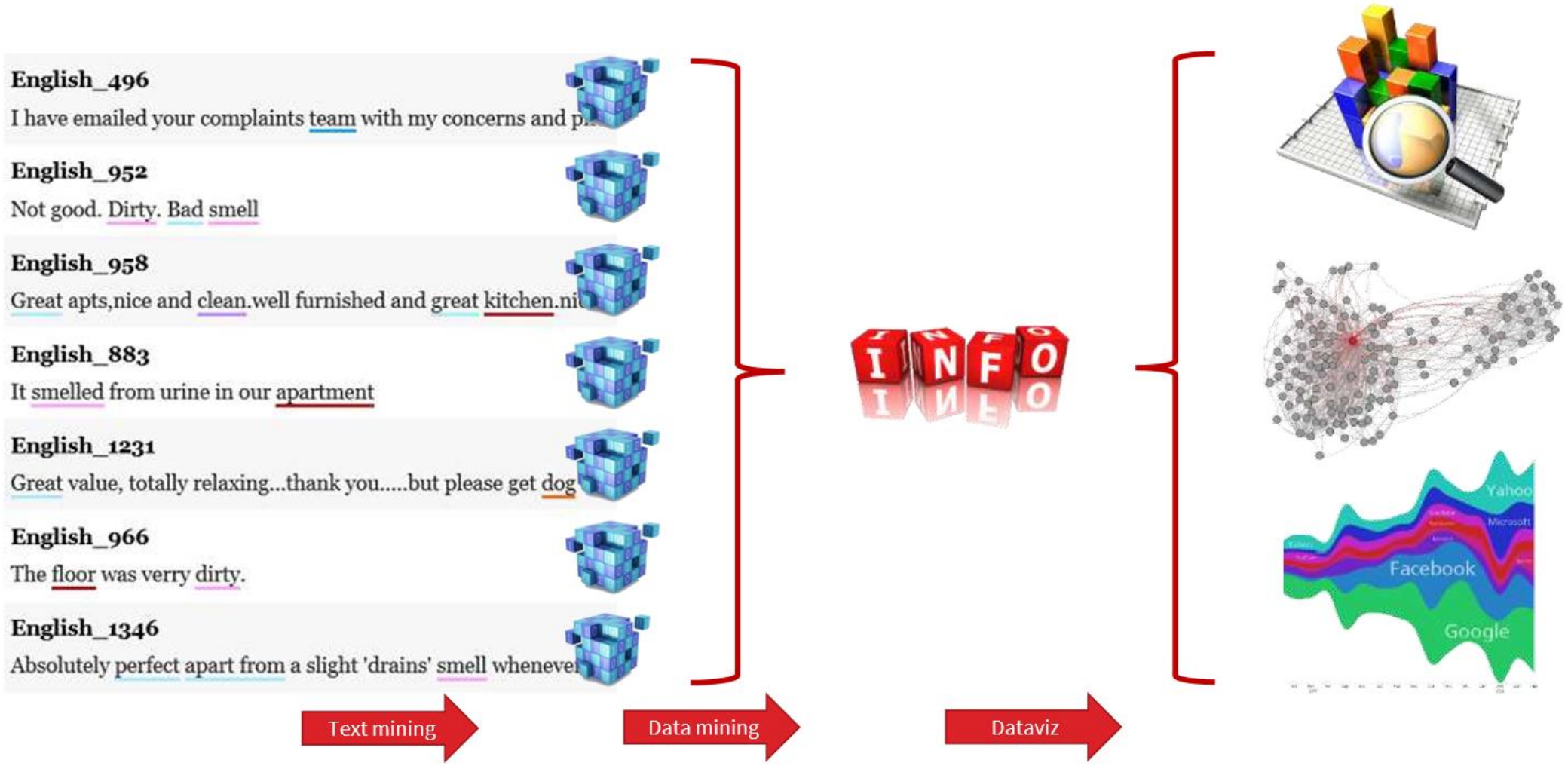
proxem

Word-sense Disambiguation (WSD)

Adaptive Skip-Gram : An unsupervised and non-parametric algorithm

About Proxem



- We need both a good recall and a good precision !

Content

- Introduction
    - Why is WSD a major concern ?
    - Application of WSD
    - Some methods for WSD
- Adaptive Skip-Gram
    - Review of Skip-Gram model
    - Learning algorithm
    - Disambiguation algorithm
- Training & Results

Introduction

There are lots of polysemous words in every language!

- English : Bank, bass, Apple, address, match …
- French : Orange, avocat, carrefour…

In english : More than 700 words have at least 10 different meanings -> Ignoring WSD can hurt precision pretty badly for some tasks.

Application of WSD

- Machine translation
  - J'ai rendez-vous avec mon avocat aujourd'hui -> Lawyer or Avocado
- Information retrieval : improve precision of queries
  - orange/total/carrefour -> Ambiguous brands in French
- Speech processing :
  - chose the right phonetization for speech synthesis -> fils / fils
  - Homophone discrimination for speech recognition -> vers / verre / ver
- POS, NER, …

Some methods for WSD

- Using Part-of-speech to disambiguate words
  - To address (Verb) and mailing address (Noun)
  - However this is not enough for most words…

- Using multi-lingual embeddings : cf Coulmance et al. , 2015
  - Apple (company) / Apple (fruit) -> Apple (company) / pomme (fruit). Hence $Apple_{english} - pomme_{french}$ will give a vector close to Tech companies

