Cue-based Lexical Information Acquisition

Núria Bel - Muntsa Padró

22-23 June 2011

INSTITUT UNIVERSITARI DE LINGÜÍSTICA APLICADA

UNIVERSITAT POMPEU FABRA

Automatic acquisition of lexical information

To find how to automatically produce lexica, i.e. repositories of language dependent lexical information, for those technologies that actually DO need this information to work:

MT, IE, Topic/Event Detection and tracking, Opinion Mining, Question Answering, Text Analytics, etc.

Some samples of entries

```
("paralelo"                AST
   ALO     "paralel"
   ATR      POST
   CL       (PF-AS PM-OS
SF-A SM-O)
   FC       (NPP)
   LY       AMENTE
   MC       ("a")
   PLC      (NG)
   PRED     (ESTAR SER)
   TA       (OBJ-P REL)
   AUTHOR   "juan"
   DATE     "31-Aug-99"
   SITE     "FB52")
```

```
("libro"                   NST
   ALO     "libr"
   CL       (PM-OS SM-O)
   GD       (M)
   KN       CNT
   PLC      (NF)
   TYN      (CNC SEM)
   AUTHOR   "juan"
   DATE     "28-Aug-99"
   SITE     "FB52")
```

```
(“amor"                    NST
   ALO     “amor"
   CL       (PF-ES)
   GD       (F)
   KN       MS
   PLC      (NF)
   TYN      (ABS)
   AUTHOR   "juan"
   DATE     "28-Aug-99"
   SITE     "FB52")
```

Entries borrowed from MT system Incyta (Metal family)

Coverage, consistency and costs of current handcrafted lexica

Agnese, 1544

Time to produce: the bottleneck

CERN's director Lew *Kowarski* (1960):

"Clearly the problem is that of speed, and since human attention and action introduce a rock-bottom bottleneck, speed can be achieved either by pouring in parallel through many bottlenecks, or by eliminating them altogether. **Either vast armies of slaves armed with templates** and desk calculators or few people operating a lot of discriminating and **thinking machinery**. The evolution is towards the elimination of humans, function by function"

(cf. P.L. Galison, *Image and logic*, 1997)

Automatic Lexical Acquisition

- Manning (1993)
- **Brent (1993)**
- Briscoe & Caroll (1997)
- Korhonen (2000)
- **Merlo and Stevenson (2001)**
- Baldwin and Bond (2003)
- **Joanis and Stevenson (2003)**
- Baldwin (2005)
- Zang and Kordoni (2006)
- Bel et al. (2007)
- **Joanis et al (2007)**
- Etc.

Brent (1993)

- Michael R. Brent. 1993. From Grammar to Lexicon: Unsupervised Learning of Lexical Syntax. Computational Linguistics, 19:203–222.

  "How can easily recognized, surface grammatical facts be used to extract from a corpus as much syntactic information as possible about individual words? …

  an approach based on two principles. **First**, rely on local morpho-syntactic cues to structure rather than trying to parse entire sentences. **Second**, treat these cues as probabilistic rather than absolute indicators of syntactic structure. Apply inferential statistics to the data collected using the cues, rather than drawing a categorical conclusion from a single occurrence of a cue."

Statistical inference, silence and noise

- Do not try to draw categorical conclusions about a word on the basis of one or a fixed number of examples. Instead, attempt to determine the distribution of exceptions to the expected correspondence between cues and syntactic frames. Use a statistical model to determine whether the co-occurrence of a verb with cues for a frame is too regular to be explained by randomly distributed exceptions.
- The cues are fairly rare, so verbs that occur fewer than 15 times tend not to occur with these cues at all.
- Further, these cues occur fairly often in structures other than those they are designed to detect.

Noise?

- For example, *record*, *recover*, and *refer* all occurred with cues for an infinitive, although none of them in fact takes an infinitive argument.

1. But I shall campaign on the Meyner record to meet the needs of the years ahead.
2. Sposato needed a front, some labor stiff with a clean record to act as business agent of the Redhook local.

Silence

Silence

in a corpus of 3,334,563 tokens, an adjective like ‘applicable’ appears 440 times, and a 37% of these co-occurring with its bound preposition ‘to’.

