UNSUPERVISED APPROACHES, SEMANTIC MODELING

Vincent Guigue

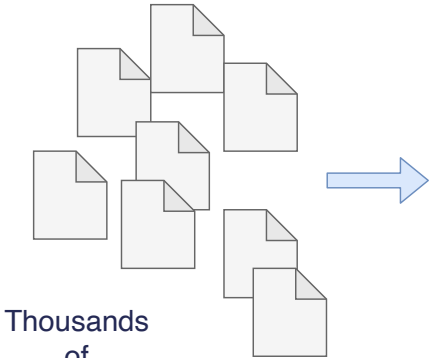
SCIENCES SORBONNE UNIVERSITÉ

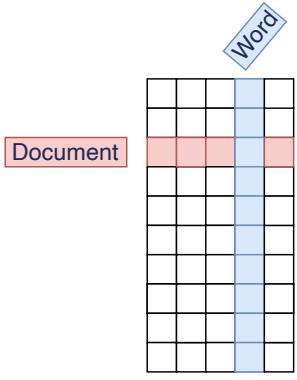
ISIR INSTITUT DES SYSTÈMES INTELLIGENTS ET DE ROBOTIQUE

MLIA Machine Learning & Deep Learning for Information Access

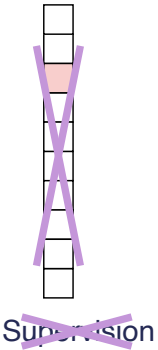
Introduction

What can we do... Without supervision?





Data matrix

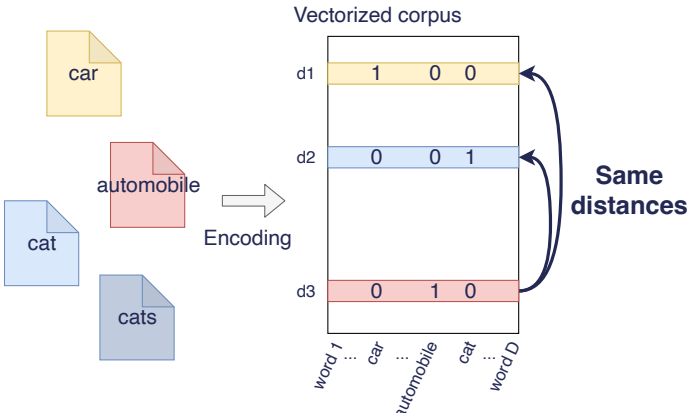


Clustering

Semantic analysis

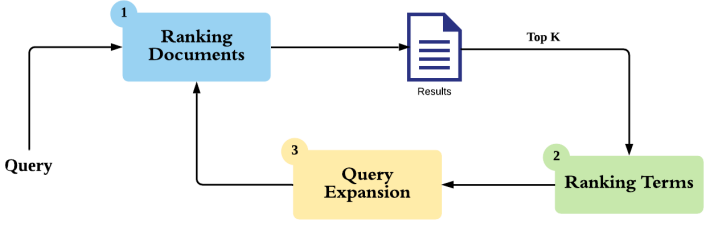
Bag of words limits

- No context modeling
  - Negative form
  - Disambiguation
- Semantic gap



Extensions

- N-gram encoding ⇒ group of words
  - *very good*
  - *not good*
  - Combinatorial dictionary ⇒ dimension issue !
- Lemmatization/stemming
  - 1 lexical stem = 1 column
- Rocchio's strategy
  - Pseudo Relevance Feedback
  - Query expansion



Semantic & ontologies

Objective

Understanding (automatically) word meaning

... And eliminating the semantic gap

$\Rightarrow$ Applications

- Information Retrieval
- Topic classification (& extraction)
- Information extraction
- Automated Summary
- Opinion classification
- ...

WordNet

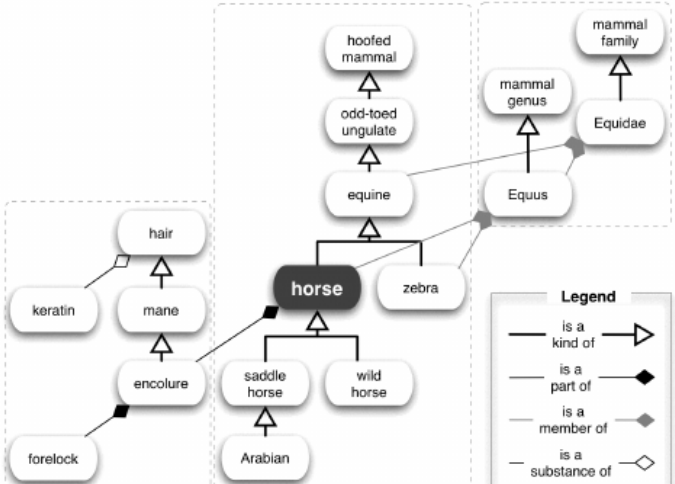
- Description: Hierarchical description of words
  - Nouns
  - Verbs
  - Adjectives

WordNet

- Description: Hierarchical description of words
  - Nouns
    - **hypernyms**: Y is a hypernym of X if every X is a (kind of) Y (canine is a hypernym of dog)
    - **hyponyms**: Y is a hyponym of X if every Y is a (kind of) X (dog is a hyponym of canine)
    - coordinate terms: Y is a coordinate term of X if X and Y share a hypernym (wolf is a coordinate term of dog, and dog is a coordinate term of wolf)
    - **meronym**: Y is a meronym of X if Y is a part of X (window is a meronym of building)
    - **holonym**: Y is a holonym of X if X is a part of Y (building is a holonym of window)
  - Verbs
  - Adjectives

WordNet

- Description: Hierarchical description of words



WordNet

- Description: Hierarchical description of words
  - Nouns
  - Verbs
    - **hypernym**: the verb Y is a hypernym of the verb X if the activity X is a (kind of) Y (to perceive is an hypernym of to listen)
    - **troponym**: the verb Y is a troponym of the verb X if the activity Y is doing X in some manner (to lisp is a troponym of to talk)
    - **entailment**: the verb Y is entailed by X if by doing X you must be doing Y (to sleep is entailed by to snore)
    - **coordinate terms**: those verbs sharing a common hypernym (to lisp and to yell)
  - Adjectives

WordNet

- Description: Hierarchical description of words
  - Nouns
  - Verbs
  - Adjectives
    - **Antomyms / Synonyms**

WordNet: Metrics

- Metrics in WordNet
  - Length of the shortest path in the graph
  - Length of the shortest path in the *synonym* graph,
  - Distance of the first common ancestor,
  - cf: Leacock Chodorow (1998), Jiang Conrath (1997), Resnik (1995), Lin (1998), Wu Palmer (1993)
- WordNet & metrics are available in NLTK

WordNet: Limits & usage

- Fully depend on static resources
  - New expressions + technical/specialized vocabulary may lack
  - Social network mining, Hashtags ...

