Introducing linguistic knowledge in sentiment analysis methods

3 types of approaches in sentiment analysis

1/ End-to-end deep learning approaches (no linguistic expertise is integrated)

3 types of approaches in sentiment analysis

2/ Knowledge-based/Rule-based methods based exclusively on linguistic expertise

*(manque|~negation-patt|(il/#NEG/y/avoir/~negation-patt))/(#PREP_DE)?/ (conseil|contact|~services-lex)**

3 types of approaches in sentiment analysis

3/ Machine learning/deep learning approaches that integrate linguistic expertise

- Using pre-processing
- Using hand-crafted features based on knowledge-based methods (hybrid methods)
  - Example : linguistic and syntactic patterns are used as inputs of supervised machine learning
    - Barrière, V., Clavel, C., Essid, E., Opinion Dynamics Modeling for Movie Review Transcripts Classification with Hidden Conditional Random Fields, Interspeech 2017

Outline

- **Pre-processing using linguistic knowledge for machine learning**
- Knowledge-based methods (hybrid methods)

Pre-processing using linguistic knowledge for machine learning

Motivation for preprocessing

	call	time	date	conference	release	meeting	corporation	earnings
document 1	2	1	3	2	1	1	1	
document 2	1		2	1	2	1	1	1
document 5		1	2		2	1	1	1
document 6	1	2	1	1	3	1	1	1
document 7	1						1	
document 8			1		1		1	1
document 9	2		1	3	1	1	1	1
document 10	2	1		1	1		1	1
document 13					1			2
document 14							3	
document 15	1			2			1	2

- **Learning the classes**

NL Pre-processing

This/PN movie/N is/VB really/RB good/JJ.

|'T'|'h'|'i'|'s'|' '|'m'|'o'|'v'|'i'|'e'|' '|'i'|'s'|' '|'r'|'e'|'a'|'l'|'l'|'y'|' '|'g'|'o'|'o'|'d'|'.'|

Convert documents into a Matrix

Learn the models corresponding to each class

- Build the vocabulary and reduce the representation space
- Group inflected forms of words around lemmas

(ex: infinitive for a verb, masculine singular for a noun)

Tokenization

I/saw/a/bat/./

- **"I saw a bat. "**
  - given a character sequence,
- tokenization is …
  - the task of chopping it up into pieces, called *tokens*

Required for TF-IDF representation and for word2vec

Tokenization

- **Methods for tokenization**
  - Simple Tokenization rule : chop on whitespace and throw away punctuation characters

I saw a bat.

(I, saw, a, bat)

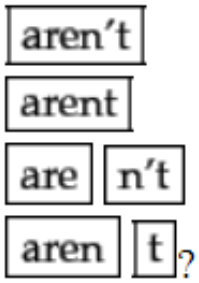
04.02.1985: I left San Francisco and went to Viêt-Nam. You were'nt born.

??? Tricky cases ???

Tokenization

I/saw/a/bat/./

- **Tricky cases**
  - Markers: dash (« - »), coma (« , »), tabulations (« »),
  - White space in « San Francisco »
  - End of sentence detection (find the dot (« . »)) : beware of acronyms E.N.S.T., numbers (3.14), and dates (02.05.2018)
  - uses of the apostrophe for possession and contractions



Tests under Acapella
Nous sommes le 02.05.2018. Il y a quelques années le nom de l'école était l'ENST ou mieux l'E.N.S.T.

Tokenization

- **More involved methods for word segmentation :**
  - Heuristic-based :
    - Use of a large vocabulary
    - Take the longest vocabulary match
      - Ex : I went to San Francisco -> (I, went, San Francisco)
    - Use some heuristics for unknown words

such methods make mistakes sometimes, and so you are never guaranteed a consistent unique tokenization.

Tokenization

- **More involved methods for word segmentation :**
  - Machine learning sequence models
    - trained over hand-segmented words
    - Ex: hidden Markov models, Conditional Random Fields, recurrent neural networks

such methods make mistakes sometimes, and so you are never guaranteed a consistent unique tokenization.

Reduce vocabulary size

- Motivation
  - Reduce vocabulary/representation space

	call	time	date	conference	release	meeting	corporation	earnings
document 1	2	1	3	2	1	1	1	
document 2	1		2	1	2	1	1	1
document 5		1	2		2	1	1	1
document 6	1	2	1	1	3	1	1	1
document 7	1						1	
document 8			1		1		1	1
document 9	2		1	3	1	1	1	1
document 10	2	1		1	1		1	1
document 13					1			2
document 14							3	
document 15	1			2			1	2

- Two options:
  - Removing tokens
  - Put together several words in one dimension

Removing tokens

more or less relevant according to the NLP task,

- Corresponding to Punctuation (??,!!, .)
  - Input : « Thank you ??!? »
  - After Tokenization :

    (Thank, you, ?, ?, !, ?)
  - After filtering:

    (Thank, you)

NB: punctuation is useful for opinion mining

Removing tokens

more or less relevant according to the NLP task,

- Belonging to
  - a predefined list (stop words) or
  - a certain type of grammatical categories (e.g. linking words, preposition, determinant, pronoun)

Input : « I like this actor but I am not convinced by his play »

Tokenization : (I, like, this, actor, but, I, am, not, convinced, by, his, play)

Filtering: (~~I~~, like, ~~this,~~ actor, ~~but,~~ I, am, not, convinced, ~~by,~~ ~~his,~~ play)

NB: linking words are useful for argument mining

Removing tokens

more or less relevant according to the NLP task,

- Hapax
  - Marginal terms (occuring once or twice) in the corpus
    - often corresponds to mispelling words
- Dates