In the same corpus, the adjective ‘favorable’ occurs 60 times, and only a 5% co-occurring with its bound preposition ‘to’,

while ‘generous’ that occurs 7 times is never found with its bound preposition ‘with’

Statistical Results in SCF's

Probability of a SCF given a Verb

F1 results

- **Briscoe and Carroll, 1997, 0.55**
- **Korhonen, 2002, 0.76, LC***
- **Chesley &Salmon-Alt, 2006, 0.66**

*** Improvement due to the use of Levin lexical classes as an informed back-off**

IN

Cue-based lexical classification

- **Merlo & Stevenson (2001),** Automatic Verb Classification Based on Statistical Distributions of Argument Structure. *Computational Linguistics,* 27:3.
- Specifically, the proposal is to automatically classify intransitive verbs into 3 Levin alternation-based classes based on argument structure properties,
- To (1) **develop statistical indicators that are able to determine the class** of an optionally intransitive verb by capturing information across its transitive and intransitive alternants. These indicators/cues/features serve as input to a (2) **machine learning algorithm, under a supervised training methodology**, which produces an automatic classification system for our three verb classes.
- The method assigns a classification to verbs that have not previously been seen in the training data.

(1) Cue-based lexical classification

- We identify occurrence contexts related to **lexical classes**, and we collect information about actual occurrences of a particular word in these contexts in order to classify a given word as belonging to a particular class or not.
- If there are motivated lexical classes, differences in the distribution of occurrence contexts can be indicators/ cues to train a classifier.
- For instance, to induce whether a noun can be classified as *mass* noun or not, its co-occurrence with particular determiners will be taken as a cue:

  some / *many mud
- Words are represented in terms of a collection of features which are taken as indicators or cues for a word to be classified as belonging to a particular class.

(2) Machine Learning methods

- Supervised Methods such as Decision Tree's, Support Vector Machines, Bayesian, Hidden Markov Models's, among others, have been used for Automatic Lexical Acquisition approached as a supervised classification problem.
- A learner is supplied with classified examples of words represented by numerical information about matched and not matched contexts.
- The exercise is to confirm that the data characterized by the lexical class motivated cues, indeed support the division into the proposed lexical classes by correctly classifying new words.

How Machine Learning?

- ML methods start having as input the target word along with a portion of the text in which it is embedded. This is "the context". This context can be pre-processed in different ways, is part-of-speech tagged.
- the input is reduced to a fixed number of features that are selected to capture information relevant to the learning task. They form the "feature vector".
- Linguistic Features can be, or a combination of both:
  - collocational: the words that occupy specific positions at the left or at the right of the target words, or information about them, i.e. POS.
  - co-occurrence: also the neighboring words but without information about the position. The number of times a particular word occurs close to the target word.

State of the Art

- Merlo and Stevenson (2001) and Joanis et al. (2007) achieved an accuracy, i.e. the number of correct classifications among all the classifications, around a 70% for the task of Levin-based lexical class classification.