- Using thesaurus such as WordNet

Those methods needs external resources that might be expensive and/or partially inaccurate : Knowledge acquisition bottleneck

Adaptive Skip-Gram

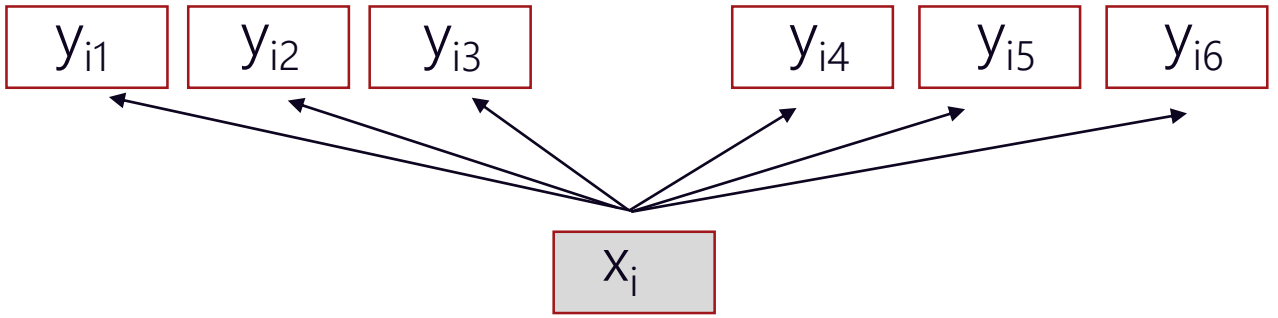
Bartunov et al. 2015

The idea behind Skip-Gram

« You shall know a word by the company it keeps » Firth (1957)

Skip-Gram : Mikolov et al. , 2013

- Input text of N words : o_1, o_2, … o_N interpreted as :
  - A sequence of input words x_1, … x_N and
  - Their context words y_1, … y_N.
- Vocabulary size V

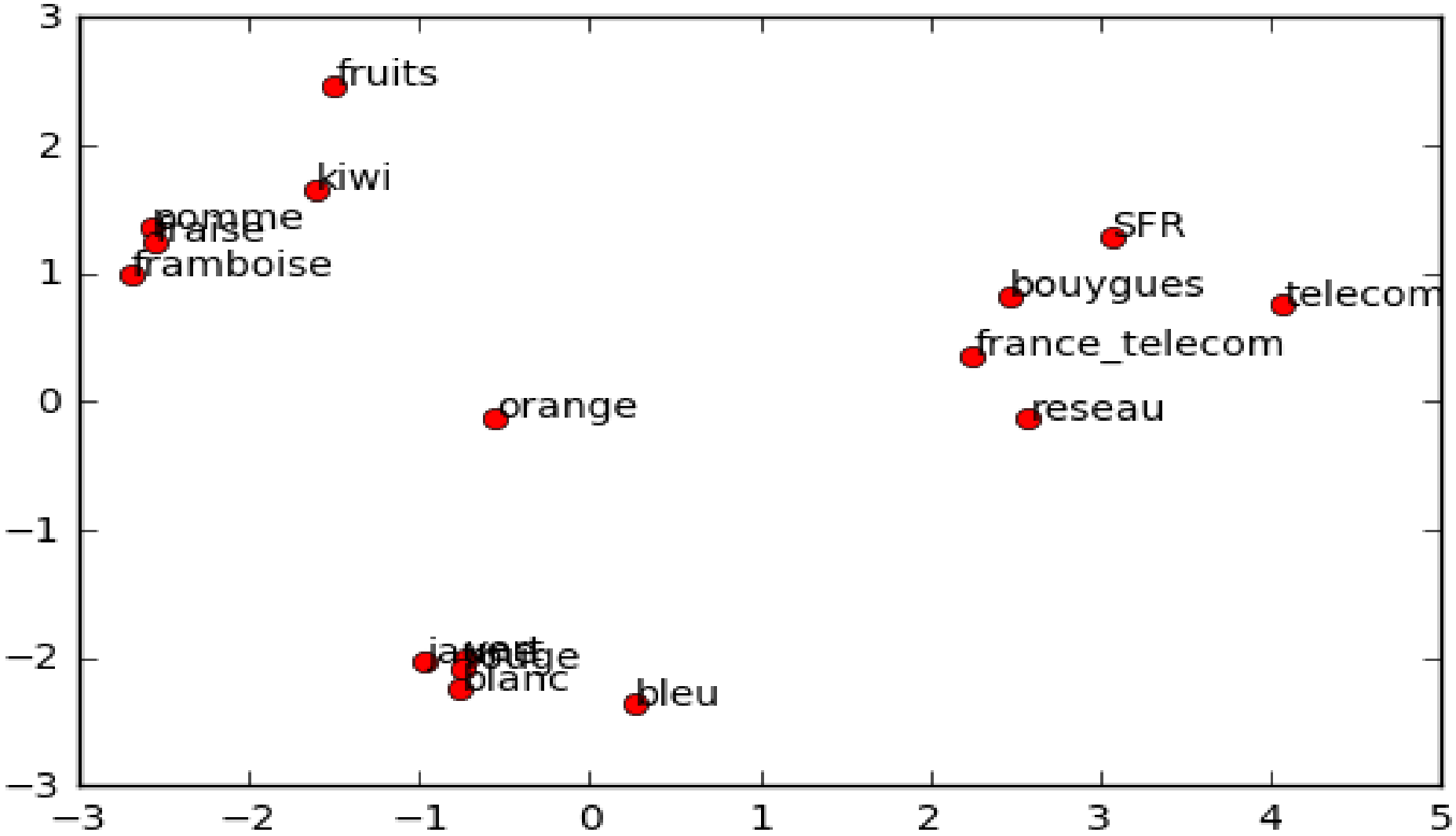


- Each word x has two vector representations : one as a center word in_x and one as a context word out_x
- Overall probability model :