Existing extensions:

- Several translations
- More generally : **a powerful diffusion tool**
  - Characterizing one part of the vocabulary
    + using WordNet to spread characterization (synonyms...)
- Applications
  - IR: Information Retrieval
  - Word Desambiguation
  - Text Classification
  - Machine Translation
  - Summarization

The General Inquirer

- Home page: `http://www.wjh.harvard.edu/~inquirer`
- List of Categories: `http://www.wjh.harvard.edu/~inquirer/homecat.htm`
- Spreadsheet: `http://www.wjh.harvard.edu/~inquirer/inquirerbasic.xls`
- Categories:
  - Positive (1915 words) and Negative (2291 words)
  - Strong vs Weak, Active vs Passive, Overstated versus Understated
  - Pleasure, Pain, Virtue, Vice, Motivation, Cognitive Orientation, etc
- Free for Research Use

Philip J. Stone, Dexter C Dunphy, Marshall S. Smith, Daniel M. Ogilvie. - MIT Press, 1966

The General Inquirer: A Computer Approach to Content Analysis

LIWC (Linguistic Inquiry and Word Count)

- Home page: `http://www.liwc.net/`
- 2300 words, >70 classes
- Affective Processes
  - negative emotion (bad, weird, hate, problem, tough)
  - positive emotion (love, nice, sweet)
- Cognitive Processes
  - Tentative (maybe, perhaps, guess), Inhibition (block, constraint)
  - Pronouns, Negation (no, never), Quantifiers (few, many)
- $30 or $90 fee

Pennebaker, J.W., Booth, R.J., & Francis, M.E. 2007. Austin, TX
Linguistic Inquiry and Word Count: LIWC

MPQA Subjectivity Cues Lexicon

- Home page: `http://www.cs.pitt.edu/mpqa/subj_lexicon.html`
- 6885 words from 8221 lemmas
  - 2718 positive
  - 4912 negative
- Each word annotated for intensity (strong, weak)
- GNU GPL

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann, EMNLP 2005
Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis

Bing Liu Opinion Lexicon

- Bing Liu's Page on Opinion Mining
  `http://www.cs.uic.edu/~liub/FBS/opinion-lexicon-English.rar`
- 6786 words
  - 2006 positive
  - 4783 negative

Minqing Hu and Bing Liu. ACM SIGKDD-2004.
Mining and Summarizing Customer Reviews

SentiWordNet

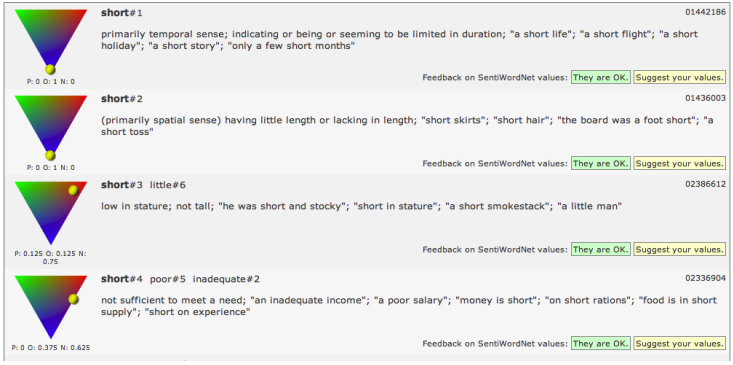
- Home page: `http://sentiwordnet.isti.cnr.it/`
- All WordNet synsets automatically annotated for degrees of:
  - positivity, negativity, and neutrality/objectiveness
- Many contexts investigated

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. LREC-2010

SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining

With an example: **short**

ADJECTIVE



Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. LREC-2010
SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining

[C. Potts] Disagreements between polarity lexicons

	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				

Building Lexicons or semantics (for sentiment analysis)

Overall Philosophy

Target:

- Extracting the meaning of words and patterns of words
- ... Namely, understanding the message and deducing the polarity

$\Rightarrow$ Building Universal Models

Important tasks and subtasks:

- Building/learning/using lexical resources
- Extracting complex sentiment patterns
- Dealing with different problems related to sentiment defintion ($(e_j, a_{jk}, so_{ijkl}, h_i, t_l)$, entity, feature, polarity, holder, time)

Stanford NLP tools : `http://nlp.stanford.edu`
Named Entity Recognition, Dependency Tree Building, POS Tagging...

Opinionated Lexicons Building

Alternative 1:

1. Getting a lexicon with synonymous (e.g. WordNet)
2. Handmade opinion reference list:
   - *good*, *poor*...
3. Diffusion of the polarity according to the synonymous graph

Alternative 2:

1. Handmade opinion reference list:
   - *good*, *poor*...
2. Diffusion with external sources:
   - corpus (with labels or not)
   - search engines

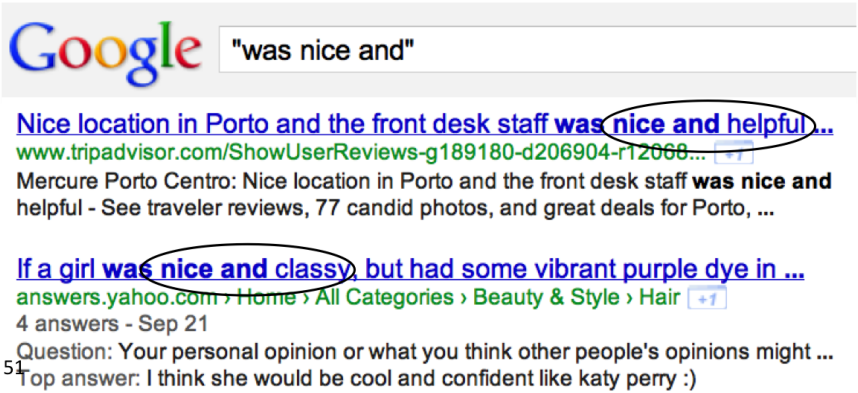
Hatzivassiloglou and McKeown 1997

Hypothesis :

- Adjectives separated by **and** $\Rightarrow$ same polarity
  - Fair **and** legitimate, corrupt **and** brutal
  - fair **and** brutal, corrupt **and** legitimate
- Adjectives separated by **but** $\Rightarrow$ different polarity
  - fair **but** brutal
- Initialization: 1336 adjectives ($\approx$ 50/50 positive/negative)

Hatzivassiloglou McKeown 1997
Predicting the Semantic Orientation of Adjectives

Expansion using external resources:



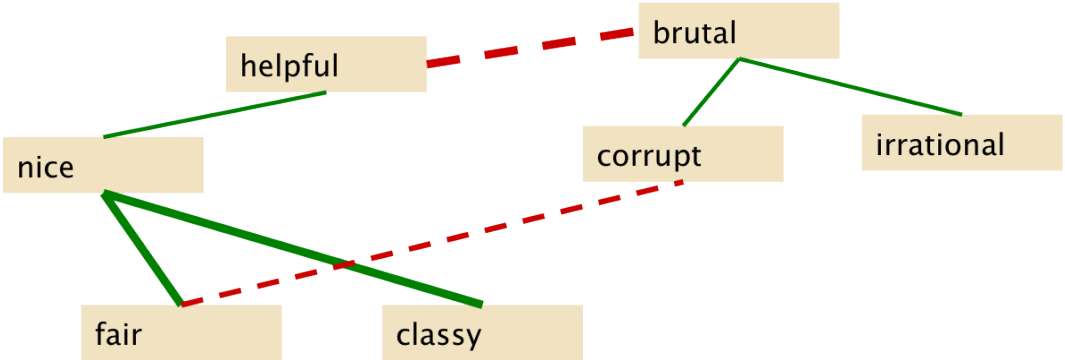
nice, helpful

nice, classy

51

Hatzivassiloglou McKeown 1997
Predicting the Semantic Orientation of Adjectives

Hatzivassiloglou and McKeown 1997



+ clustering

Hatzivassiloglou McKeown 1997
Predicting the Semantic Orientation of Adjectives

Results :

- **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

- Initialization from an annotated corpus
  (user reviews)

  ☆☆☆☆☆ **The iPhone 4S: a smartphone and a whole lot more**, September 30, 2012
  By **SophieK** (Palo Alto, CA) - See all my reviews
  This review is from: Apple iPhone 4S 16GB (White) - AT&T (Electronics)
  I finally made the transition to the Apple iPhone 4S after over two years of a few highs and countless lows with an old Motorola Droid (model A855), which now serves as a paper weight. I'll make this short and sweet.