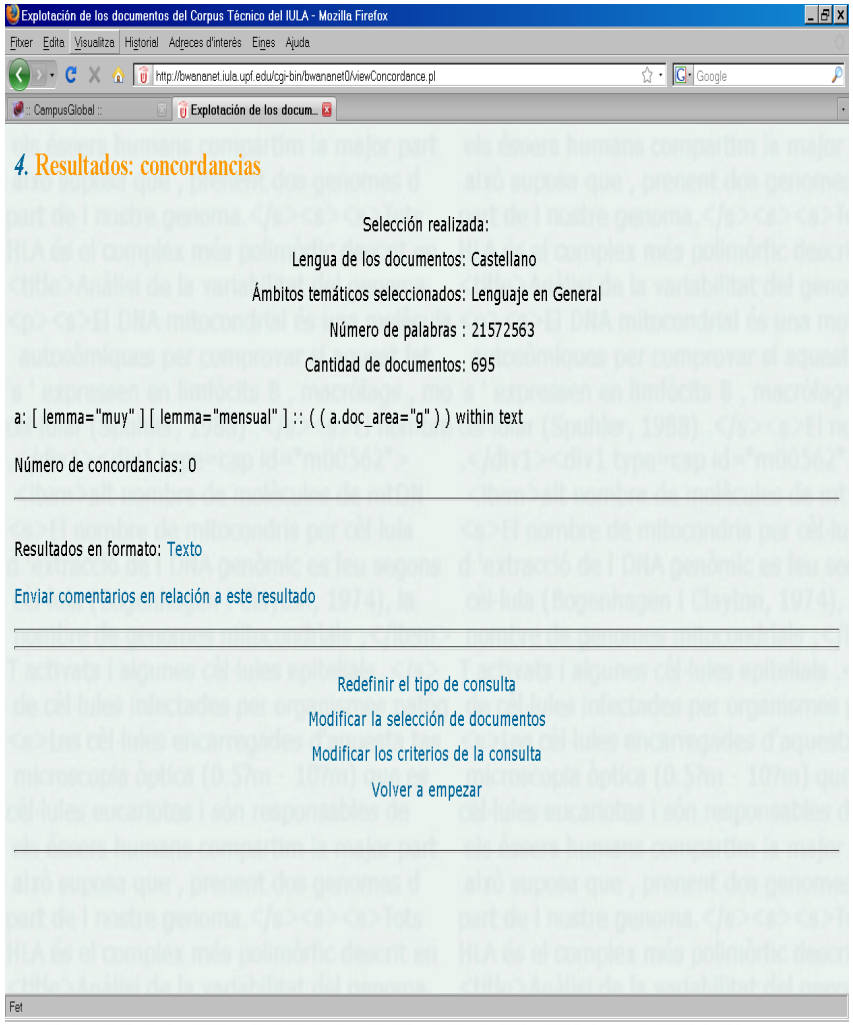
Acquiring Lexico-semantic Information

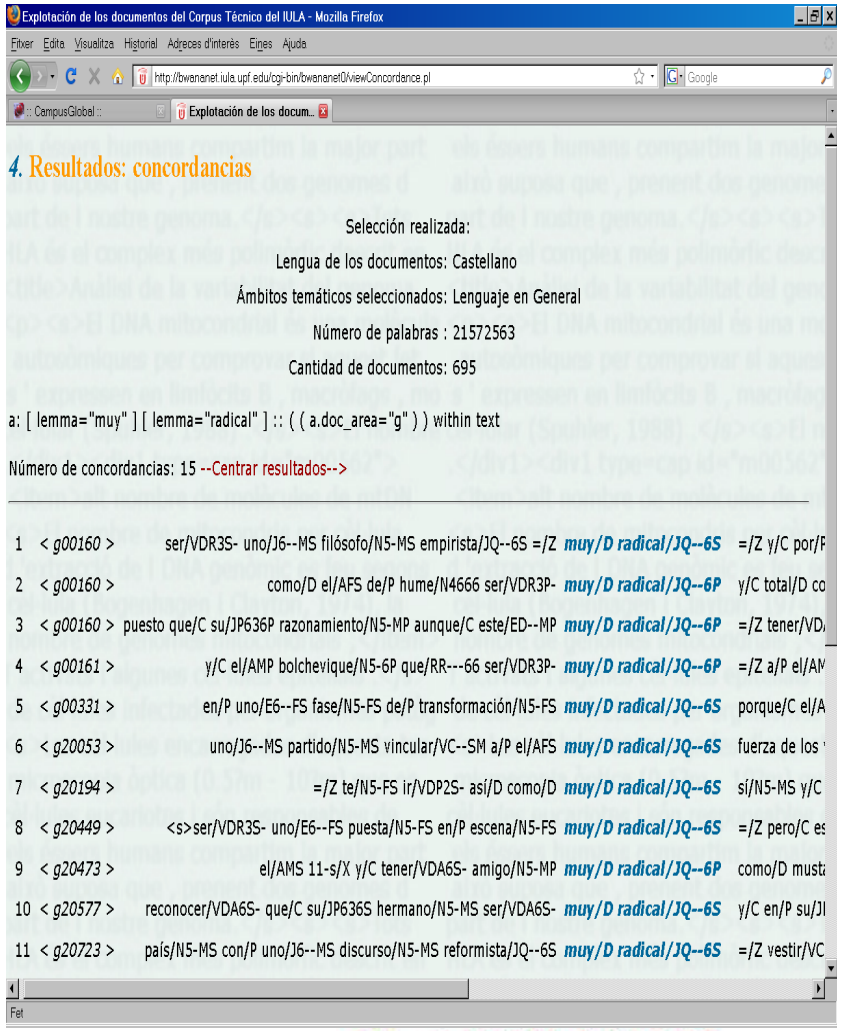
- Lexico-semantic information too?
- Types of words? Relational adjectives, mass nouns?
- Internal Structure Properties: Events, HUM, LOC..

Session 2

Distributional Hypothesis and cue identification: the linguistic facts and their implementation as Regular Expressions

mensual (monthly) vs. *radical*





Lexical types can be said to be built on these differences

TYPE / SF	PreNom	Pred	Grad	Prep	Sent
a_adv_event e.g. *principal* ('main')	yes	no	no	no	no
a_rel_pred e.g. *accidental* ('accidental')	no	yes	no	yes	no
a_rel_npred e.g. *administrativo* ('administrative')	no	no	no	no	no
a_qual_trans e.g. *común* ('common')	yes	yes	yes	yes	no
a_qual_int e.g. *agresivo* ('agressive')	yes	yes	yes	no	no
a_qual_sent e.g. *consciente* ('conscientious')	no	yes	yes	yes	yes