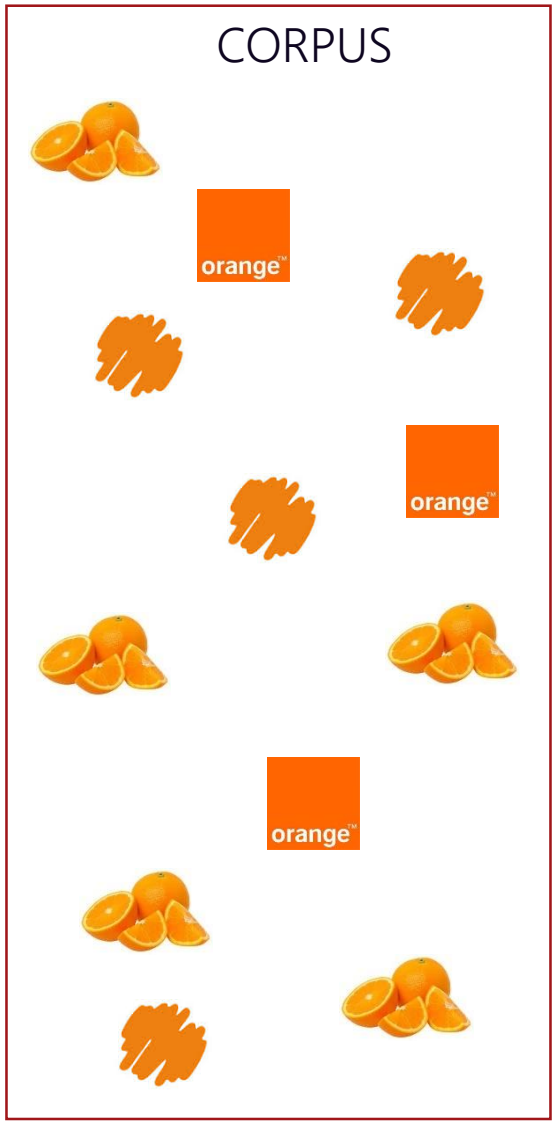
$$p(Y \mid X, \theta) = \prod_{i=1}^{N} \prod_{j=1}^{C} p(y_{ij} \mid x_i, \theta) \qquad \text{with} \qquad p(y \mid x, \theta) = \frac{\exp(in_x^T out_y)}{\sum_{v=1}^{V} \exp(in_x^T out_v)}$$

C is the context window size often chosen to be sampled from [1,5] for each new word x_i