NB: useful for Named Entity Recognition

Put together several words in one dimension

- **Gather inflectional forms and derivationally related forms**
  - Inflectional forms : a change in or addition to the form of a word that shows a change in the way it is used in sentences
  - Morphological derivation : the process of forming a new word from an existing word, often by adding a prefix or suffix, such as -ness or un-

PRACTICE :

ENGLISH :

propose inflectional forms of « dog », « sit »

propose derivational form of happy

FRENCH : donner les flexions du verbe « jouer »

Put together several words in one dimension

- **Gather inflectional forms and derivationally related forms of a word around …**
  - Their stems -> stemming
    - « cherchons » -> « cherch »
  - Their lemmas -> lemmatization
    - am, are, is => be
    - car, cars, car's, cars' => car

Put together several words in one dimension

- **Stemming and lemmatization are based on**
  - a morphological analysis of the words

Morphological analysis : an analysis of word internal structure

Morpheme : minimal linguistic unit carrying a sense (abstract unit)

Morphologic processes : flection, declension, conjugation, derivation (anti-constitu-tionn-elle-ment)

Put together several words in one dimension

- Stemming :
  - a crude heuristic process that chops off the ends of words (removal of derivational affixes)
  - How? Ex: Porter's algorithm based on morphological rules [Porter, 1980]

(F)	Rule			Example		
	SSES	→	SS	caresses	→	caress
	IES	→	I	ponies	→	poni
	SS	→	SS	caress	→	caress
	S	→		cats	→	cat

PRACTICE:

What is the stem of the word « frontal » in French?

Put together several words in one dimension

– Example of stemmer outputs

Sample text: Such an analysis can reveal features that are not easily visible from the variations in the individual genes and can lead to a picture of expression that is more biologically transparent and accessible to interpretation

Lovins stemmer: such an analys can reve featur that ar not eas vis from th vari in th individu gen and can lead to a pictur of expres that is mor biolog transpar and acces to interpres

Porter stemmer: such an analysi can reveal featur that ar not easili visibl from the variat in the individu gene and can lead to a pictur of express that is more biolog transpar and access to interpret

Paice stemmer: such an analys can rev feat that are not easy vis from the vary in the individ gen and can lead to a pict of express that is mor biolog transp and access to interpret

Put together several words in one dimension

- **Lemmatization** :

> "the boy's cars are different colors" =>
> "the boy car be different color"

- use of a vocabulary and morphological analysis of words,
- to return the *lemma:*
  - the base or dictionary form of a word: .
    - am, are, is => be
    - car, cars, car's, cars' => car

- NB: syntactic analysis is required to disambiguate some cases
  - Ex: « les poules du couvent couvent »
    - Couvent -> couvent (noun) ou couver (verb)

PRACTICE:
What are the stem and the lemma of the word « saw » in English?

Put together several words in one dimension

- **Lemmatization – existing tools**
  - For French
    - Treetagger
    - Xerox, Brill [Brill, 1995]
    - LIA_Tag, macaon http://macaon.lif.univ-mrs.fr/index.php?page=home-en
  - For English:
    - NLTK : http://www.nltk.org/
    - Treetagger

Xerox

[10/1996, 10/1997]

La	le	+DET_SG
petite	petit	+ADJ2_SG
ferme	ferme	+NOUN_SG
du	de=le	+PREP_DE
père	père	+NOUN_SG
Fouchard	Fouchard	+NOUN_INV
se	se	+PC
trouvait	trouver	+VERB_P3SG
au sortir du	au sortir de=le	+PREP
défilé	défilé	+NOUN_SG
.	.	+SENT

Lemma

Part of Speech

Put together several words in one dimension

- **Stemming vs. Lemmatization**
  - What is the best choice?
    - It depends on the language
      - Ex: stemming works well in German

Outline

- Pre-processing using linguistic knowledge for machine learning
- **Knowledge-based methods (hybrid methods)**

Knowledge-based/rule-based methods - Introduction

Could be used as resources to make hand-crafted features for machine learning approaches

Rule-based methods

Describe the information to be extracted (business concept, topic, opinion)

1/ using resources (ex: WordNet), Building lexicons, ontologies,

2/ define linguistic/syntactic patterns (ex :automaton, grammars)

« manque de qualité de service »

« il n'y a vraiment pas eu de contact », …

Concept

DIS**SATISFACTION**

Rule-based methods

- Linguistic/syntactic patterns
  - Regular expressions that can call
    - Lemmas : ex. *« avoir »*
    - Syntactic categories (parts of speech) : « *#PREP_DE* » « *#NEG* »
    - Predefined lexicons: « *~services-lex* »

« manque de qualité de service »

« il n'y a vraiment pas eu de contact », …

→ Concept DIS**SATISFACTION**

*(manque|~negation-patt|(il/#NEG/y/avoir/~negation-patt))/(#PREP_DE)?/ (conseil|contact|~services-lex)**

* syntax of TEMIS tool

Rule-based methods using regular expressions

- **Syntaxe courante (Unix, perl, etc.)**

Expression	Langage accepté
`r*`	0 ou plusieurs `r`
`r+`	1 ou plusieurs `r`
`r?`	0 ou 1 `r`
`[abc]`	`a` ou `b` ou `c`
`[a-z]`	N'importe quel caractère dans l'intervalle `a...z`
`.`	N'importe quel caractère sauf `\n`
`[^s]`	N'importe quel caractère sauf ceux de `s`
`r{m,n}`	Entre `m` et `n` occurences de `r`
`r1 r2`	La concaténation de `r1` et `r2`

Expression	Langage accepté
`r1 | r2`	`r1` ou `r2`
`(r)`	`r`
`^r`	`r` en début de ligne
`r$`	`r` en fin de ligne
`"s"`	Le string `s`
`\c`	Le caractère `c`
`r1 / r2`	`r1` quand il est suivi de `r2`

- `[a-zA-z]` Une lettre.
- `[0-9]` Un chiffre.
- `a[^A-Za-z]b` Un `a`, suivi d'un caractère non alphabétique, suivi d'un `b`.
- `^Monsieur` `Monsieur` en début de ligne.
- `[a-zA-Z]([a-zA-Z]|[0-9])*` Un identifiant Pascal. ...