  What I love:
  1. The awesome camera, especially when paired with the Camera+ app, allows me to keep my bulky DSLR at home when I need a good, serviceable scenery shot for social

  - Part of Speech analysis
  - Adjectives annotated from document label
- frequential filtering

Usage: one step of their summarization system:

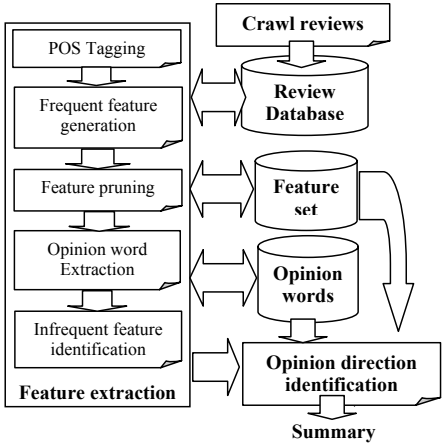


Figure 1: The opinion summarization system

Pointwise Mutual Information ,Turney, 2002

1. Documents ⇒ small patterns (=phrases)

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

2. Phrases evaluation
   - Positive phrases co-occur more with *excellent*
   - Negative phrases co-occur more with *poor*
3. Score aggregation at the document level

Turney, ACL 2002

Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised

Pointwise Mutual Information ,Turney, 2002

1. Documents ⇒ small patterns (=phrases)

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

2. Phrases evaluation
   - Positive phrases co-occur more with *excellent*
   - Negative phrases co-occur more with *poor*
3. Score aggregation at the document level

But how to measure co-occurrence?

Turney, ACL 2002

Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised

Mutual Information:

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

kind of similarity between X et Y.

Pointwise Mutual Information:

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

How much more do events x and y co-occur than if they were independent? (i.e. $PMI = 0$ in case of independence)

Probabilities estimation with Altavista:

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

Sentence Polarity

$$Pol(s) = PMI(s, "excellent") - PMI(s, "poor")$$

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

PMI [Turney, 2002] : Results

Positive Reviews:

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

Negative Reviews:

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

$\Rightarrow$ External resources: finding some pattens that are topic-related and not universal

PMI [Turney, 2002] : Results

- 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...)

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 %

Key points:

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

Same methodology as Turney... But introducing other analysis axes :

$$\text{Evaluative factor: } EVA(m) = \frac{d(m, bad) - d(m, good)}{d(good, bad)} \quad (1)$$

$$\text{Potency factor: } POT(m) = \frac{d(m, weak) - d(m, strong)}{d(strong, weak)} \quad (2)$$

$$\text{Activity factor: } ACT(m) = \frac{d(m, passive) - d(m, active)}{d(active, passive)} \quad (3)$$

Quantitative results: 61% $\rightarrow$ 71%
Qualitative analysis: comparison with the General Inquirer

J. Kamps, MJ Marx, R.J Mokken et M. De Rijke, LREC 2004
Using wordnet to measure semantic orientations of adjectives

LSA: Latent Semantic Analysis

- Modeling: Word count (and BoW storage)

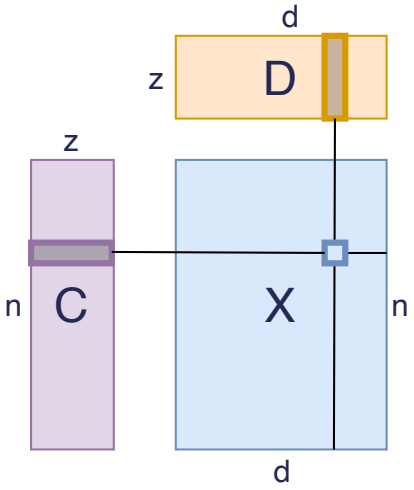
$$X = \begin{matrix} & & \mathbf{t}_j \\ & & \downarrow \\ \mathbf{d}_i \rightarrow & & \end{matrix} \begin{pmatrix} x_{1,1} & \cdots & x_{1,D} \\ \vdots & \ddots & \vdots \\ x_{N,1} & \cdots & x_{N,D} \end{pmatrix}$$

- Basic proposal: semantics = metrics = similarity between columns in BoW

$$s(j,k) = \langle t_j, t_k \rangle, \qquad \text{Normalized: } s_n(j,k) = \cos(\theta) = \frac{\mathbf{t}_j \cdot \mathbf{t}_q}{\|\mathbf{t}_j\| \, \|\mathbf{t}_q\|}$$

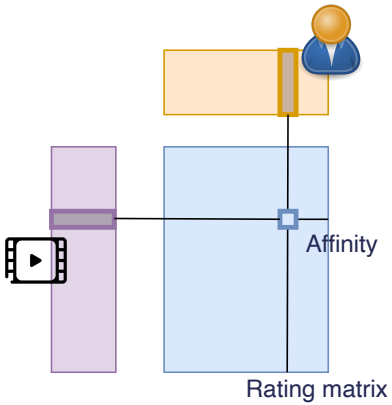
  - If two terms appear in the same document, they are similar

Matrix factorization = basic tool to understand the data



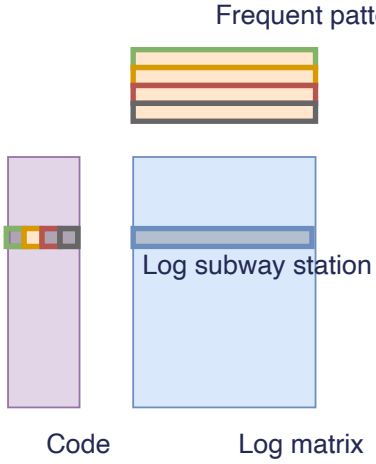
- Extract a compact representation
  - for words
  - for documents
- = focus on high-energy phenomenon
  - Eliminate noise in the data
- Optimal data compression [Mean Square criterion]

Matrix factorization = basic tool to understand the data



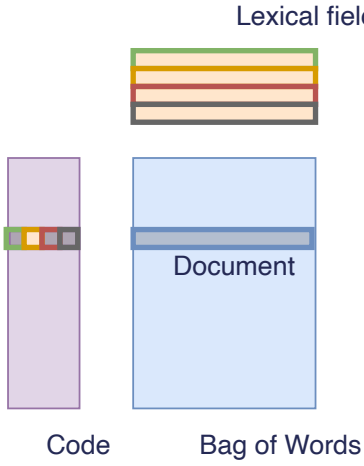
- Extract a compact representation
  - for words
  - for documents
- = focus on high-energy phenomenon
  - Eliminate noise in the data
- Optimal data compression [Mean Square criterion]

Matrix factorization = basic tool to understand the data



- Extract a compact representation
  - for words
  - for documents
- = focus on high-energy phenomenon
  - Eliminate noise in the data
- Optimal data compression [Mean Square criterion]

Matrix factorization = basic tool to understand the data



- Extract a compact representation
  - for words
  - for documents
- = focus on high-energy phenomenon
  - Eliminate noise in the data
- Optimal data compression [Mean Square criterion]

