INGI
Département d'ingénierie informatique

Comparative study of statistical word sense discrimination techniques

Marie-Catherine de Marneffe
Pierre Dupont

Computing Science Department, INGI – UCL

Outline

- What is word sense disambiguation?
- Techniques used for disambiguation: three categories
- Discrimination: two different models
  - ***Naïve Bayes model***
  - ***Vector based model***
- Results on a corpus
  (six months from the New York Times News)

Contributions

- Unified formalism
- Extensions of the models
  - ***Naïve Bayes model***:
    global strategy
  - ***Vector based model*** (Schütze's model):
    implementation of the Gaussian model

Disambiguation is helpful in many linguistic applications

The purpose of disambiguation is to ***determine*** the exact ***sense*** of an ***instance*** of an ambiguous word according to its ***particular use***.

Useful in any linguistic application where word sense matters such as

- automatic translation,
- text categorization,
- speech understanding,
- *etc*.

Three categories of disambiguation techniques

contexts
sense informants } $\Rightarrow$ **senses**

1. ***supervised techniques***
   Semantically *tagged* corpus (= training corpus)
   Senses defined by *semantic tags*
2. ***dictionary based techniques***
   Untagged corpus (= training corpus)
   Senses from a *dictionary* or a *thesaurus*
3. ***unsupervised techniques***
   Fully unsupervised: only an untagged corpus
   Senses ?? ***discrimination***

Statistical vs linguistic approach

The models presented are **based on statistics**.

$\downarrow$

seems **contradictory to linguistics**

But the data has poor linguistic information
(the only one is the stop-list).

Perspective:

To improve the models by adding linguistic information

How to discriminate? One generic model to be instantiated

w	:	an ambiguous word,
$s_1, \ldots, s_K$	:	the K possible senses of w,
$c_1, \ldots, c_I$	:	contexts of the I instances of w in a training corpus,
$v_1, \ldots, v_J$	:	J possible informants.

$$\hat{k} = \underset{k}{\operatorname{argmax}}\, P(s_k|c) = \underset{k}{\operatorname{argmax}}\, P(s_k)P(c|s_k)$$

Two "instances" of the generic model are

1. Naïve Bayes model (discrete)
2. Vector-based model (continuous)

A context window is used to defined the contexts

sense ??

$\downarrow$

— — — — — — — **w** — — — — — — —

informants = words in the context window

The only linguistic information is a stop-list.

Naïve Bayes model: contexts are bags of independent words

contexts (c_i)	=	bag of **informants**
informants (v_j)	=	content words belonging to the context window (i.e. words $\notin$ stop-list)

Independent informants because of the **Naïve Bayes assumption**:

$$P(c_i|s_k) = P(\{v_j \in c_i\}|s_k) = \prod_{v_j \in c_i} P(v_j|s_k)$$

The parameters of the model are $P(v_j|s_k)$ and $P(s_k)$
Estimated to maximize the likelihood of the training corpus via the EM algorithm

Naïve Bayes model: decision rule

The ***final decision*** rule is:

$$\hat{k} = \underset{k}{\operatorname{argmax}}\, P(s_k|c) = \underset{k}{\operatorname{argmax}}\, P(s_k) \prod_{v_j \in c} P(v_j|s_k)$$

Naïve Bayes model: decision rule

The ***final decision*** rule is:

$$\hat{k} = \underset{k}{\operatorname{argmax}}\, P(s_k|c) = \underset{k}{\operatorname{argmax}}\, P(s_k) \prod_{v_j \in c} P(v_j|s_k)$$

Generic final decision rule :

$$\hat{k} = \underset{k}{\operatorname{argmax}}\, P(s_k|c) = \underset{k}{\operatorname{argmax}}\, P(s_k) P(c|s_k)$$

Choice of the informants: 2 strategies

- a **local** strategy
  The informants are defined for ***each*** ambiguous word: they are all the content words occurring in all the context windows around the different instances of w.

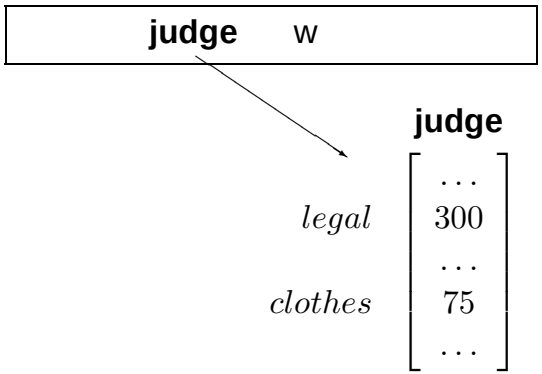
- a **global** strategy
  The informants are the same for ***all*** ambiguous words: they are the most frequent words of the corpus (stop-list excluded).