Tiré de http://www.ulb.ac.be/di/ssd/ggeeraer/lg/extexpreg_print.pdf

Rule-based methods using regular expressions - Practice

- **Give the regular expression parsing the sentences fulfilling the following criteria:**
  - The first word is an uppercase
  - The last character is a full stop
  - The sentence contains one or more words (characters a...z et A...Z) separated by a space.

Check regexp : http://www.regexplanet.com/advanced/java/index.html
https://regex101.com/

From http://www.ulb.ac.be/di/ssd/ggeeraer/lg/extexpreg_print.pdf

Rule-based methods using regular expressions - Practice

- **Give the regular expression parsing the sentences fulfilling the following criteria:**
  - The first word is an uppercase
  - The last character is a full stop
  - The sentence contains one or more words (characters a...z et A...Z) separated by a space.

Expression	Langage accepté
`r*`	0 ou plusieurs `r`
`r+`	1 ou plusieurs `r`
`r?`	0 ou 1 `r`
`[abc]`	`a` ou `b` ou `c`
`[a-z]`	N'importe quel caractère dans l'intervalle `a...z`
`.`	N'importe quel caractère sauf `\n`
`[^s]`	N'importe quel caractère sauf ceux de `s`
`r{m,n}`	Entre `m` et `n` occurences de `r`
`r1 r2`	La concaténation de `r1` et `r2`

Expression	Langage accepté
`r1 | r2`	`r1` ou `r2`
`(r)`	`r`
`^r`	`r` en début de ligne
`r$`	`r` en fin de ligne
`"s"`	Le string `s`
`\c`	Le caractère `c`
`r1 / r2`	`r1` quand il est suivi de `r2`

- `[a-zA-z]` Une lettre.
- `[0-9]` Un chiffre.
- `a[^A-Za-z]b` Un `a`, suivi d'un caractère non alphabétique, suivi d'un `b`.
- `^Monsieur` `Monsieur` en début de ligne.
- `[a-zA-Z]([a-zA-Z]|[0-9])*` Un identifiant Pascal. ...

Tiré de http://www.ulb.ac.be/di/ssd/ggeeraer/lg/extexpreg_print.pdf

Tools

- Unitex : http://www-igm.univ-mlv.fr/~unitex/
- Grammars of NLTK
- Gate

External Lexical Resources called by lexicon-based rules

External Lexical Resources called by rules

- **Wordnet : lexical database**
  - Retrieve information on word meaning/sense
  - Core idea :
    - A word can have several meanings (ex: « bat »)
    - groups English words into *synsets*
    - Synsets : set of synonyms

PRACTICE :

Let's search the word « estimable » on Wordnet website for English
http://wordnetweb.princeton.edu/perl/webwn

Q1 : how many senses are existing for this word?
Q2 : what is the size of the synset of **estimable#2?**

Resources

- **Wordnet : lexical database**
  - Synsets : set of synonyms
    - Synonyms : words that are interchangeable in some context without changing the truth value of the proposition
    - Synsets include simplex words as well as collocations like "eat out" and "car pool.“
    - The meaning of a synset is further clarified with a short definition and one or more usage examples

Example :

good, right, ripe – (most suitable or right for a particular purpose; "a good time to plant tomatoes"; "the right time to act"; "the time is ripe for great sociological changes")

Resources

- **Wordnet : lexical database**
  - All synsets are connected to other synsets by means of semantic relations:
    - Ex: canine is a hypernym of dog
    - *window* is a meronym of *building*

PRACTICE
On wordnet
To see the semantic relation click on the S

Version française : Wordnet Libre du Français (WOLF) :
http://alpage.inria.fr/~sagot/wolf.html

Specific resources for opinions

- SentiWordNet http://sentiwordnet.isti.cnr.it/
  - Relies on Wordnet: lexical database
    - Principle : *synsets*
    - English Version: http://wordnetweb.princeton.edu/perl/webwn
    - French Version: Wordnet Libre du Français (WOLF) : http://alpage.inria.fr/~sagot/wolf.html

Lexicon of opinions in English

- SentiWordNet http://sentiwordnet.isti.cnr.it/
  - Principle : add to each synset a positive score, a negative score AND an objective score between 0 and 1
    - [estimable(J,3)] “may be computed or estimated”

      `Pos 0 Neg 0 Obj 1`
    - [estimable(J,1)] “deserving of respect or high regard”

      `Pos .75 Neg 0 Obj .25`

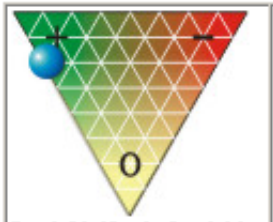
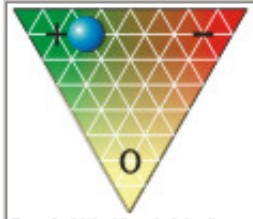
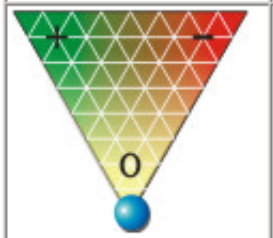
Lexicon of opinions in English

- SentiWordNet

estimable | Search word | show position

Adjective

3 senses found.