SVD : Singular Value decomposition

- In NLP : SVD = LSA: Latent Semantic Analysis
- Idea : grouping similar documents / learning a representation of documents

$$
\begin{array}{cccccc}
 & X^T & = & U & \Sigma & V^T \\
 & \mathbf{d}_i & & & & \hat{\mathbf{d}}_i \\
 & \downarrow & & & & \downarrow \\
\mathbf{t}_j \rightarrow & \begin{pmatrix} x_{1,1} & \dots & x_{1,N} \\ \vdots & \ddots & \vdots \\ x_{D,1} & \dots & x_{D,N} \end{pmatrix} & = & \left(\begin{pmatrix} \\ \mathbf{u}_1 \\ \\ \end{pmatrix} \dots \begin{pmatrix} \\ \mathbf{u}_l \\ \\ \end{pmatrix} \right) & \begin{pmatrix} \sigma_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_l \end{pmatrix} & \begin{pmatrix} (& \mathbf{v}_1 &) \\ & \vdots & \\ (& \mathbf{v}_l &) \end{pmatrix}
\end{array}
$$

- Good news: functions well on sparse matrices

Factorization = robustness & clustering ability

S. Deerwester, et al., JSIS 1990
Indexing by latent semantic analysis

Selecting the k greatest singular values gives a rank-k approximation of the occurence matrix.

- Each $\mathbf{u} \in \mathbb{R}^D$ is a weight vector associated to the vocabulary
- The base $\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$ is orthogonal
  - Each $\mathbf{u}$ corresponds to a different **lexical field**
- The new document representation $\mathbf{v}$ is a weight vector associated to the lexical fields
  - Clustering issue: the strongest weight gives the document class

Thomas K. Landauer, Peter W. Foltz et Darrell Laham, Discourse Processes, vol. 25, 1998
Introduction to Latent Semantic Analysis

Many applications

Usages:

- Clustering (each eigen vector describes a *topic)*
- Semantics: words have a representation over the topics
- IR Improvement:
  - Query expansion based on the topic definition
  - Detection of polysemic terms
- new representation $\Rightarrow$ new metrics
  - opportunities in question answering
    - Finding the part of a document relating to a specific topic
  - Automated summarization
    - Document segmentation + sentence extraction
  - TDT : Topic detection & Tracking

LSA Limits

- Fully based on BOW: no word dependency modeling
  - issues regarding negative formulation
  - depends on document sizes
  - Not robust to stop words
    - associated to high singular values
    - + appear in many topics
- Topic modeling is link to a corpus
  - problem with rare words in small corpus
  - bias of the corpus

- Still a BOW modeling

$$X = \begin{matrix} & & \mathbf{t}_j \\ & & \downarrow \\ \mathbf{d}_i \rightarrow & \end{matrix} \begin{pmatrix} x_{1,1} & \dots & x_{1,D} \\ \vdots & \ddots & \vdots \\ x_{N,1} & \dots & x_{N,D} \end{pmatrix}$$

- Algorithm that scale up well
  - Possible **on-line** version of the algorithm
  - Can be linked to chinese restaurant / indian buffet process
    - $\Rightarrow$ Discover k in an online process
- Orthogonality is not longer enforced

New vision of k-means :

- k clusters
- A priori probabilities : $\pi_k = p(\theta_k)$
- Probability of a word in a cluster : $p(w_j|\theta_k) = \mathbb{E}_{d \in \mathcal{D}_k}[w_j]$
- Document hard assignment in a cluster: $p(\theta_k|d_i) = 1/0$

$$y_i = \arg\max_k p(\theta_k)p(d_i|\theta_k) = \arg\max_k \log(\pi_k) + \sum_{w_j \in d_i} \log p(w_j|\theta_k)$$

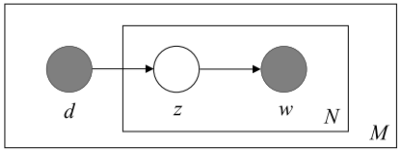
$$y_i = \arg\max_k \sum_j t_{ij}\theta_{jk}, \text{ with } :\theta_{jk} = \log p(w_j|\theta_k) \text{ and uniform prior}$$

Algorithm:

Init. Random or expert knowledge
C/E Cluster assignment
M Parameter update (mean re-computation)

Probabilistic Latent Semantic Analysis

- Idea: CEM $\Rightarrow$ EM (more complex / finer)
- All documents belongs to all clusters... With a weight $p(z|d)$
- Graphical model



- Doc d is drawn from $P(d)$
- Topic z is drawn from $P(z|d)$
- Word w is drawn from $P(w|z)$

We estimate the following parameters:

- $p(d)$
- $p(\alpha|d)$
- $p(w|\alpha)$

Maximizing the log-likelyhood:

$$\mathcal{L} = \sum_{d=1}^{D} \sum_{w=1}^{W} n(d, w) \log P(d, w)$$

- **E**xpectation (probability of the missing variables)
- **M**aximization

Maximizing the log-likelyhood:

$$\mathcal{L} = \sum_{d=1}^{D} \sum_{w=1}^{W} n(d, w) \log P(d, w)$$

- **E**xpectation (probability of the missing variables)

$$P(\alpha|d, w) = \frac{P(d)P(\alpha|d)P(w|\alpha)}{\sum_{\alpha' \in \mathcal{A}} P(d)P(\alpha'|d)P(w|\alpha')}$$

- **M**aximization

Maximizing the log-likelyhood:

$$\mathcal{L} = \sum_{d=1}^{D} \sum_{w=1}^{W} n(d, w) \log P(d, w)$$

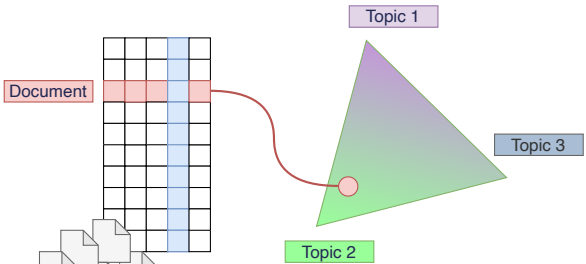
- Expectation (probability of the missing variables)
- Maximization

$$P(d) = \frac{\sum_{w \in \mathcal{W}} n(d, w)}{\sum_{d' \in \mathcal{D}} \sum_{w \in \mathcal{W}} n(d', w)}$$

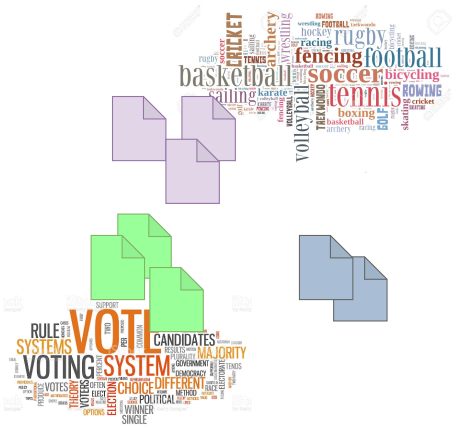
$$P(\alpha|d) = \frac{\sum_{w \in \mathcal{W}} n(d, w) P(\alpha|d, w)}{\sum_{\alpha' \in \mathcal{A}} \sum_{w \in \mathcal{W}} n(d, w) P(\alpha'|d, w)}$$

$$P(w|\alpha) = \frac{\sum_{d \in \mathcal{D}} n(d, w) P(\alpha|d, w)}{\sum_{w' \in \mathcal{W}} \sum_{d \in \mathcal{D}} n(d, w') P(\alpha|d, w')}$$