Outline

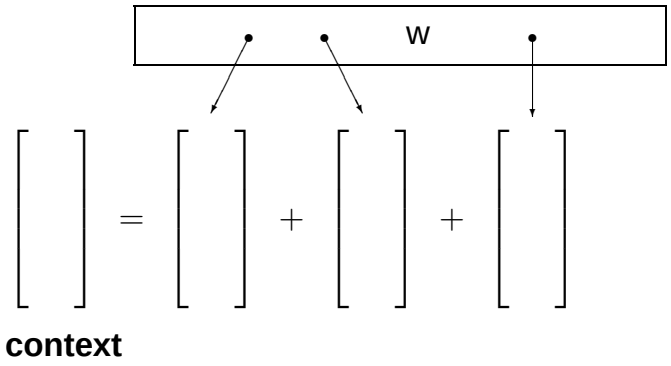
- What is word sense disambiguation?
- Techniques used for disambiguation: three categories
- Discrimination: two different models
  - Naïve Bayes model
  - ***Vector based model***
- Results on a corpus
  (six months from the New York Times News)

Vector-based model: informants are represented by vectors

Each **informant** is represented by a vector of co-occurrence frequencies in the context window between the informant and other words.



Vector-based model: contexts = sum of informants

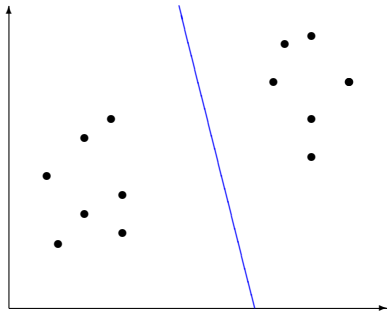


context

Similar contexts form clusters in the vector space

Similar context vectors can be seen as forming clusters in vector space.

Each ***cluster*** represents ***one sense*** of the ambiguous word.



How clusters are defined?

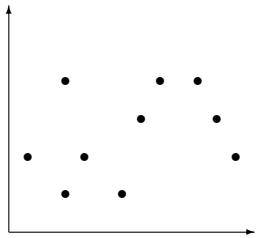
K-means

(1) Initialization: fixes K means randomly

(2) Iteration:

– Given the K means, vectors are assigned to their closest mean
(euclidean distance).

– Cluster means are then recomputed.



How clusters are defined?

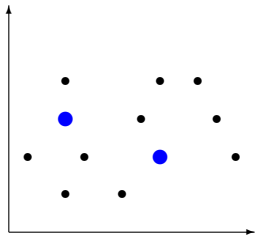
K-means

(1) Initialization: fixes K means randomly

(2) Iteration:

– Given the K means, vectors are assigned to their closest mean
(euclidean distance).

– Cluster means are then recomputed.



How clusters are defined?

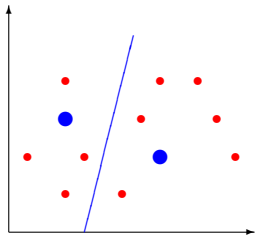
K-means

(1) Initialization: fixes K means randomly

(2) Iteration **0**:

– **Given the K means, vectors are assigned to their closest mean**
(euclidean distance).

– Cluster means are then recomputed.



How clusters are defined?

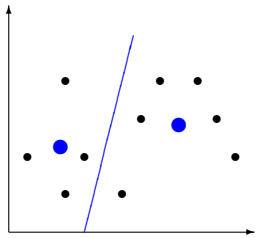
K-means

(1) Initialization: fixes K means randomly

(2) Iteration **0**:

– Given the K means, vectors are assigned to their closest mean
(euclidean distance).

– **Cluster means are then recomputed.**



How clusters are defined?

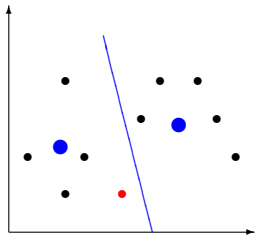
K-means

(1) Initialization: fixes K means randomly

(2) Iteration **1**:

– **Given the K means, vectors are assigned to their closest mean**
(euclidean distance).

– Cluster means are then recomputed.



How clusters are defined?

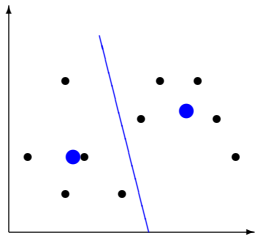
K-means

(1) Initialization: fixes K means randomly

(2) Iteration **1**:

– Given the K means, vectors are assigned to their closest mean
(euclidean distance).