 P = 0.75, N = 0, O = 0.25	estimable(1) *deserving of respect or high regard*
 P = 0.625, N = 0.25, O = 0.125	honorable(5) good(4) respectable(2) estimable(2) *deserving of esteem and respect; "all respectable companies give guarantees"; "ruined the family's good name"*
 P = 0, N = 0, O = 1	computable(1) estimable(3) *may be computed or estimated; "a calculable risk"; "computable odds"; "estimable assets"*

main page

(c) Andrea Esuli 2005 - andrea.esuli@isti.cnr.it

Lexicon of opinions in English

- **Wordnet affect**
  - Selecting a subset of wordnet Affective label + valence

Etiquette affective	Exemples de synsets associés
Emotion	nom ANGER#1, verbe FEAR#1
Mood	nom ANIMOSITY#1, adjectif AMIABLE#1
Trait	nom AGGRESSIVENESS#1, adjectif COMPETITIVE#1
Cognitive State	nom CONFUSION#2, adjectif DAZED#2
Physical State	nom ILLNESS#1, adjectif ALL IN#1
Edonic Signal	nom HURT#3, nom SUFFERING#4
Emotion-Eliciting Situation	nom AWKWARDNESS#3, adjectif OUT OF DANGER#1
Emotional Response	nom COLD SWEAT#1, verbe TREMBLE#2
Behaviour	nom OFFENSE#1, adjectif INHIBITED#1
Attitude	nom INTOLERANCE#1, nom DEFENSIVE#1
Sensation	nom COLDNESS#1, verbe FEEL#3

Tiré de https://www.proxem.com/Download/Research/BDL-CA07-WordNet_et_son_ecosysteme-Francois_Chaumartin.pdf

Lexicon of opinions in English

- **LIWC (Linguistic Inquiry and Word Count) - Pennebaker, J.W., Booth, R.J., & Francis, M.E. (2007).**
- **Home page: http://www.liwc.net/**
- **2300 mots, >70 classes**
- **Version française : http://sites.univ-provence.fr/wpsycle/outils_recherche/liwc/FrenchLIWCDictionary_V1_1.dic**

French LIWC

Tableau 1
Les 80 descripteurs analysés par le LIWC 2007 version anglaise (extrait de Pennebaker et al., 2007 ; NB : entre parenthèses l'effectif de radicaux présents dans le dictionnaire anglais).

Processus linguistiques	Processus psychologiques	Préoccupations personnelles	Dimensions du langage oral	Ponctuation
Total de mots	Processus sociaux (465)	Travail (327)	Consentement (30)	Total
Mots par phrase	Famille (64)	Accomplissement (186)	Phatiques (8)	Points
Mots du dictionnaire	Amis (37)	Loisirs (229)	Remplisseurs (9)	Virgules
Mots de plus de 6 lettres	Humains (61)	Maison (93)		Doubles points
Total de mots fonctionnels (464)	Processus affectifs (915)	Argent (173)		Points virgules
Total des pronoms (116)	Émotions positives (406)	Religion (159)		Points d'interrogation
Pronoms personnels (70)	Émotions négatives (499)	Mort (62)		Points d'exclamation
1er personne du singulier (12)	Anxiété (91)			Tirets
1er personne du pluriel (12)	Colère (184)			Guillemets
2e personne (20)	Tristesse (101)			Apostrophes
3e personne du singulier (17)	Processus cognitifs (730)			Parenthèses
3e personne du pluriel (10)	Perspicacité (195)			Autres ponctuations
Pronoms impersonnels (46)	Causation (108)			
Articles (3)	Divergence (76)			
Verbes (383)	Tentative (155)			
Verbes auxiliaires (144)	Certitude (83)			
Verbes au passé (145)	Inhibition (111)			
Verbes au présent (169)	Inclusion (18)			
Verbes au futur (48)	Exclusion (17)			
Adverbes (69)	Processus perceptifs (273)			
Prépositions (60)	Vue (72)			
Conjonctions (28)	Audition (51)			
Négations (57)	Toucher (75)			
Quantifieurs (89)	Processus biologiques (567)			
Nombres (34)	Corps (180)			
Jurons (53)	Santé (236)			
	Sexualité (96)			
	Alimentation (111)			
	Relativité (638)			
	Mouvement (168)			
	Espace (220)			
	Temps (239)			

A. Piolat et al. / Psychologie française 56 (2011) 145–159

Lexicon-based methods for sentiment analysis

Could be used as resources to make hand-crafted features for machine learning approaches

The simplest rule : Keyword spotting

- The text is classified in the category of opinions corresponding to the presence of opinionated words (using lexicons)
- « je suis content » => positive

- **Limits :**
  - Do not process negation
    - « je ne suis pas content » => positive
  - Ignore words that are implicitly positive or negative
    - « le réchauffement climatique » (global warming)

More complex rules

- **In order to deal with :**
  - negation processing (I don't like this movie)
  - modifiers and intensifiers (the plot is not very good)
  - conditional tense
  - discourse markers

- **Ex: Taboaba et al. : Lexicon-Based methods for sentiment analysis**
- **Principle:**
  - Attributes an SO (Semantic Orientation) between -5 and 5 to adjectives, nouns, verbs and adverbs

Word
monstrosity
hate (noun and verb)
disgust
sham
fabricate
delay (noun and verb)
determination
inspire
inspiration
endear
relish (verb)
masterpiece

Table 1
Examples of words in the noun and verb dictionaries.

Word	SO Value
monstrosity	−5
hate (noun and verb)	−4
disgust	−3
sham	−3
fabricate	−2
delay (noun and verb)	−1
determination	1
inspire	2
inspiration	2
endear	3
relish (verb)	4
masterpiece	5

More complex rules

- **Taboaba et al. : Lexicon-Based methods for sentiment analysis**
- **Principle:**
  - Intensifier management: modification of SO

Table 3
Percentages for some intensifiers.