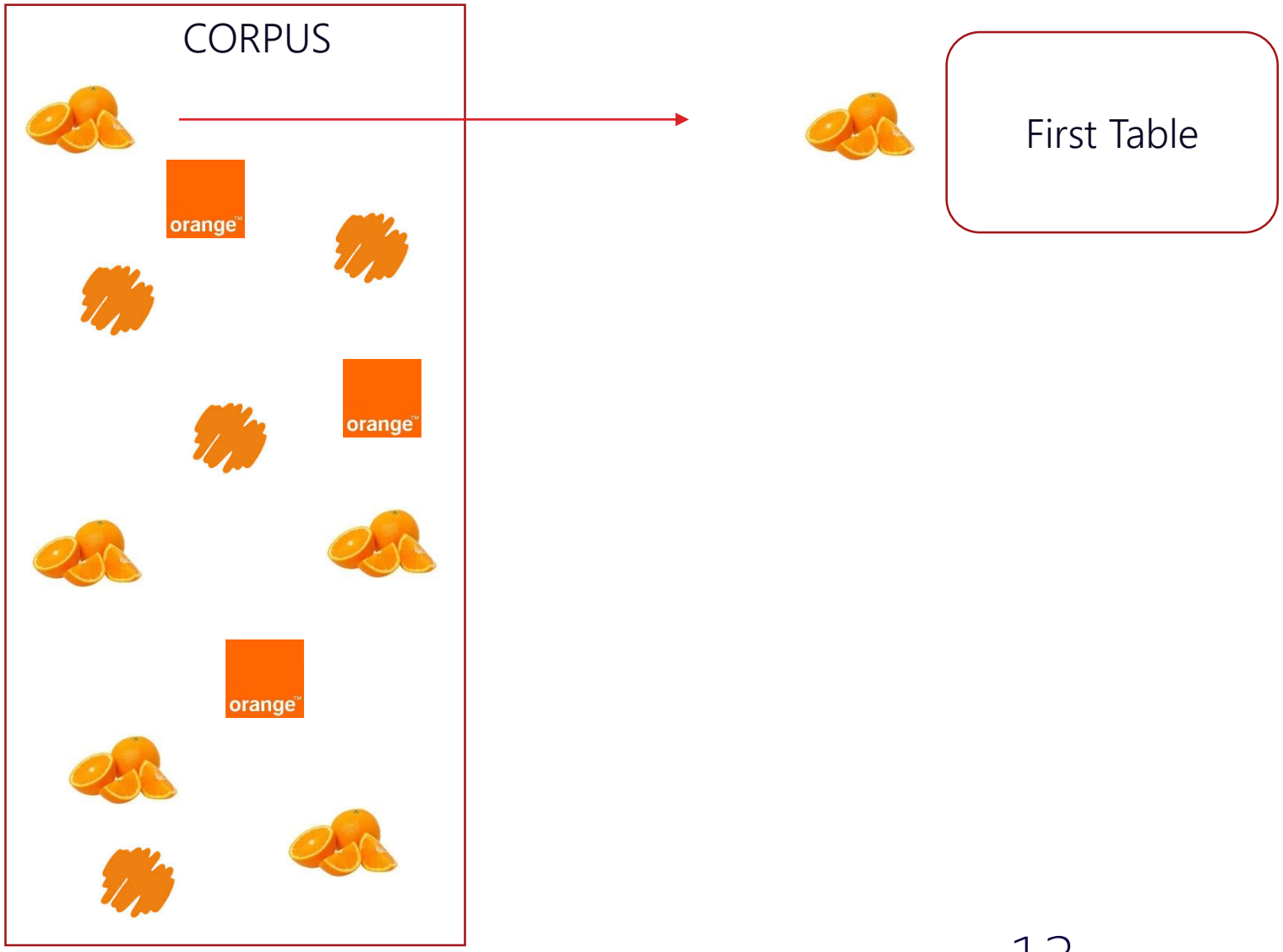
Skip-Gram : visualisation



AdaGram : Chinese Restaurant process

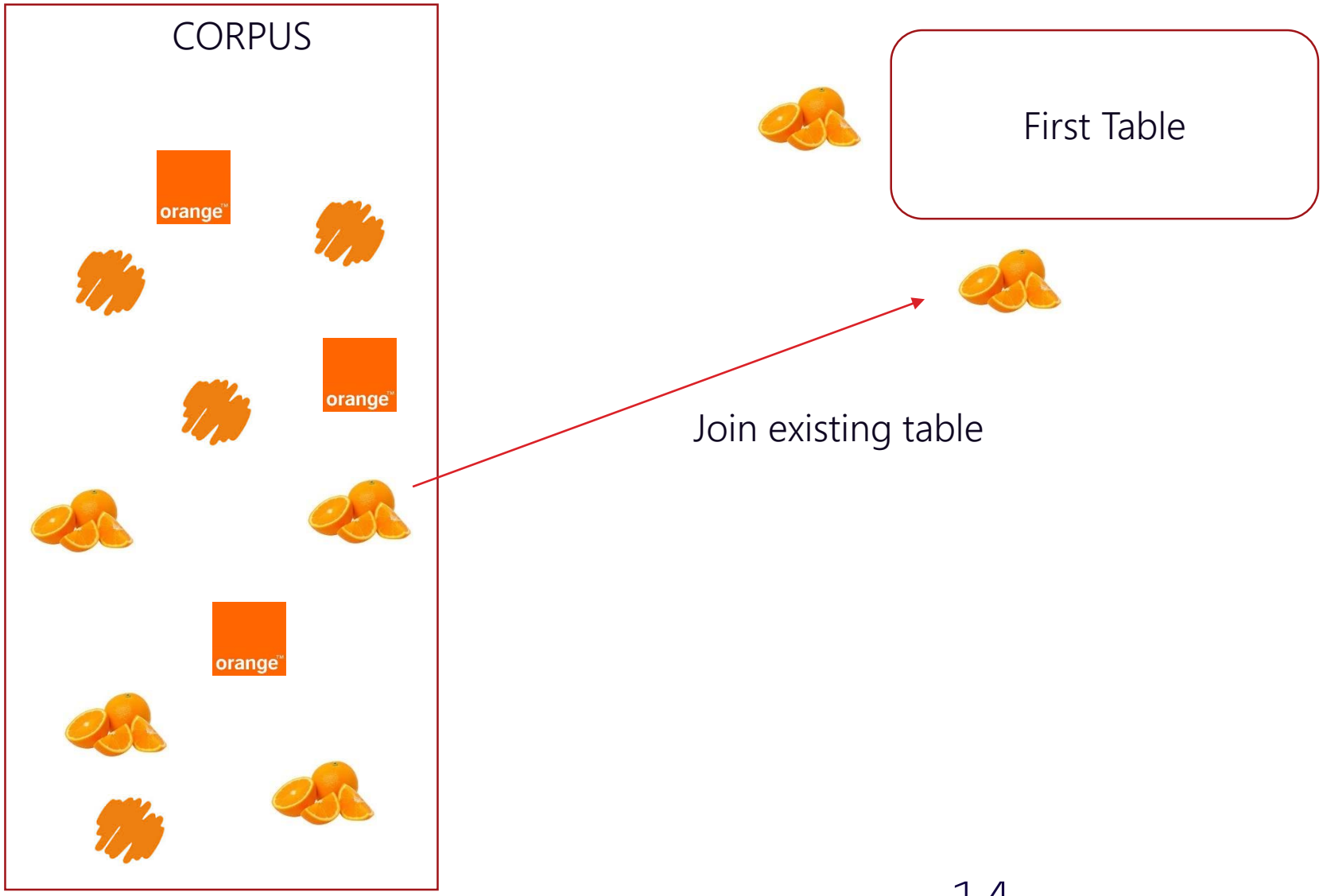


AdaGram : Chinese Restaurant process

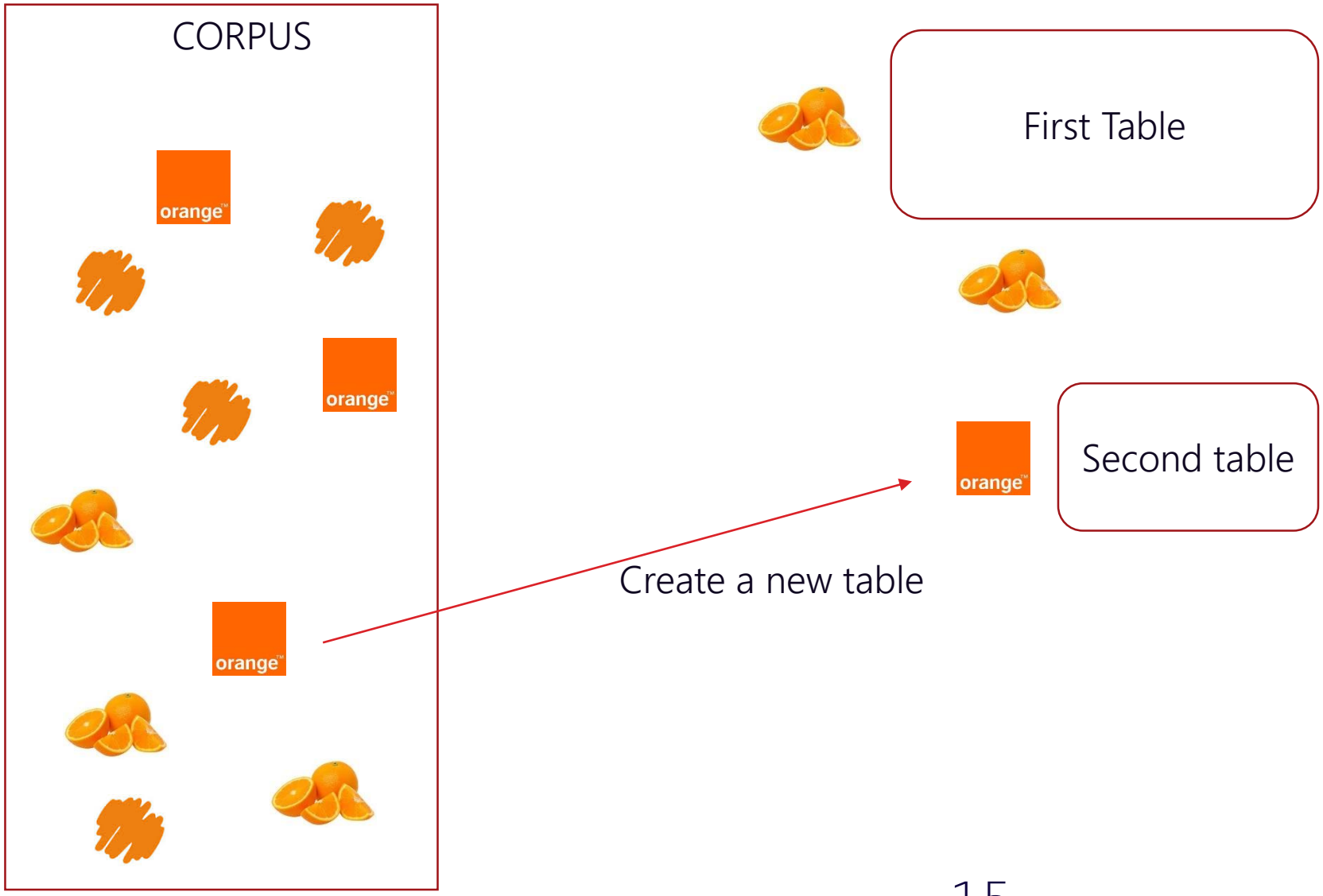


First Table

AdaGram : Chinese Restaurant process

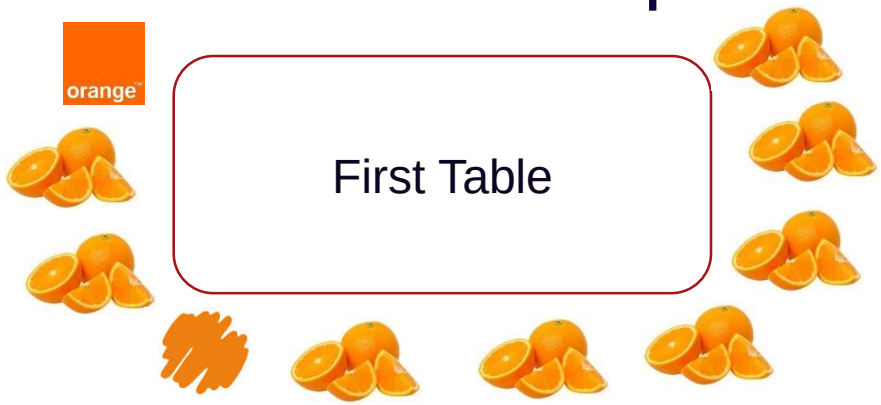


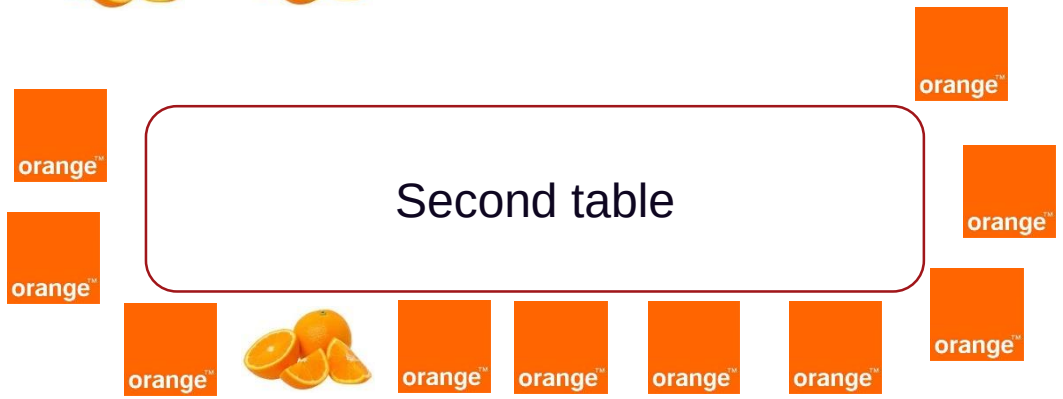
AdaGram : Chinese Restaurant process



AdaGram : Chinese Restaurant process

CORPUS

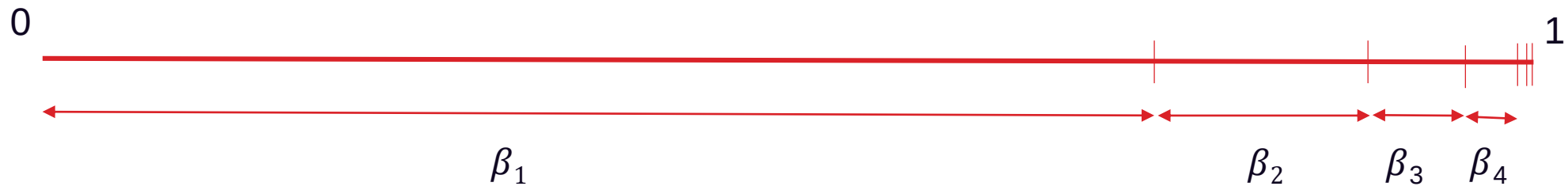




AdaGram : Bartunov et al. 2015

- We will learn one embedding per meaning for center words only.
- The discrete random variable z correspond to the assigment of a word to a meaning. We put a Dirichlet process (DP) prior on this r.v. The breaking-stick formulation is as follow :