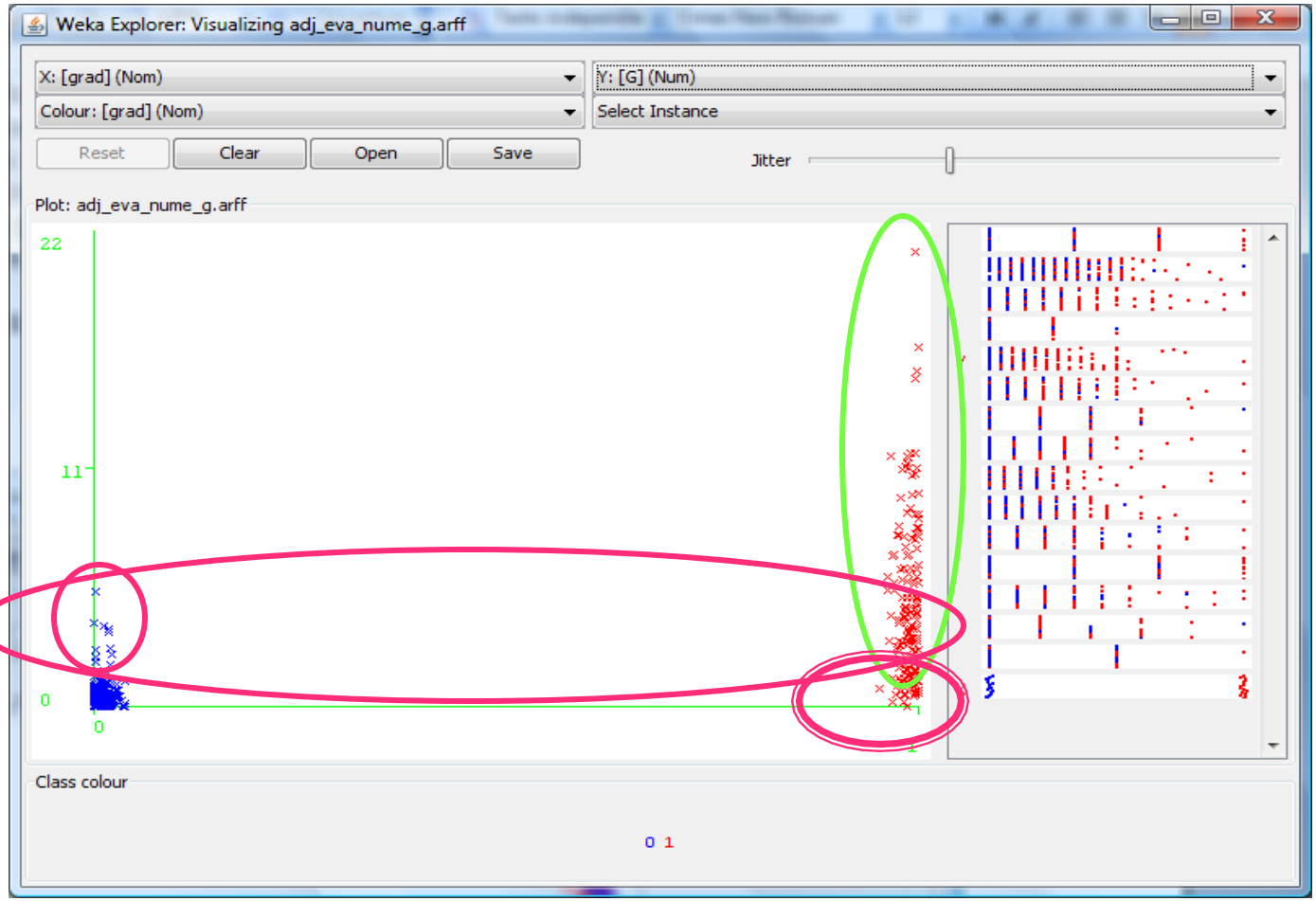
Practicalities

- **Word's occurrences are converted into attribute vectors. We use lemmatized pos tagged corpus and Regular Expression implemented patterns to build vectors.**
- `# 5 gradable_adj_muy`

  `5&<&((casi|demasiado|tan|muy|bastante|poco|menos|más|(esencial|virtual|absoluta|práctica|especial|extremada|alta|fuerte)mente)\/[ED][A-Z0-9-]+)\s\#`
- @data

  15,2,8,4,0,**8**,1,0,3,1,3,3,0,1

Analysing data: 444 adjectives and "Gradual"



Using Machine Learning: DT's

- Decision Trees as classifiers: C4.5 (Quinlan, 1993). DT's perform a general to specific search in a feature space, selecting the most informative attributes for a tree structure. The goal is to select the minimal set of attributes that efficiently partitions the feature space into the classes of observations and assemble them into a tree.
- Mitchell (1987) recommends their use for:
  - Instances are represented by attribute-value pairs
  - Outputs can be discrete values, yes/no
  - The training data may contain errors
  - The training data may contain missing values
- Besides, DTs allow for inspection of the results that can be interpreted easily.