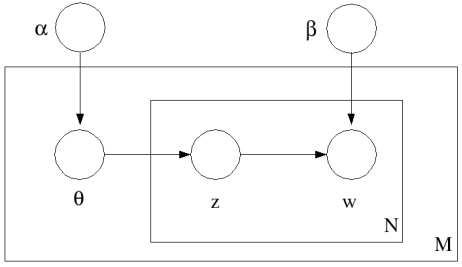
PLSA: results



Analyse sémantique



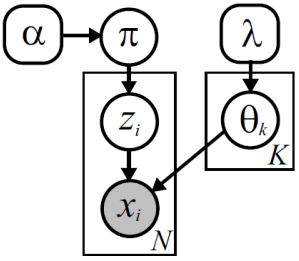
Latent Dirichlet Allocation:

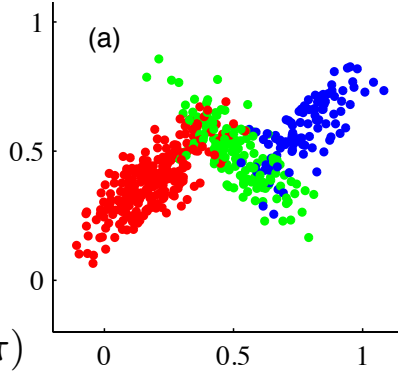


- Idea: adding a prior on the topic distribution
  - A document is supposed to belong to a topic **strongly or not**
- Learning through Gibbs sampling ($\sim$ MCMC)

not to be confused: LDA: Latent Dirichlet Allocation *vs* Linear Discriminant

On an example:





$$\theta_k = \{\mu_k, \Sigma_k\}$$

$$p(z_i \mid \pi) = \mathrm{Cat}(z_i \mid \pi)$$

$$p(x_i \mid z_i, \mu, \Sigma) = \mathcal{N}(x_i \mid \mu_{z_i}, \Sigma_{z_i})$$

Given mixture weights $\pi^{(t-1)}$ and cluster parameters $\{\theta_k^{(t-1)}\}_{k=1}^K$ from the previous iteration, sample a new set of mixture parameters as follows:

1. Independently assign each of the N data points x_i to one of the K clusters by sampling the indicator variables $z = \{z_i\}_{i=1}^N$ from the following multinomial distributions:

$$z_i^{(t)} \sim \frac{1}{Z_i} \sum_{k=1}^{K} \pi_k^{(t-1)} f(x_i \mid \theta_k^{(t-1)}) \, \delta(z_i, k) \qquad Z_i = \sum_{k=1}^{K} \pi_k^{(t-1)} f(x_i \mid \theta_k^{(t-1)})$$

2. Sample new mixture weights according to the following Dirichlet distribution:

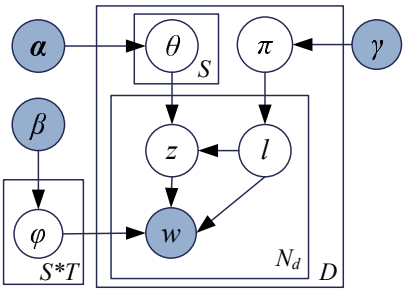
$$\pi^{(t)} \sim \mathrm{Dir}(N_1 + \alpha/K, \ldots, N_K + \alpha/K) \qquad N_k = \sum_{i=1}^{N} \delta(z_i^{(t)}, k)$$

3. For each of the K clusters, independently sample new parameters from the conditional distribution implied by those observations currently assigned to that cluster:

$$\theta_k^{(t)} \sim p(\theta_k \mid \{x_i \mid z_i^{(t)} = k\}, \lambda)$$

When λ defines a conjugate prior, this posterior distribution is given by Prop. 2.1.4.

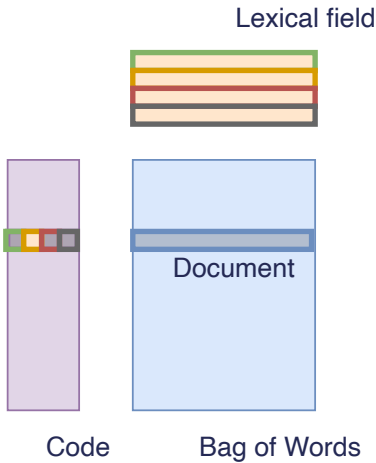
- Graphical models = easy to adapt



- For each document d, choose a distribution $\pi_d \sim \text{Dir}(\gamma)$.
- For each sentiment label l under document d, choose a distribution $\theta_{d,l} \sim \text{Dir}(\alpha)$.
- For each word w_i in document d
  - choose a sentiment label $l_i \sim \text{Mult}(\pi_d)$,
  - choose a topic $z_i \sim \text{Mult}(\theta_{d,l_i})$,
  - choose a word w_i from $\varphi_{z_i}^{l_i}$, a Multinomial distribution over words conditioned on topic z_i and sentiment label l_i.

Conclusion on statistical semantic analysis

1. Quantitative results
   - Clustering
   - Major issue with frequent words
   - Human required in the loop (init., cluster selection, etc...)
   - Evaluation issue (purity, perplexity, ...)
2. Qualitative analysis
   - Word similarity
   - Lexical field extraction



Word2Vec algorithm

The distributional hypothesis [Harris et al. 1954]

Word that appear in similar contexts in text tend to have similar meanings.

he curtains open and the moon shining in on the barely
ars and the cold , close moon " . And neither of the w
rough the night with the moon shining so brightly , it
made in the light of the moon . It all boils down , wr
surely under a crescent moon , thrilled by ice-white
sun , the seasons of the moon ? Home , alone , Jay pla
m is dazzling snow , the moon has risen full and cold
un and the temple of the moon , driving out of the hug
in the dark and now the moon rises , full and amber a
bird on the shape of the moon over the trees in front
But I could n't see the moon or the stars , only the
rning , with a sliver of moon hanging among the stars
they love the sun , the moon and the stars . None of
the light of an enormous moon . The plash of flowing w
man 's first step on the moon ; various exhibits , aer
the inevitable piece of moon rock . Housing The Airsh
oud obscured part of the moon . The Allied guns behind

Textual data = different analysis level

- At the **document** scale:
  - $\Rightarrow$ Several efficient algorithm to classify/categorize
- At the **sentence** scale:
  - Segmentation
  - POS, NER, SRL...
  - $\Rightarrow$ Sequential approaches (HMM, CRF)
- At the **word** scale:
  - Defining / understanding the word
  - Linking the word with antonym, synonym, etc, ...
  - Binding the word to a lexical field
  - $\Rightarrow$ Latent semantics / linguistic resources

$\Rightarrow$ How deep learning / representation learning can improve our previous solutions?

NLP & representation learning

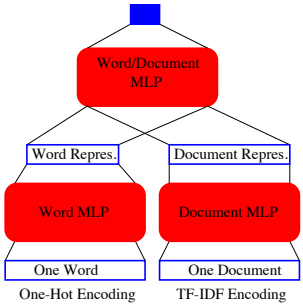
- 1990 Matrix factorization :
  - PCA is a historical way to learn representations
    - Criterion = reconstruction
  - In NLP $\Rightarrow$ SVD / PCA [Deerwester, 1990]
- 2005 First neural architecture for text representation
  - an alternative to PLSA [Keller, 2005]
- 2008 Convolutional Neural architecture for text
  - precursor of modern architectures
  - multi-tasks [Collobert, 2008]
- 2012 The Word2Vec wave
  - Qualitative & cheap word embeddings [Mikolov, 2012]
- 2013 Manifest for representation learning [Bengio, 2013]
- 2015 Seq2seq paradigm [Luong, 2015]

$\Rightarrow$ Also an approximate schedule of the presentation

Task =
Is the word w present is document d?