$$p(z = k \mid x, \beta) = \beta_{xk} \prod_{r=1}^{k-1} (1 - \beta_{xr}) \quad , \qquad p(\beta_{xk} \mid \alpha) = Beta\,(\beta_{xk} \mid 1, \alpha) \qquad \forall k$$



- Advantages of the DP :
  - Insure that the number of meaning for a word is finite with probability 1
  - Compatible with variational inference methods for posterior approximation

AdaGram : Intractability and solution

- Overall model similar to Skip-Gram but with meanings added for center words:

$$p(Y, Z, \beta \mid X, \alpha, \theta) = \left(\prod_{x=1}^{V} \prod_{k=1}^{\infty} p(\beta_{xk} \mid \alpha) \right) \prod_{i=1}^{N} \left[p(z_i \mid x_i, \beta) \prod_{j=1}^{C} p(y_{ij} \mid z_i, x_i, \theta) \right]$$

- Marginal likelihood of the model : $\log p(Y \mid X, \theta, \alpha) = \log \int \sum_{Z} p(Y, Z, \beta \mid X, \theta, \alpha) d\beta$

Intractable

- Two principal methods for approximating the posterior distribution
  - Monte-Carlo Markov Chains
  - Variational inference

Variational Inference in a nutshell

- Hypotheses :
  - Bayesian model with observed variables (x), latent variables (z) and hyperparameter α.
  - Posterior distribution : p(z | x, α) intractable.
- Idea : Approximate the posterior by a simpler distribution (fully factorized = mean field variational inference) : q(z | θ)
- Method : Maximisation problem using KL divergence between true posterior and approximation

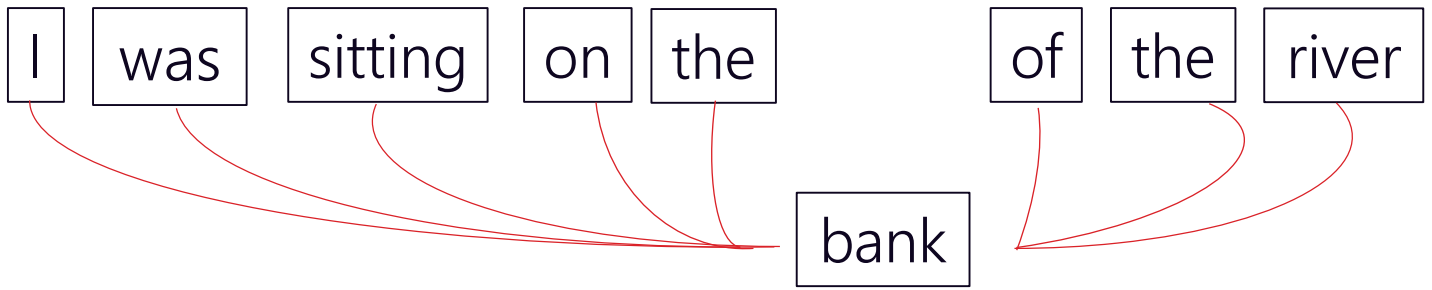
$$KL\,(q\,||\,p\,) = E_q\left[\log\left(\frac{q(z\mid\theta)}{p(z\mid x,\ \alpha)}\right)\right] = \log(p(x\mid\alpha)) - \underbrace{\left(E_q[\log(p(x,z\mid\alpha)] - E_q\left[\log\left(q(z\mid\theta)\right)\right]\right)}_{\text{Evience lower bound (ELBO)}} \geq 0$$