Intensifier	Modifier (%)
slightly	−50
somewhat	−30
pretty	−10
really	+15
very	+25
extraordinarily	+50
(the) most	+100

EXO :
If *sleazy* has an SO at 3, what is the SO of *somewhat sleazy ?*
If excellent has an SO at 5, what is the SO of *most excellent* ?

More complex rules

- **Taboaba et al. : Lexicon-Based methods for sentiment analysis**
- **Principle:**
  - Negation processing :
    - switch negation for the simplest cases (good(+3), not good(-3))
    - Search for negation in more complicated cases
      - Ex: « Nobody gives a good performance in this movie »
  - Manage the « Irrealis blocking »: ex: « would »
    - « This should have been a great movie »(SO = 3 -> SO =0)

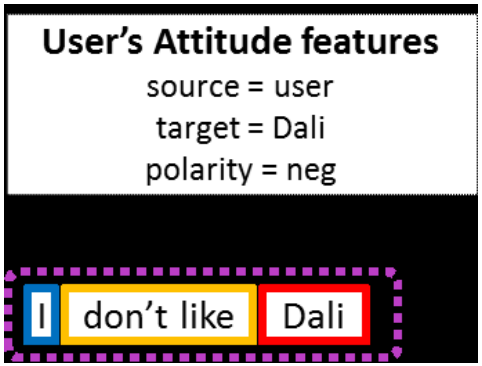
Syntactic-based rules for sentiment analysis

Syntactic-based rules for sentiment analysis

- Objectives:
  - To identify more complex structures and phenomena relying on theoretical models of sentiment
- Rules that are built on the syntactic structure of the sentence and on lexical resources
  - Compositional approaches
  - Syntactic patterns

Theoretical models to formalize the structure of a sentiment

- Example : Appraisal theory from systemic functional linguistics [Martin and White, 2005]
  - An appraisal expression is a source that evaluates a target -> 3 components.



Theoretical models to formalize the structure of a sentiment

- Properties of the model
  - Identify the opinion target and source
  - Distinguish evaluative stances : Affect/Judgment/Appreciation
    - Affect : personal reaction referring to an emotional state
    - Judgment : assigning qualities - e.g. tenacity - to individuals according to normative principles
    - Appreciation : Evaluation of an object - e.g. a product or a process

    - PRACTICE :
      - ‘The shop assistant’s behavior was really unfriendly’
      - ‘He is frightened of spiders’
      - ‘Plastic bags are environment unfriendly’

‘He is frightened of spiders’

-> Affect : personal reaction referring to an emotional state

‘The shop assistant’s behavior was really unfriendly’

-> Judgment : assigning qualities - e.g. tenacity - to individuals according to normative principles

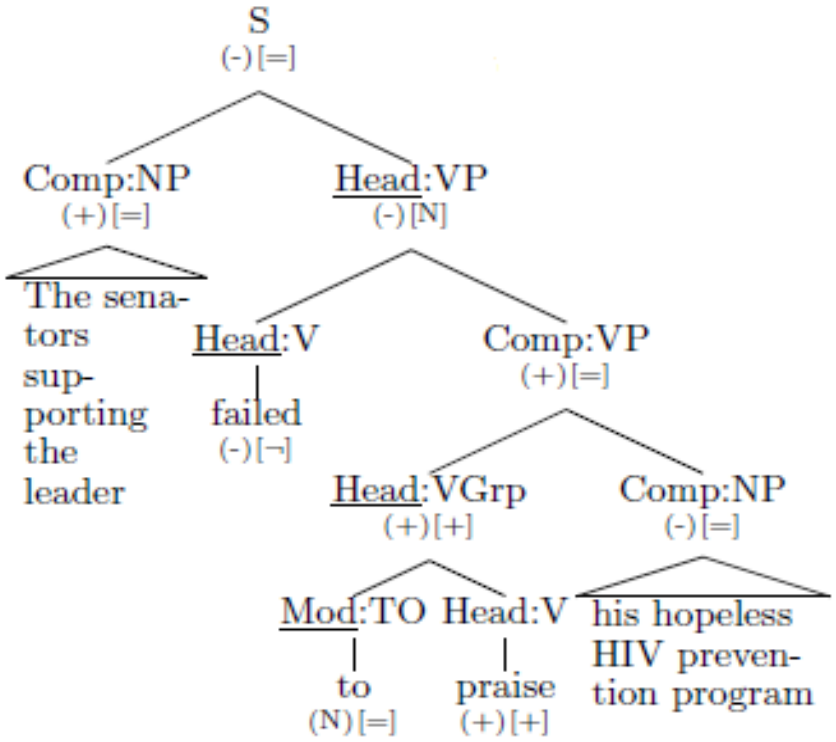
‘Plastic bags are environment unfriendly’

-> Appreciation : Evaluation of an object - e.g. a product or a process

Compositional approach

- Representation of the sentence by constituents [Moilanen 2007] :

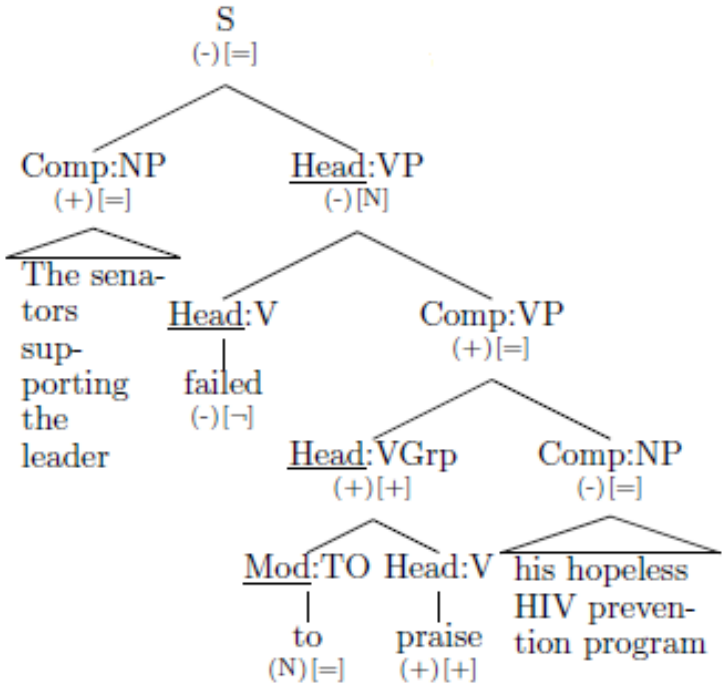
« The senators supporting the leader failed to praise his hopeless HIV prevention program »



