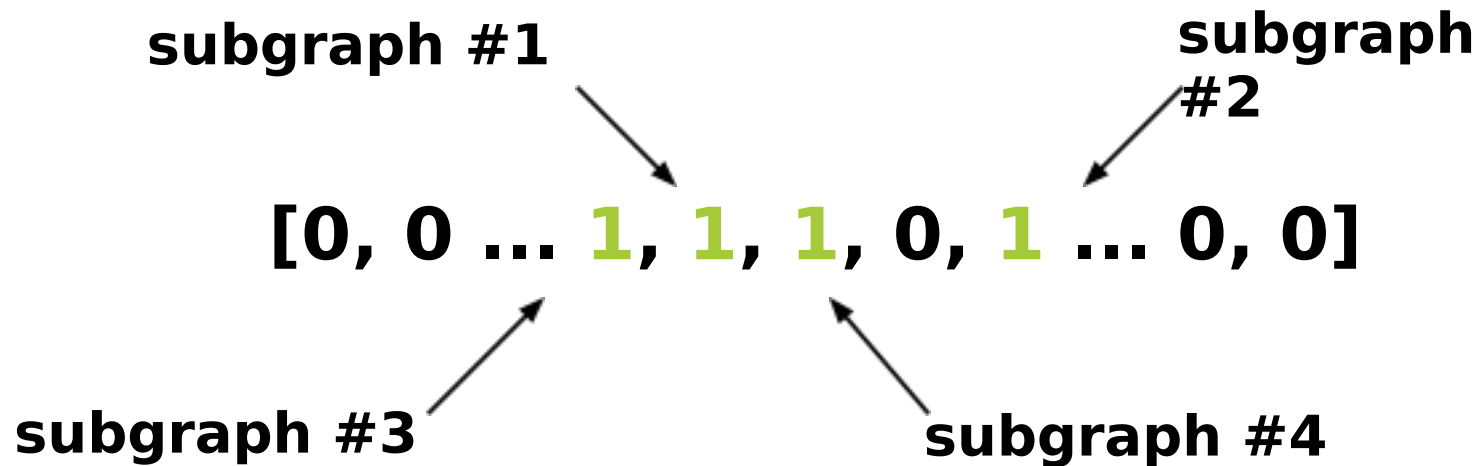
Next, is **combining nodes** with partial meaning

Subgraph text representation



**dependency tree subgraph text
representation**

Feature vector



It can be used the same way as the bag-of-words model,
e.g.: to train an **SVM** classifier

- Le Cloud Computing, rupture technologique dans la gestion des données et la conception des systèmes informatiques,
- Le Cloud Computing substrat technologique et technique facilement assimilable,
- Les approches Big Data sont complexes, multiformes et encore en voie de définition.
- Le Data Mining et la Visualisation de Données préexistants et associés au Big Data n'ont pas encore donné pleinement la mesure de leurs potentialités.

Le rapport du Mac Kinsey Global Institute

<http://www.economie.gouv.fr/files/rapport-mckinsey-company.pdf>,
paru en 2012, positionne le Big Data et le Big Analytics parmi
les 10 technologies stratégiques pouvant créer un impact fort
au sein des entreprises pour les 3 prochaines années à venir.

- Le traitement de telles données imposent de faire appel à des technologies de rupture à tous les niveaux : infrastructure, stockage, analyse et visualisation.
- Pour résoudre le problème de (« scalabilité »), il faut de nouvelles technologies associant **Big Data**, **Big Analytics**, **Visual Analytics** et **Cloud Computing**.

- Le traitement de telles données imposent de faire appel à des technologies de rupture à tous les niveaux : infrastructure, stockage, analyse et visualisation.
- Pour résoudre le problème de (« scalabilité »), il faut de nouvelles technologies associant **Big Data**, **Big Analytics**, **Visual Analytics** et **Cloud Computing**.

Mais il y aura toujours de la place pour les petits...dans des niches qui vont se multiplier...

The screenshot shows the homepage of Short Édition, a website dedicated to short stories and literary content. The header features the logo "shortédition" with the tagline "AVEC MOINS DE 2000 MOTS", a search bar, and a "LIBRAIRIE EN LIGNE" badge. Navigation tabs include "NOUVELLES", "BD COURTES", "POÈMES", and "TRÈS TRÈS COURTS".

The main content area is divided into several sections:

- ICI, ON AIME LA LITTÉRATURE COURTE**: A large banner with the text "Tout ce qui se lit d'un seul trait, en moins de 20 minutes. C'EST PAR ICI" and an image of a person's legs in a skirt and stockings.
- GRAND PRIX DU COURT**: An orange banner with a ribbon icon and the text "Participez à notre Prix. C'EST PAR ICI".
- LE BLOG**: A section titled "INTERVIEW DU JEUDI : CÉLINE LAURENT-SANTRAN" dated "LE 03 AVRIL", featuring a photo of a woman.
- HISTOIRES DE LECTURE**: A section titled "FESTIVAL DES CONVERSATIONS 2014 AVEC ORANGE" with a colorful graphic of speech bubbles and characters.
- À DÉCOUVRIR**: A section titled "NOUVELLES" featuring "Fantasme" (36 min, 455 lectures, 12 votes) and "RÊVES" by Pingouin, with a small author photo.