Close to LSA / PLSA paradigm: compressing the original data matrix through embeddings

$\Rightarrow$ Learning a **language model**



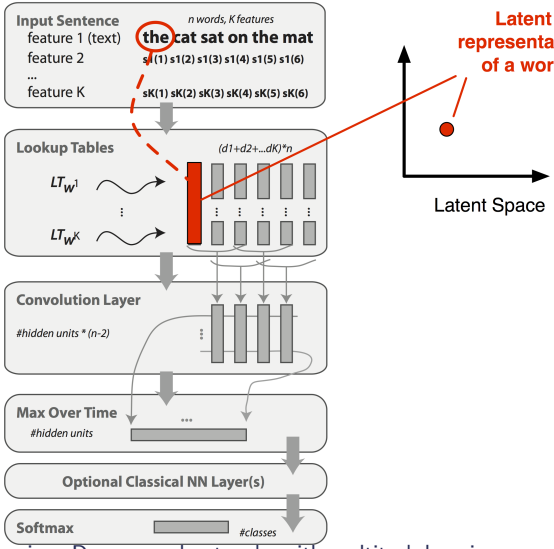
Conclusion:

Words & documents representations are more efficient than PLSA ones for several TREC information retrieval tasks

M. Keller, S. Bengio, ICANN 2005
A neural network for text representation

- **Lookup table** concept = table of embeddings
- State of the art on : POS, NER, SRL
- Quite difficult to set...
- ... But open source + **open embeddings**
- Based on torch...

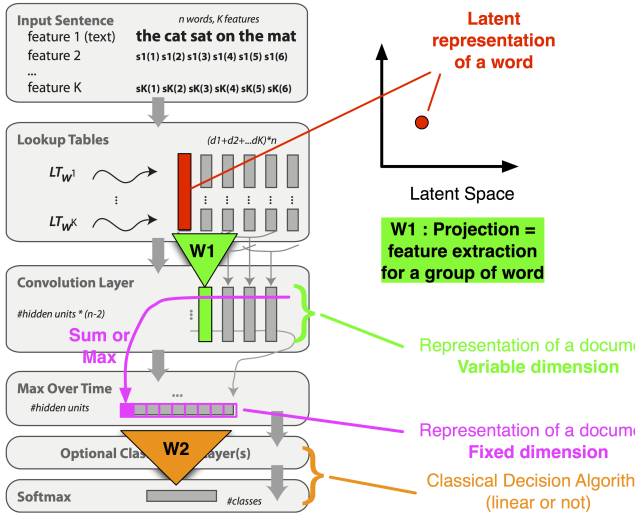
By Collobert



R. Collobert, J. Weston ICML 2008
A unified architecture for natural language processing: Deep neural networks with multitask learning

- **Lookup table** concept = table of embeddings
- State of the art on : POS, NER, SRL
- Quite difficult to set...
- ... But open source + **open embeddings**
- Based on torch...

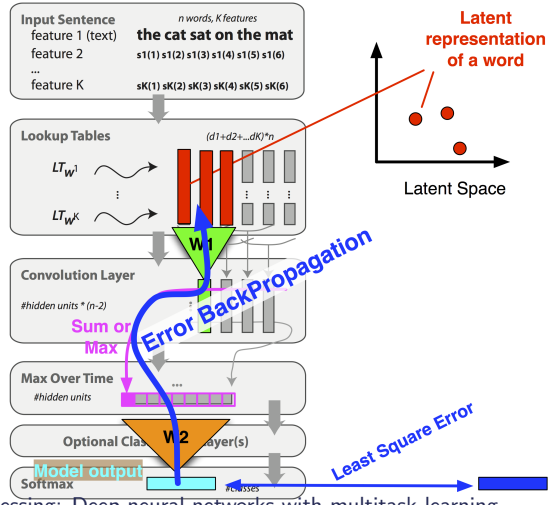
By Collobert



R. Collobert, J. Weston ICML 2008
A unified architecture for natural language processing: Deep neural networks with multitask learning

- **Lookup table** concept = table of embeddings
- State of the art on : POS, NER, SRL
- Quite difficult to set...
- ... But open source + **open embeddings**
- Based on torch...

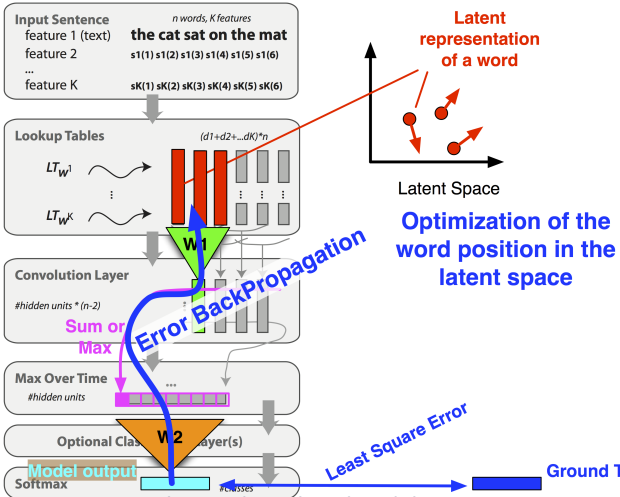
By Collobert



R. Collobert, J. Weston ICML 2008
A unified architecture for natural language processing: Deep neural networks with multitask learning

Convolutional Neural Architecture

- **Lookup table** concept
  = table of embeddings
- State of the art on :
  POS, NER, SRL
- Quite difficult to set...
- ... But open source +
  open embeddings
- Based on torch...
  By Collobert



R. Collobert, J. Weston ICML 2008
A unified architecture for natural language processing: Deep neural networks with multitask learning

Several important informations

- Embedding are learned **keeping the sentence structure**
- Embeddings benefit from multi-tasks
- Learning is slow...

Our embeddings have been trained for about 2 months, over Wikipedia.

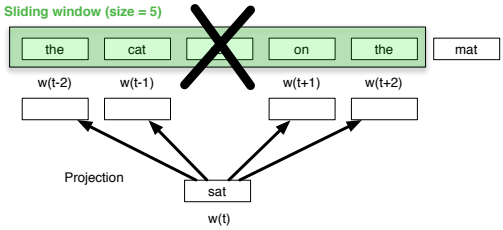
`https://ronan.collobert.com/senna/`

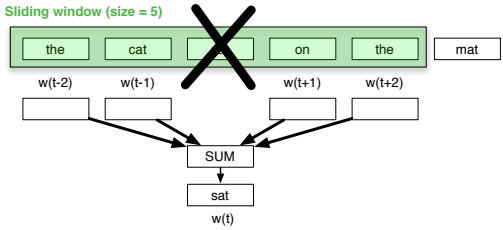
- ... But inference is fast & efficient

Task	Benchmark		Performance	Timing (s)
Part of Speech (POS)	(Toutanova et al, 2003)	(Accuracy)	97.29%	3
Chunking (CHK)	CoNLL 2000	(F1)	94.32%	2
Name Entity Recognition (NER)	CoNLL 2003	(F1)	89.59%	2
Semantic Role Labeling (SRL)	CoNLL 2005	(F1)	75.49%	36
Syntactic Parsing (PSG)	Penn Treebank	(F1)	87.92%	74