Compositional approach

- **Calculates the overall polarity of an output constituent from the input constituents**

- Propagation rules:
  - the polarity of a neutral constituent is "erased" by that of a non-neutral constituent
  - {(+)(N)} → (+)
  - {(-)(N)} → (-)
  - Ex: POS(to admire) his behavior -> POS (to admire his behavior)
- Inversion rules:
  - (+) → (-) ; (-) → (+) in order to deal with negation,
  - for example : never POS(succeed)-> NEG(never succeed)
- Polarity conflict resolution rules:
  - when the two polarities are conflicting at different levels of the syntactic structure
  - Ex POS



Syntactic pattern

- Ex1:

Pattern : It-LinkVb-ADJ-CLAUSE

LinkVB : Linking Verb (verbs that link the subject to a subject complement)

CLAUSE : a group of words containing a subject and predicate and functioning as a member of a complex sentence. The sentence "When it rained they went inside" consists of two clauses: "when it rained" and "they went inside."

Ex : 'It was distressing to hear her talking like that'
LinkVB : 'was'
CLAUSE: 'to hear her talking like that'
ADJ linked to expression of opinion phenomena: distressing

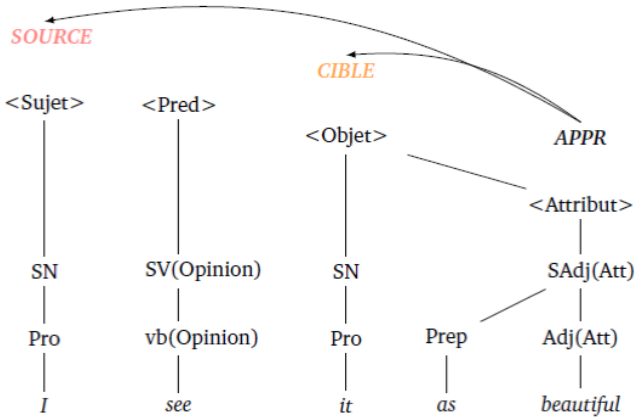
Distinguish different opinion phenomena

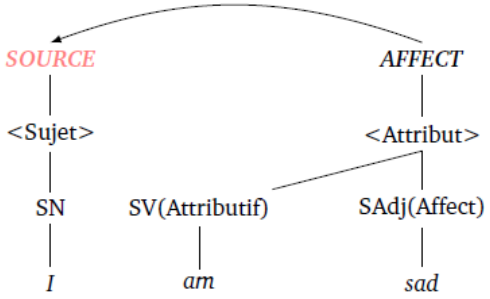
- depending on the categories of used adjectives -> distinguish Affect, judgment and appreciation
  - Affect : personal emotional state
    - 'It was distressing to hear her talking like that'
    - adjective referring to the source's inner feeling
  - Judgment : social or ethical appraisal of other's behavior
    - 'It was reasonable to incur that expense'
    - adjective referring to ethical norm
  - Appreciation : evaluation of phenomena
    - 'It was wonderful talking to you the other day'
    - adjective referring to good/bad scale

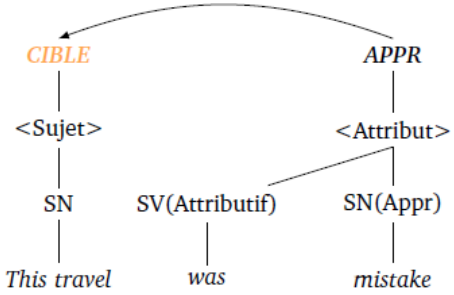
Bednarek, Monika. "Language patterns and ATTITUDE." *Functions of language* 16.2 (2009): 165-192.

Syntactic pattern

- Ex2: identifying syntagms
  - referring to an evaluation with an attribute position
  - using or not a prepositional syntagm (‘I see’)







C. Langlet and C. Clavel, Improving social relationships in face-to-face human-agent interactions: when the agent wants to know users likes and dislikes , in ACL 2015

Schemas From Caroline Langlet’s Phd dissertation

Attribution of chunk semantic role (target and source)

- Target/Source/Evaluation structures can take many different forms
  - ‘I am sad’
  - ‘This travel was a mistake’
  - ‘I see it as beautiful’

PRACTICE : identify the target/source/evaluation tokens in the previous sentences. Is the evaluation an affect, a judgment or an appreciation?

Rules for attribution of chunk semantic role (target and source)

- Target/Source/Evaluation structures can take many different forms
  - SOURCE-EVALUATION (Affect) ‘I am sad’
  - TARGET-EVALUATION (Appreciation) ‘This travel was a mistake’
  - SOURCE-TARGET-EVALUATION (Appreciation) ‘I see it as beautiful’

Syntagm referring to an attitude with an attribute position using or not a prepositional syntagme (I see)

How to do a syntactic analysis?

For example, for syntactic-based rules for sentiment analysis

Part of Speech tagging

- **Task for the attribution of grammatical categories of a token**

INPUT:

Profits soared at Boeing Co., easily topping forecasts on Wall Street, as their CEO Alan Mulally announced first quarter results.