– **Cluster means are then recomputed.**



How clusters are defined?

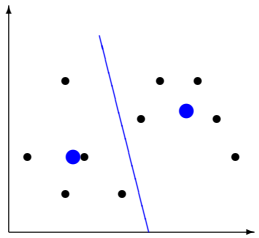
K-means

(1) Initialization: fixes K means randomly

(2) Iteration **2**:

– **Given the K means, vectors are assigned to their closest mean**
(euclidean distance).

– Cluster means are then recomputed.



How clusters are defined?

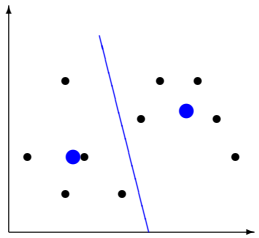
K-means

(1) Initialization: fixes K means randomly

(2) Iteration **2**:

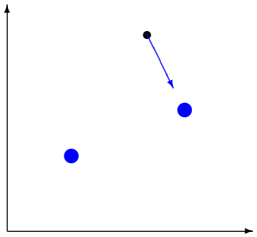
– Given the K means, vectors are assigned to their closest mean
(euclidean distance).

– **Cluster means are then recomputed.**



Sense discrimination

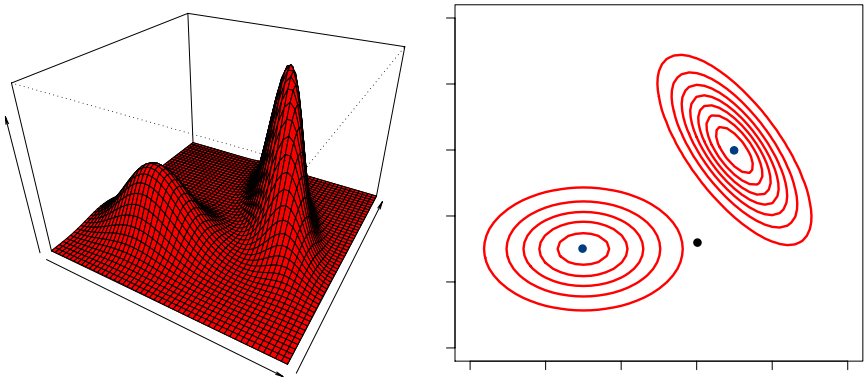
A context representing a nieuw instance will be assigned to the closest cluster mean.



A distribution of the cluster can be assumed

Mixture of K Gaussian components

Each cluster is assumed to follow a Gaussian distribution.



The ***final decision*** rule is:

$$\hat{k} = \underset{k}{\operatorname{argmax}}\, P(s_k)P(c|s_k) = \underset{k}{\operatorname{argmax}}\, P(s_k)f_k(\vec{x})$$

In practice

The experiments are done on six months from the 1997 New York Times News:

$\sim$ 500 megabytes

$\sim$ 80 million words

$\sim$ 500 thousand different words

We follow Schütze's choices in **parameters setting**:

– context window size

– number of informants (vocabulary size)

for **comparison purposes**.

The number of senses is fixed: K = 2.

Use of pseudowords to test performance

To test performance, a large number of instances has to be disambiguated by hand.

To avoid that, we can generate artificially ambiguous words, called *pseudowords*.

A ***pseudoword*** is the concatenation of two or more natural words.

For example: *banana-moon*.

Performance measure

We have considered 5 pseudowords.

The models include **random initialization**.

↓

experiments repeated **10 times**
with different random seeds

↓

mean and ***standard deviation***
for each pseudoword

The ***accuracy*** gives the total proportion of correctly labeled instances for both senses.

Naïve Bayes: global strategy performs better in average

Vector Model: Gaussian is worthy

	Naïve Bayes		*Vector Model*	
	Local	Global	K-Means	Gaussian
Average accuracy (%)	63.7	68.3	64.1	75.9
Standard deviation (%)	7.2	5.0	5.9	0.2
CPU Time (sec)	.4	.5	10.5	36.7

The Gaussian model offers the best average accuracy with computational costs still tractable.

Conclusions

– Description of both models in a unified statistical formalism in order to stress the similarities and differences

– Vector-based model can be improved when a real Gaussian model is estimated

– Interest of a global selection of content words to characterize the context of an ambiguous instance in the Naïve Bayes model

Future work

– Influence of the parameters
(stop-list, context window size, vocabulary size, value of the number of clusters)

– Influence of Latent Semantic Indexing in the vector model (SVD)

– Automatic determination of an optimal value of the number of clusters

– Addition of linguistic information such as

stemming
POS tagging
semantically more relevant context window

– Integration in an application (text categorization)