- + Costly embeddings have been made available to the community

Word2Vec : extracting local semantics





Skip-Gram (& *negative sampling* implementation) : easy learning

- Local analysis
- *predictive* criterion : estimating missing value
- Sliding window = local context C of word w

Mikolov, Sutskever, Chen, Corrado, Dean, NIPS 2013 (arXiv 2012)
Distributed representations of words and phrases and their compositionality

Word2Vec : extracting local semantics

```
he curtains open and the moon shining in on the barely
ars and the cold , close moon " . And neither of the w
rough the night with the moon shining so brightly , it
made in the light of the moon . It all boils down , wr
 surely under a crescent moon , thrilled by ice-white
sun , the seasons of the moon ? Home , alone , Jay pla
m is dazzling snow , the moon has risen full and cold
un and the temple of the moon , driving out of the hug
 in the dark and now the moon rises , full and amber a
bird on the shape of the moon over the trees in front
 But I could n't see the moon or the stars , only the
rning , with a sliver of moon hanging among the stars
 they love the sun , the moon and the stars . None of
the light of an enormous moon . The plash of flowing w
man 's first step on the moon ; various exhibits , aer
 the inevitable piece of moon rock . Housing The Airsh
oud obscured part of the moon . The Allied guns behind
```

(1) Initialisation aléatoire des positions des mots

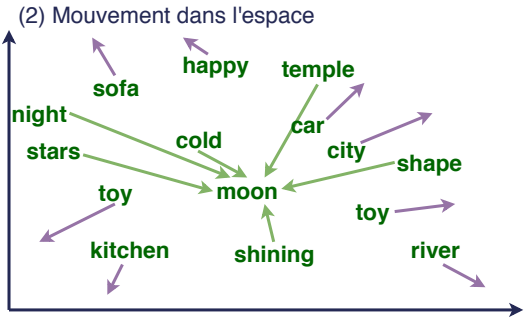
Skip-Gram (& *negative sampling* implementation) : easy learning

- Local analysis
- *predictive* criterion : estimating missing value
- Sliding window = local context C of word w

Mikolov, Sutskever, Chen, Corrado, Dean, NIPS 2013 (arXiv 2012)
Distributed representations of words and phrases and their compositionality

Word2Vec : extracting local semantics

he curtains open and the moon shining in on the barely
ars and the cold , close moon " . And neither of the w
rough the night with the moon shining so brightly , it
made in the light of the moon . It all boils down , wr
surely under a crescent moon , thrilled by ice-white
sun , the seasons of the moon ? Home , alone , Jay pla
m is dazzling snow , the moon has risen full and cold
un and the temple of the moon , driving out of the hug
in the dark and now the moon rises , full and amber a
bird on the shape of the moon over the trees in front
But I could n't see the moon or the stars , only the
rning , with a sliver of moon hanging among the stars
they love the sun , the moon and the stars . None of
the light of an enormous moon . The plash of flowing w
man 's first step on the moon ; various exhibits , aer
the inevitable piece of moon rock . Housing The Airsh
oud obscured part of the moon . The Allied guns behind



Skip-Gram (& *negative sampling* implementation) : easy learning

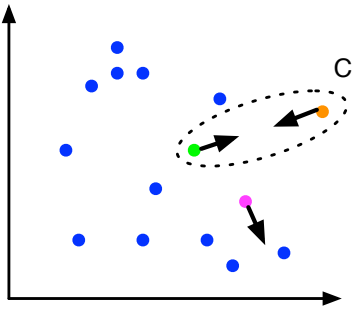
- Local analysis
- *predictive* criterion : estimating missing value
- Sliding window = local context C of word w

Mikolov, Sutskever, Chen, Corrado, Dean, NIPS 2013 (arXiv 2012)
Distributed representations of words and phrases and their compositionality

- Given word w and local contexts C:

$$\text{Idée SG: } \arg\max_{\theta} \prod_{C} \prod_{w \in C} p(C|w;\theta)$$

- $p(D = 1|w_i, w_j; \theta) \Rightarrow$ proba. that w_i and w_j occur in the same context



$$\arg\max_{\theta} \prod_{i,j \in C} p(D = 1|w_i, w_j; \theta) + \underbrace{\prod_{i,j \in \bar{C}} p(D = 0|w_i, w_j; \theta)}_{\text{Negative Sampling}}$$

Goldberg, Levy, arXiv 2014
word2vec Explained: Deriving Mikolov et al.'s Negative-Sampling Word-Embedding Method

Hammer, NN 2002
Generalized Relevance Learning Vector Quantization

$$\arg\max_{\theta} \prod_{i,j \in C} p(D=1|w_i, w_j; \theta) + \underbrace{\prod_{i,j \in \bar{C}} p(D=0|w_i, w_j; \theta)}_{\text{Negative Sampling}}$$

- Using logistic function : $p(D=1|w_i, w_j) = \frac{1}{1+\exp(-\mathbf{z}_i \mathbf{z}_j)}$
- Global log-likelihood : $\arg\max_{Z} \left(\sum_{i,j \in C} \log \sigma(\mathbf{z}_i \cdot \mathbf{z}_j) + \sum_{i,j \in \bar{C}} \log \sigma(-\mathbf{z}_i \cdot \mathbf{z}_j) \right)$

σ : fct sigmoide, C : Set of Cooccurences, $\bar{C}$: Set of Non-Cooc

- Stochastic Gradient Descent + triplet loss
- Frequent word subsampling trick : picking words with $p(w_i) = 1 - \sqrt{\frac{t}{freq(w_i)}}, \qquad t = 10^{-5}$

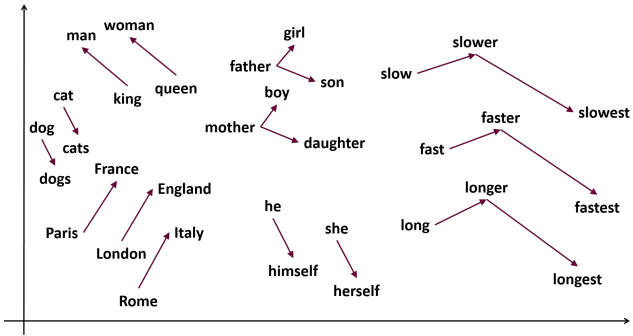
W2V: Parametrization & results

- Choosing a large latent space: $z \in [500,\ 1000]$
  (vs PLSA/LDA $z \in [10,\ 100]$)
- Operate on large & small corpora...
- ... very fast !