OUTPUT:

Profits/N soared/V at/P Boeing/N Co./N ,/, easily/ADV topping/V forecasts/N on/P Wall/N Street/N ,/, as/P their/POSS CEO/N Alan/N Mulally/N announced/V first/ADJ quarter/N results/N ./.

N = Noun
V = Verb
P = Preposition
Adv = Adverb
Adj = Adjective
...

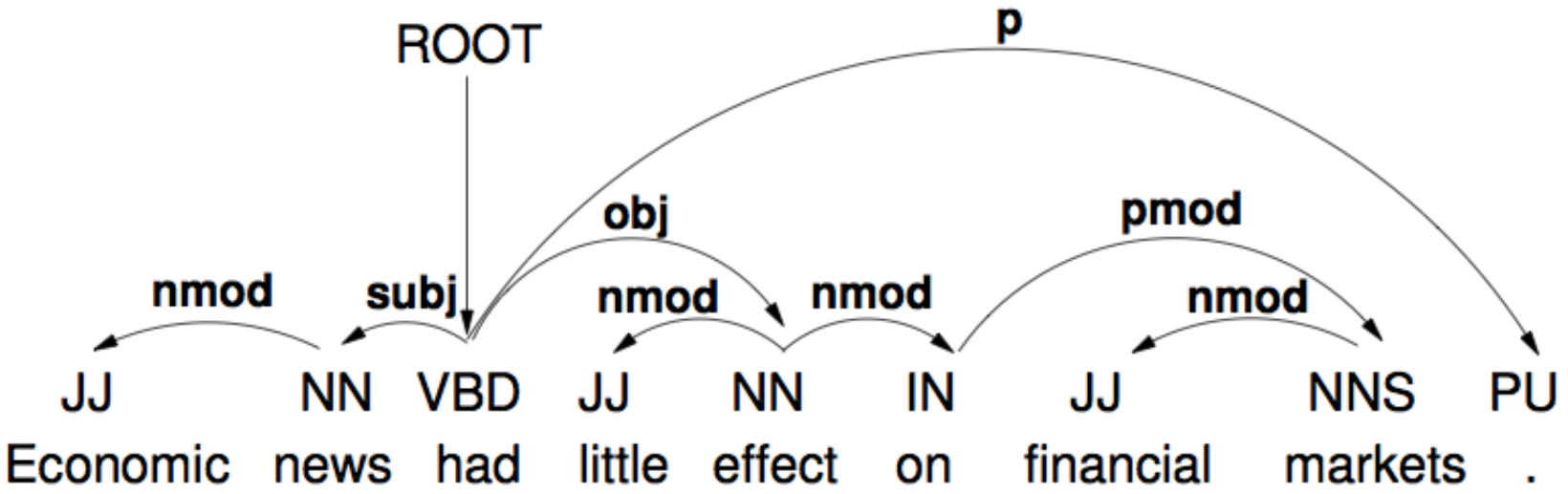
Chunking task

[John] [talked] [to the children][about drugs]

- **Detection of syntactic components :**
  - Noun phrase (groupe nominal), verbal phrase (noyau verbal), *etc.*
  - Borderline detection
  - Phrase labelling

Dependency parsing task

- Describe the syntactic structure of a sentence in terms of the words (or lemmas) in a sentence and an associated set of directed binary grammatical relations that hold among the words
- Relations among the words are illustrated with directed, labeled arcs from heads to dependents
- Ex : Economic is a nominal modifier of news



Methods and challenges for syntactic analysis

- **2 types of methods**
  - Based on linguistic expertise
  - Machine learning:
    - From a BIG labelled corpus/database
    - Learn the probabilities for the different transitions between syntactic categories

Modelling linguistic expertise for POS tagging/chunking/dependency parsing

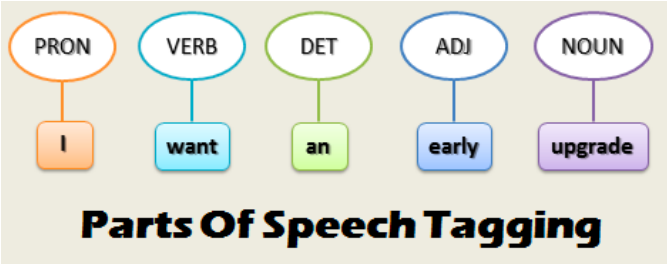
- **Example :**
  - DET/PRO V -> PRO V
  - NP (Noun Phrase) : DET ADJ* NN ADJ*
- **Strengths :**
  - Readable rules,
  - Errors are easier to understand
- **Weaknesses**
  - Not robust to noisy inputs and out of vocabulary words
  - Time-consuming

Machine learning for POS tagging

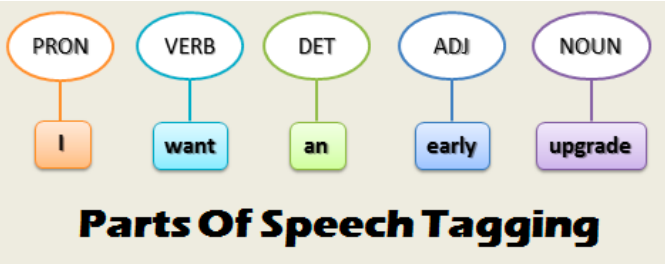
- **Problem formulation for Hidden Markov Models**

$$\begin{array}{lccccccccc} M = & \cdot & \cdot & \cdot & w_{i-2} & w_{i-1} & w_i & \cdot & \cdot & \text{mots} \\ E = & \cdot & \cdot & \cdot & e_{i-2} & e_{i-1} & e_i & \cdot & \cdot & \text{étiquettes} \end{array}$$