The URL in the browser is "short-edition.com/prix/histoires-de-lecture-2014".

LA RECHERCHE EN TAL

Si fondées et la production de textes sont naturelles pour les êtres humains, il n'en va pas de même pour les ordinateurs : les langues naturelles relèvent à la formalisation complète qui constitue le préalable habituel à un traitement algorithmique. Cette situation motive des recherches de fond sur le traitement automatique des langues, recherches menées dans le public (universités, organismes de recherche) comme dans le privé (centres de recherche des grandes entreprises informatiques) en linguistique et en informatique.

En linguistique, les chercheurs modèlent plus complètement et plus finement les propriétés des langues à tous leurs niveaux (morphologie, syntaxe, sémantique, pragmatique) en ayant en ligne de mire une formalisation suffisante pour permettre leur implémentation en machine. Ce travail inclut la réalisation de grammaires formelles et de grands dictionnaires utilisés par des programmes informatiques.

En informatique, il faut mettre au point des algorithmes d'analyse et de génération de langues naturelles, mais aussi des formalismes de représentation des connaissances linguistiques associant des contraintes combinatoires

d'expressivité, de calculabilité et de robustesse. For alléu, de la même façon que les êtres humains acquièrent des connaissances linguistiques par l'expérience quotidienne et la lecture, certaines méthodes à base statistique parcourent le contenu de grandes collections de textes pour engendrer automatiquement des connaissances qui seraient utiles à produire ou produire d'autres textes.

Une part importante de ces travaux se situe à la rencontre de ces disciplines et nécessite une double compétence.

L'ENSEIGNEMENT DU TAL EN FRANCE

L'enseignement universitaire du Traitement Automatique des Langues se répartit uniformément sur les territoires français, quel bien ou niveau (donc qu'il s'agit de niveau Master). Le TAL est une discipline fondamentalement interdisciplinaire entre informatique et linguistique, les LRF de rattachement des formations sont pour moitié des formations en Sciences du Langage, pour l'autre moitié en informatique, toutes adossées à des laboratoires de recherche ayant une compétence forte en TAL.

Selon les formations, une maîtrise reconnue à ce jour, le TAL est enseigné comme un composant d'une formation professionnalisante ou comme une finalité en soi de formation professionnelle elle-même de recherche. Dans un cas comme dans l'autre, les responsables de formation s'accordent sur le fait que, sans des sciences du langage ou de l'informatique, les étudiants ont un fort taux d'abandon dans l'activité, notamment parce que les enseignements évoluent au gré des avancées de la discipline.

La visibilité des enseignements et des métiers visés permet de répondre que besoins du monde professionnel tout en assurant un renouvellement et un accroissement de la communauté scientifique universitaire en traitement automatique des langues. La demande de stagiaires et de professionnels ne décroît pas malgré la crise actuelle, et justifie pleinement la pérennisation des formations en traitement automatique des langues.



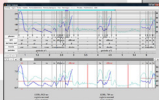
L'enseignement du TAL en France

Une application du TAL



LE TAL EN QUELQUES CHIFFRES

- Le TAL en France c'est :
 - 27 formations universitaires (source IFRIL);
 - 8 métiers typés : ergonômiste, chef de projet, consultant avant-vente, ingénieur informaticien, ingénieur assurance qualité, terminologue, veilleur technologique, ingénieur support (source TIC-RELIANCE);
 - 5 secteurs d'activité : moteurs de recherche intelligents, gestion de contenu, applications vocales et multimédias, e-learning et traduction automatique (source TIC-RELIANCE);
- Une place dans le trio des leaders Européen sur le marché du TAL (source Barou van Dam Information Management 2008);
- de 82 à 150 sociétés (sources APIL 2006 et ministère de la Culture 2007);
- 500 professionnels de haut niveau en ingénierie linguistique appartenant tant au secteur de la recherche publique qu'au secteur industriel (ministère de la Culture 2007).



Le global leader

Traduction automatique
Communication homme-machine
Extraction d'information
Analyse de données
Représentation des connaissances



ATALA

Association
pour le Traitement
Automatique
des Langues
www.atala.org

Traitement Automatique

des Langues

<http://www.atala.org/-Adhesion->

| | | |
|-------|----------------------|---------------------------------|
| 20 € | Etudiants / Chômeurs | Présentation d'un justificatif |
| 40 € | Individuels | Moyen de paiement personnel |
| 60 € | Professionnels | Moyen de paiement professionnel |
| 200 € | Institutionnels | Adhésion de soutien |