Model (training time)	Redmond	Havel	ninjutsu	graffiti	capitulate
Collobert (50d) (2 months)	conyers lubbock keene	plauen dzerzhinsky osterreich	reiki kohona karate	cheesecake gossip dioramas	abdicate accede rearm
Turian (200d) (few weeks)	McCarthy Alston Cousins	Jewell Arzu Ovitz	- - -	gunfire emotion impunity	- - -
Mnih (100d) (7 days)	Podhurst Harlang Agarwal	Pontiff Pinochet Rodionov	- - -	anaesthetics monkeys Jews	Mavericks planning hesitated
Skip-Phrase (1000d, 1 day)	Redmond Wash. Redmond Washington Microsoft	Vaclav Havel president Vaclav Havel Velvet Revolution	ninja martial arts swordsmanship	spray paint grafitti taggers	capitulation capitulated capitulating

a is to *b* what *c* is to ??? $\Leftrightarrow$ $\mathbf{z}_b - \mathbf{z}_a + \mathbf{z}_c$

Syntactical property (1):



$\mathbf{z}_{woman} - \mathbf{z}_{man} \approx \mathbf{z}_{queen} - \mathbf{z}_{king}$

$\mathbf{z}_{kings} - \mathbf{z}_{king} \approx \mathbf{z}_{queens} - \mathbf{z}_{kings}$

Query:

$\mathbf{z}_{woman} - \mathbf{z}_{man} + \mathbf{z}_{king} = \mathbf{z}_{req}$

Nearest neighbor:

$\arg\min_i \|\mathbf{z}_{req} - \mathbf{z}_i\| = \text{queen}$

a is to b what c is to ??? $\Leftrightarrow$ $\mathbf{z}_b - \mathbf{z}_a + \mathbf{z}_c$

Syntactical property (2):

Query:

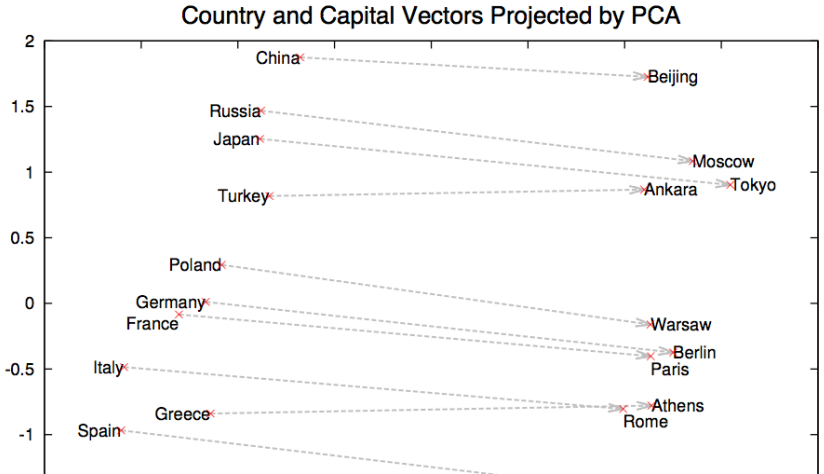
$\mathbf{z}_{easy} - \mathbf{z}_{easiest} + \mathbf{z}_{luckiest} = \mathbf{z}_{req}$

Nearest neighbor:

$\arg\min_i \|\mathbf{z}_{req} - \mathbf{z}_i\| = \text{lucky}$

a is to *b* what *c* is to ??? $\Leftrightarrow$ $\mathbf{z}_b - \mathbf{z}_a + \mathbf{z}_c$

Semantic Property (1):



a is to *b* what *c* is to ??? $\Leftrightarrow$ $\mathbf{z}_b - \mathbf{z}_a + \mathbf{z}_c$

Semantic Property (2)

Czech + currency	Vietnam + capital	German + airlines	Russian + river	French + actress
koruna	Hanoi	airline Lufthansa	Moscow	Juliette Binoche
Check crown	Ho Chi Minh City	carrier Lufthansa	Volga River	Vanessa Paradis
Polish zolty	Viet Nam	flag carrier Lufthansa	upriver	Charlotte Gainsbourg
CTK	Vietnamese	Lufthansa	Russia	Cecile De

Table 5: Vector compositionality using element-wise addition. Four closest tokens to the sum of two vectors are shown, using the best Skip-gram model.

Why is it working?

- Predicting instead of counting... Not the good explanation!
- Learning a local semantics !

$\Rightarrow$ GloVe $\approx$ PLSA + local context + embedding based implem.

- $X \in \mathbb{R}^{V \times V}$ word co-occurrence matrix
- X_{ij} frequency of word i co-occurring with word j
- $X_i = \sum_k X_{ik}$ total number of occurrences of word i in corpus
- $P_{ij} = P(j|i) = \frac{X_{ij}}{X_i}$ a.k.a. probability of word j occurring within the context of word i
- $w \in \mathbb{R}^z$ a word embedding of dimension z
- $\tilde{w} \in \mathbb{R}^z$ a context word embedding of dimension z

Baroni, Dinu & Kruszewski, ACL 2014
Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors

W2V : Strengths & weaknesses

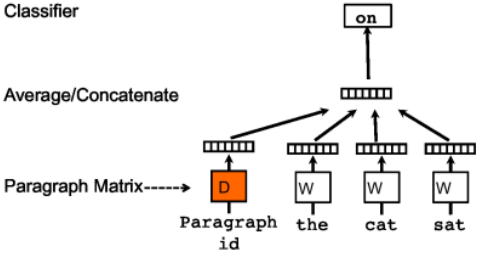
- Incredible impact for a badly written article...
- ... Associated to strong results / Lightweight algorithm & powerful implem.
- ... & available pretrained embeddings

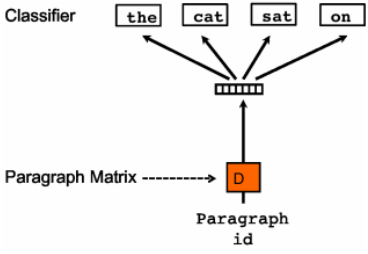
One remaining question: we learned a powerful semantics at the **word level**... How scaling to the **sentence** or **document level**?

W2V : Strengths & weaknesses

- Incredible impact for a badly written article...
- ... Associated to strong results / Lightweight algorithm & powerful implem.
- ... & available pretrained embeddings

One remaining question: we learned a powerful semantics at the **word level**... How scaling to the **sentence** or **document level**?





Q. Le, T. Mikolov, ICML 2014
Distributed representations of sentences and documents

- Incredible impact for a badly written article...
- ... Associated to strong results / Lightweight algorithm & powerful implem.
- ... & available pretrained embeddings

One remaining question: we learned a powerful semantics at the **word level**... How scaling to the **sentence** or **document level**?

Or simple averaging of word embeddings:

+ great results on small word groups
− poor results on larger groups
  - quickly converge to a central abstract point of the latent space

W2V : Strengths & weaknesses

- Incredible impact for a badly written article...
- ... Associated to strong results / Lightweight algorithm & powerful implem.
- ... & available pretrained embeddings

One remaining question: we learned a powerful semantics at the **word level**... How scaling to the **sentence** or **document level**?

1. Aggregate multiple words associated to a single entity
   - *Pointwise Mutual Information* threshold:

$$\text{score}(w_i, w_j) = \frac{\text{count}(w_i w_j) - \delta}{\text{count}(w_i) \times \text{count}(w_j)}.$$

2. Include new terms in the dictionary before running word2vec

Newspapers			
New York	New York Times	Baltimore	Baltimore Sun
San Jose	San Jose Mercury News	Cincinnati	Cincinnati Enquirer
NHL Teams			
Boston	Boston Bruins	Montreal	Montreal Canadiens
Phoenix	Phoenix Coyotes	Nashville	Nashville Predators
NBA Teams			
Detroit	Detroit Pistons	Toronto	Toronto Raptors
Oakland	Golden State Warriors	Memphis	Memphis Grizzlies
Airlines			
Austria	Austrian Airlines	Spain	Spainair
Belgium	Brussels Airlines	Greece	Aegean Airlines
Company executives			
Steve Ballmer	Microsoft	Larry Page	Google
Samuel J. Palmisano	IBM	Werner Vogels	Amazon