  - Labelled corpus :
    - sequences of pairs (word, syntactic category)
  - Graphical model, markov chain
    - learn the probabilities for the different transitions between the nodes of the graph (e.g. between syntactic categories)



POS tagging with Hidden Markov Models



- Simplifying hypothesis:
  - Markov assumption : first-order markov chain (the probability of a particular state depends only on the previous state)

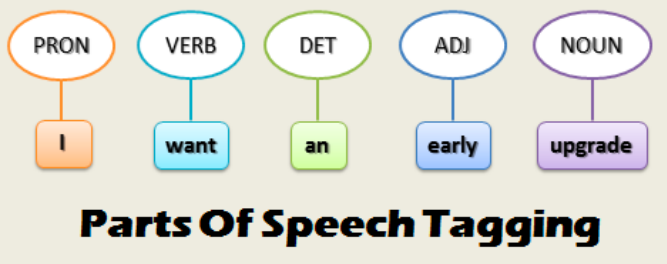
$$P(q_i|q_1 \ldots q_{i-1}) = P(q_i|q_{i-1})$$

  - Conditionally to the labels, words are independent :

$$p(w_i|e_1 \ldots e_i, w_1 \ldots w_i) = p(w_i|e_i)$$

POS tagging with Hidden Markov Models

- Training : learning the model λ = (**A, B, Π**) from a labeled corpus
  - A : CxC transition state matrix
    - e.g. probability to have a VERB after a DET
  - B : NxC matrix of the probabilities of the observation i in state j
    - e.g. probability to generate « want » if the state is a verb
  - Distribution Π of the initial state : vector of length C
    - E.g. probability to begin with a PRON



N: number of distinct observations (vocabulary size)
C: number of grammatical categories

POS tagging with Hidden Markov Models

$$
\begin{array}{lccccccccc}
M = & \cdot & \cdot & \cdot & w_{i-2} & w_{i-1} & w_i & \cdot & \cdot & \text{mots} \\
E = & \cdot & \cdot & \cdot & e_{i-2} & e_{i-1} & e_i & \cdot & \cdot & \text{étiquettes}
\end{array}
$$

- Training :
  - Learning the model λ = (**A, B, Π**) from a labeled corpus
- Decision:
  - find the best sequence E that maximizes the model for the sequence of words M

Syntactic analysis – Remaining challenges

- **Challenges**
  - Disambiguation : « La petite brise la glace » ?
  - Capable of handling mistakes and typos

Towards a better choice of sentiment categories

Task description and challenges

- Challenges:
  - choose the relevant phenomenon according to the application and the data [Clavel and Callejas, 2016]
- Examples
  - Detect when the student is frustrated or bored in e-learning system or detect when the user is upset in a human-machine dialogue system => emotion

Challenges

- Examples
  - Detect depressed persons at home in robot companion systems => mood

  - Detect extrovert personality in job interview data => personality traits

Challenge

- Opinion/emotion classes
  - Define among the wide variety of existing Sentiment/opinion-related phenomena
    - Emotion, opinion, sentiment, mood, attitude, interpersonal stance, personality traits, affect, judgment, appreciation, argumentation, engagement
  - … the classes that will be recognized by the system (anger vs. discontent) or that will give access the contents of the corpus for a better understanding of the behavior of the system

Challenge : Opinion/emotion classes

- Scherer's definitions [Scherer, 2005]
  - Emotion: short phenomenon, physiological reaction, appraisal of a major event (stimulus)
  - Mood: diffuse non-caused low-intensity long-duration change in subjective feeling
  - Interpersonal stances: affective stance toward another person in a specific interaction
  - Attitudes: enduring, affectively colored beliefs, dispositions towards objects or persons
  - Personality traits: stable personality dispositions and typical behavior tendencies
- PRACTICE : link the following terms to the most relevant phenomenon
  - liking, gloomy, contemptuous, jealous, sad

Standardization

- Well defined for emotions
  - Emotion : Emotion Markup Language
  - http://www.w3.org/TR/2014/REC-emotionml-20140522/
- **Still little about opinions and sentiments**
  - Opinion and sentiments
    http://www.w3.org/community/sentiment/ : Linked Data
  - Models for Emotion and Sentiment Analysis Community Group

Materials to go further

- NLP in general
  - **https://nlp.stanford.edu/IR-book**
  - **From Miha Grcar “Text mining and Text stream mining tutorial**
  - **Foundations of Statistical Natural Language Processing Christopher D. Manning and Hinrich Schütze**
  - **Lecture from Stanford http://cs224d.stanford.edu/lecture_notes/notes1.pdf**

Materials to go further

- NLP and deep learning
  - Deep Natural Language Processing course offered in Hilary Term 2017 at the University of Oxford.
  - Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. Deep learning. MIT press, 2016.
  - Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." Advances in neural information processing systems. 2013.

Materials to go further

- Tools :
  - word2vec from Google https://code.google.com/p/word2vec/ tutorial from
  - tensorflow https://www.tensorflow.org/tutorials/word2vec
  - Other representation : Glove http://nlp.stanford.edu/projects/glove/
- Sentiment definitions
  - [MUN 14] MUNEZERO M. D., SUERO MONTERO C., SUTINEN E., PAJUNEN J., “Are They Different? Affect, Feeling, Emotion, Sentiment, and Opinion Detection in Text”, IEEE Transactions on Affective Computing, 2014.

Materials to go further

- In French :
  - Une petite introduction au traitement automatique des langues naturelles par François Yvon http://perso.limsi.fr/Individu/anne/coursM2R/intro.pdf
  - Introduction au TALN et à l'ingénierie linguistique par Isabelle Tellier http://www.lattice.cnrs.fr/sites/itellier/poly_info_ling/info-ling.pdf