Cues, classification and state-of-the-art results

- Merlo and Stevenson (2001) selected very specific cues ad-hoc for classifying verbs into a number of Levin (1993) based verbal classes: anymacy of the subject, passives, ...
- Baldwin (2005) used general features, such as the *pos* tags of neighboring words for type classification.
- Joanis et al. (2007) used the frequency of filled syntactic positions or slots, tense and voice of occurring verbs, etc., to describe the whole system of English verbal classes.
- Difficult to compare results, but .. an accuracy of about 70%

Classifying Gradual Adjectives

DT C4.5 as implemented in Weka. Occurrences got from IULA's **economy corpus 1.091.314 tokens**

=== Summary ===

Correctly Classified Instances	409	**92.1171 %**
Incorrectly Classified Instances	35	7.8829 %
Kappa statistic	**0.8388**	
Total Number of Instances	444	

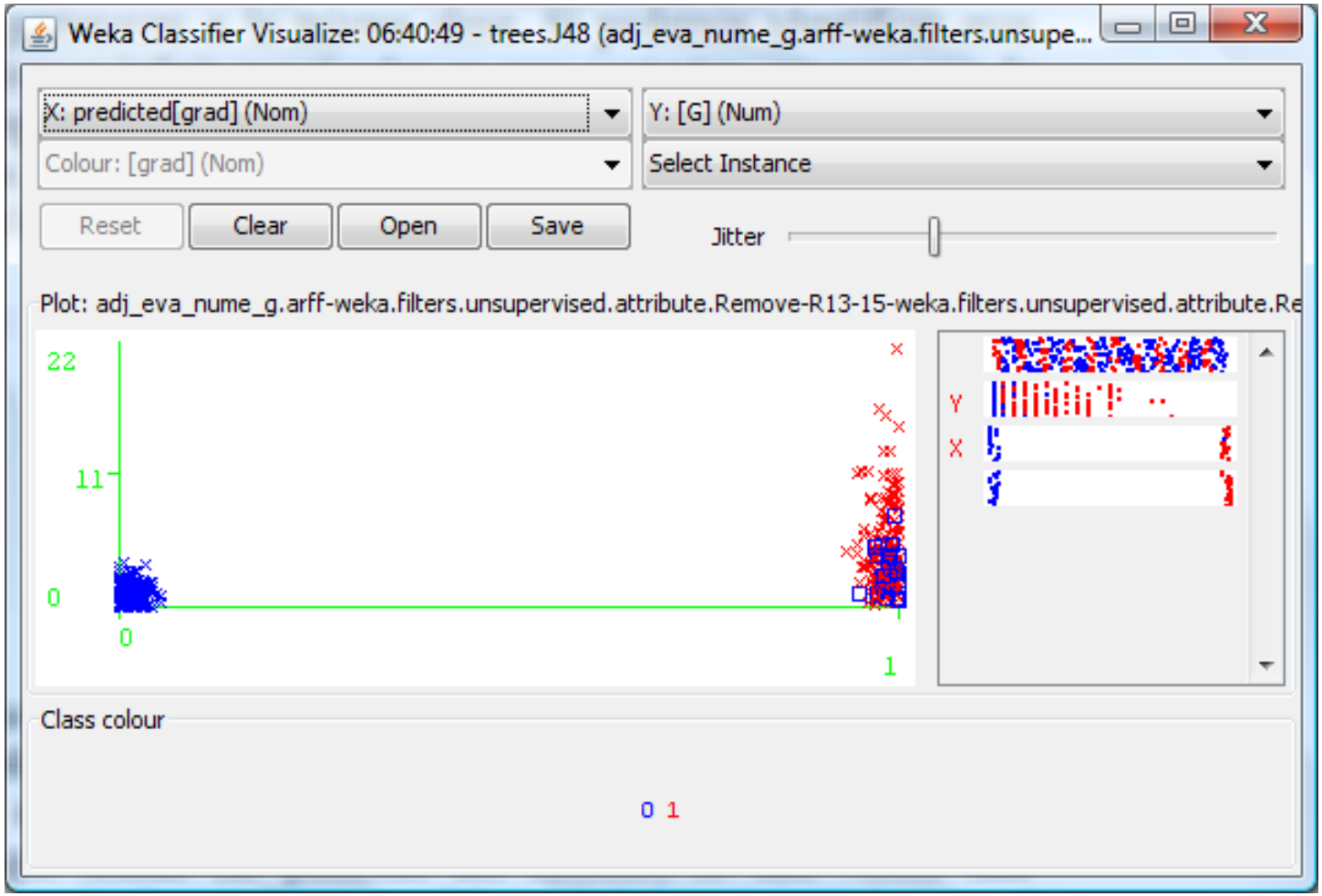
=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.88	0.012	0.992	0.88	0.932	0.912	0
0.988	**0.12**	0.836	0.988	0.906	0.912	1