Maximising ELBO → Reducing KL divergence → q « close » to p

AdaGram : Disambiguation procedure

Prior probabilities (given by the corpus)

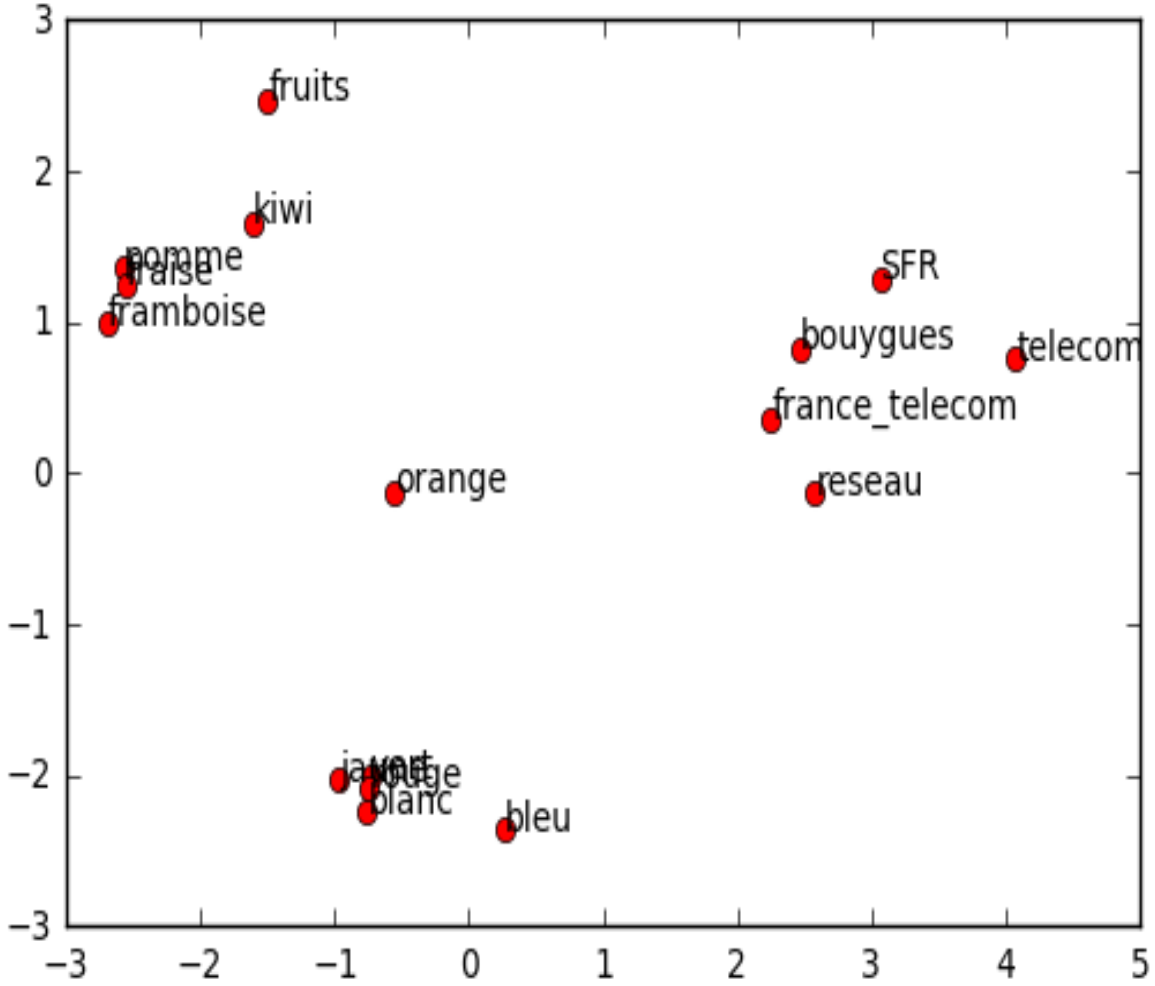
Financial institution	River bank
62%	38%