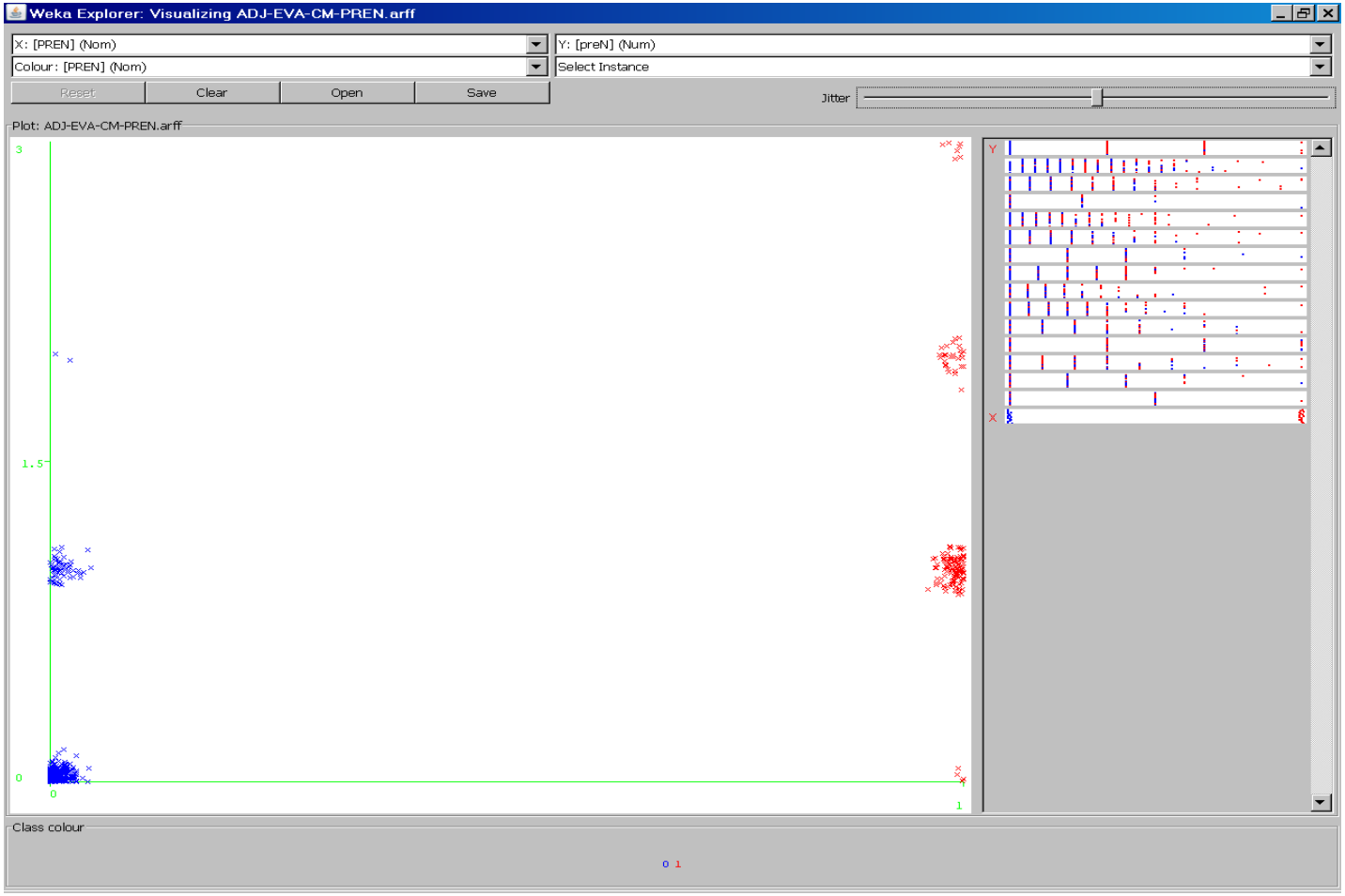
=== Confusion Matrix ===

a	b	<-- classified as
241	33	a = 0
2	168	b = 1

Graphically



Pre-Nominal feature and cues: a noiser scenario



Samples of noise and its impact

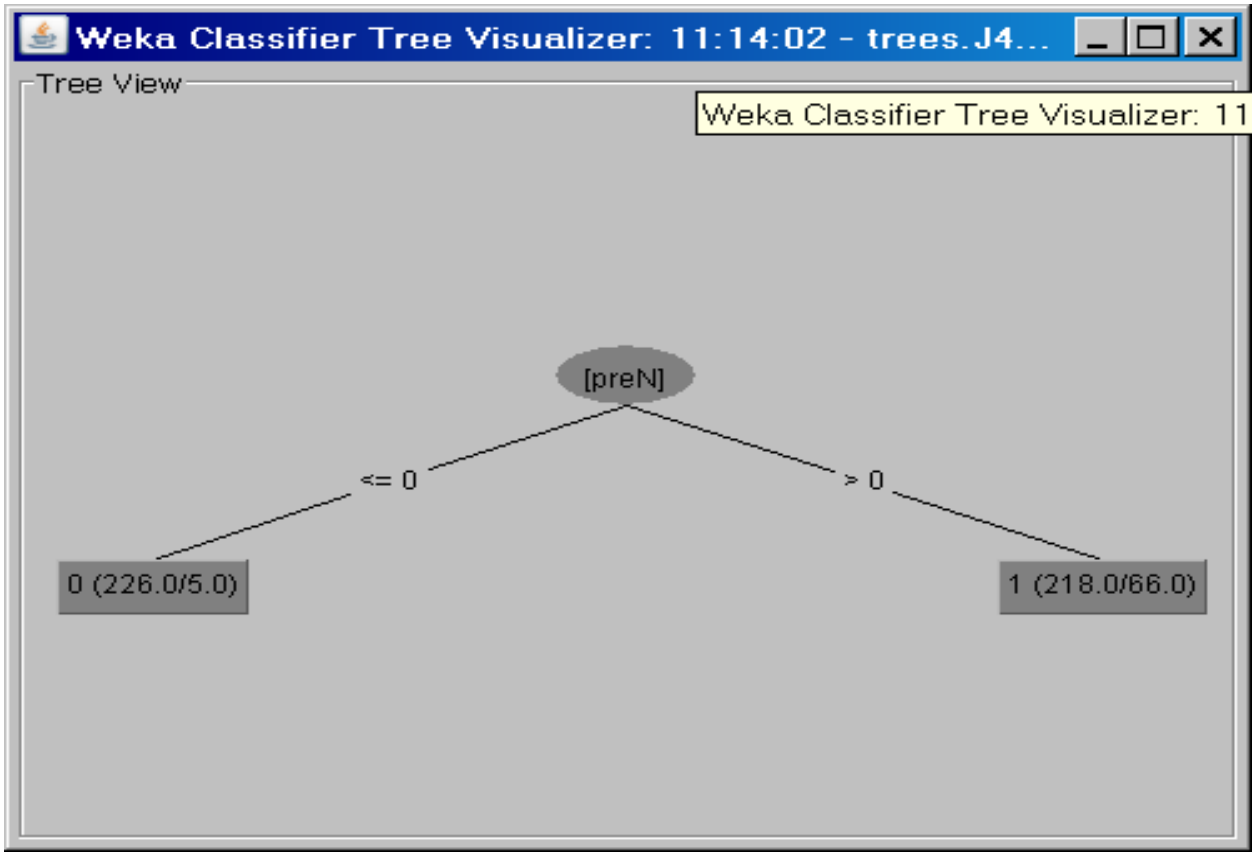
- What is noise?:
  - "tal que maximiza sus beneficios **conjuntos sujeto** a la ..."
  - "basta con considerar como **conjunto origen** el conjunto .."
- **Impact in DT Classification:**

Correctly Classified Instances	**373**	**84.009 %**
Incorrectly Classified Instances	71	15.991 %
Kappa statistic	0.6785	
Total Number of Instances	444	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Prec	Recall	F-Meas	Class
0.77	0.032	0.978	0.77	0.862	0
0.968	0.23	**0.697**	0.968	0.811	1

Experiment-1 Decision Tree



Correlation Experiments: Prenom+G

Correctly Classified Instances	390	87.8378 %
Incorrectly Classified Instances	54	12.1622 %
Kappa st		
Total Nur		

Experiment-1

Correctly Classified Instances	**373**	**84.009 %**
Incorrectly Classified Instances	71	15.991 %
Kappa statistic	0.6785	