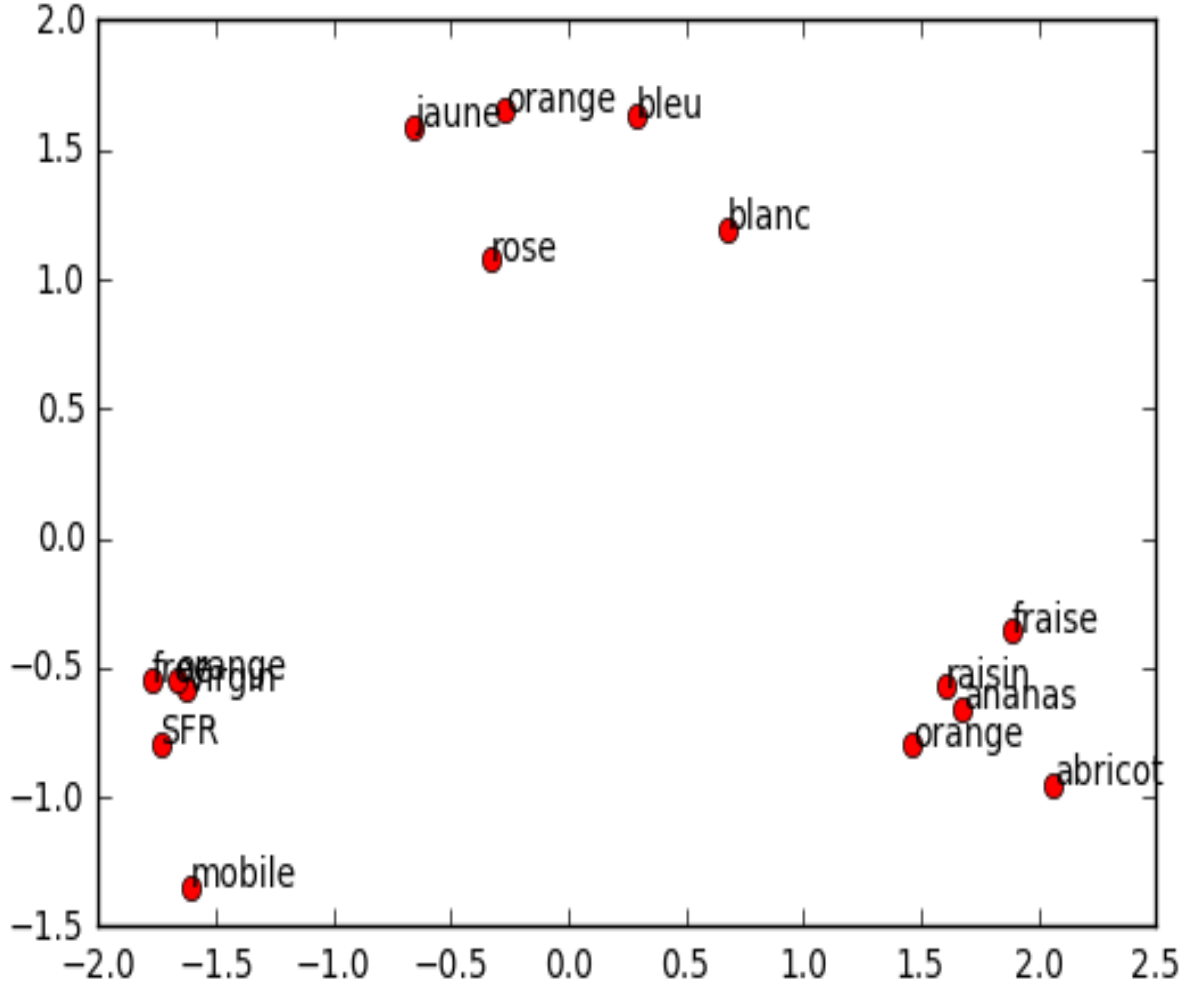
98%

2%

AdaGram : Visualisation



SkipGram



AdaGram

Training & Results

Training

	Number of tokens	Max number of meanings	Epochs	Training time
French Wikipedia	Arround 1 billion	5 or 10	2	Approximately 14h without GPU (28h for 10 meanings)
English Wikipedia	Arround 2 billions	5 or 10	2	Approximately 20h without GPU (40h for 10 meanings)

AdaGram is approximately T times slower than Skip-Gram where T is the maximal number of meanings to be learned

Results (English)

Bank and Apple: closest vector of most probable meanings and prior probability

Bank 2 : 48,3	Bank 2 : 37,1	Bank 3 : 7,0
Business	Side	Robbing
Firm	Slope	Hold-up
Transaction	Shore	Robbery
Company	River	Armed
Shareholding	Floodplain	Policeman
Corporation	Confluence	Shoplifter

Apple 1 : 38,2	Apple 2: 25,9	Apple 3 : 24,0
Itunes	Peach	Pear
Amazon	Avocado	Milk
Store	Pecan	Cake
Youtube	Fruit	Oak
Billboard	Citrus	Soup
Spanishcharts	honey	Corn

Results (French disambiguation)

- First sentence : « … j'ai pris un verre de couleur **orange** qui avait deux …. »
- Second sentence « …que j'achète une carte recharge **orange** et qu'elle ne fonctionne pas …. »

	Prior probability	Posterior probability : sentence 1	Posterior probability : sentence 2
Orange : cassis, agrumes , …	**28,4 %**	23,3 %	6,2 %
Orange : jus, pampril, ananas …	24,1 %	6,1 %	18,6 %
Orange : SFR, free, mobile	23,8 %	17,2 %	**65,5** %
Orange : rose, fluo jaune	17,3 %	**53,4 %**	7,4 %

References

- Sergey Bartunov, Dmitry Kondrashkin, Anton Osokin, and Dmitry Vetrov. 2015. *Breaking sticks and ambiguities with adaptive skip-gram.*
- David M. Blei, Michael I. Jordan. 2006. *Variational inference for dirichlet process mixtures.*
- David M. Blei, Alp Kucukelbir, Jon D. McAuliffe, 2016. *Variational Inference : a review for statisticians.*
- Martin J. Wainwright, Michael I. Jordan. 2008. *Graphical Models, exponential families and variational inference.*
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, Jeff Dean. 2013. *Distributed representations of words and phrases and their compositionality.*
- Jocelyn Coulmance, Jean-marc Marty, Guillaume Wenzek, Amine Benhalloum. 2015. *Trans-Gram, Fast cross-lingual Word-embeddings.*

Thank you for your attention. Questions?

Proxem – 105 rue La Fayette – 75010 Paris