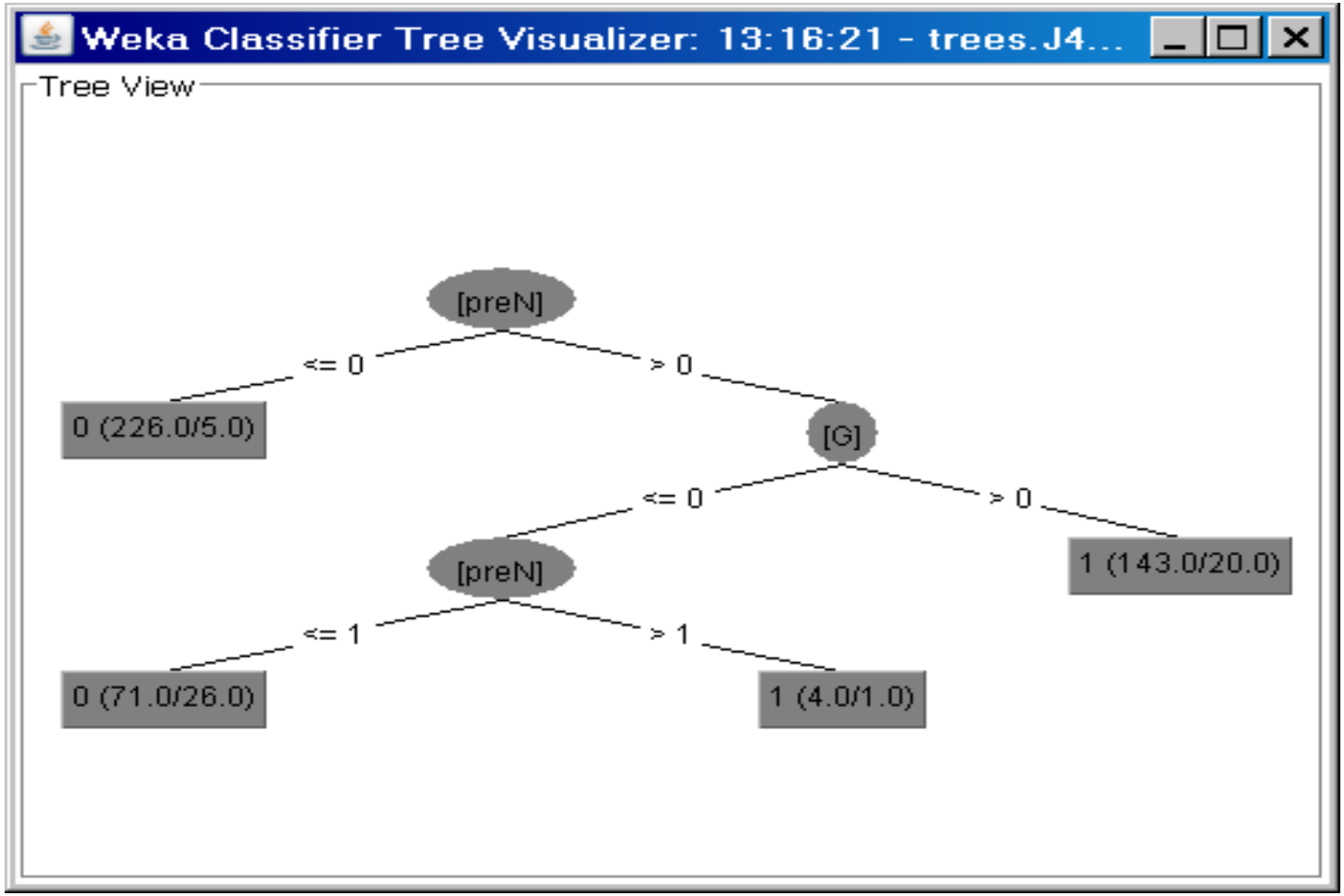
=== Detail

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.93	0.217	0.887	0.93	0.908	0.904	0
0.783	0.07	0.86	0.783	0.82	0.904	1

=== Confusion Matrix ===

a	b	<-- classified as
267	20	a = 0
34	123	b = 1

Experiment-2F DT



A more complex case: Classifying Nouns

- We have used the Spanish Resource Grammar (Marimon et al. 2007) typology for identifying the features and cues that justify lexical nominal types.

TYPE / SF	Mass	Count	Intrans	Trans	Prep
n_int_mass e.g. acidez ('acidity')	Yes	No	Yes	No	No
n_int_mass_count e.g. aceite ('oil')	Yes	Yes	Yes	No	No
n_int_count e.g. mesa ('table')	No	Yes	Yes	No	No
n_trans_count e.g. traductor ('translator')	No	Yes	No	Yes	No

Features and cues for Nouns

- For countable=yes, plural morphology
- For mass=yes, to be the singular head of a noun phrase without determiner occurring immediately after some verbs "hay barro" ('there is mud') and the co-occurrence of the noun in singular with certain quantifiers: más ('more'), menos ('less') and bastante ('enough').
- For trans=yes,
  - nominalization suffixes such as "-ción", "-sión" and "-miento",
  - definiteness in the complements, e.g. "aceleración de la economía" (*acceleration of the economy*) vs. "mesa de juegos" (*table of games*)
  - two PPs introduced by the preposition *de* ('of') as in "la colección de coches de mi hermano' (*the collection of cars of my brother*).
  - And to find the bound preposition of complements, we used a pattern for each possible preposition found after the noun in question "el acceso a" (*the access to*).

Classifying Nouns as Mass=yes/no

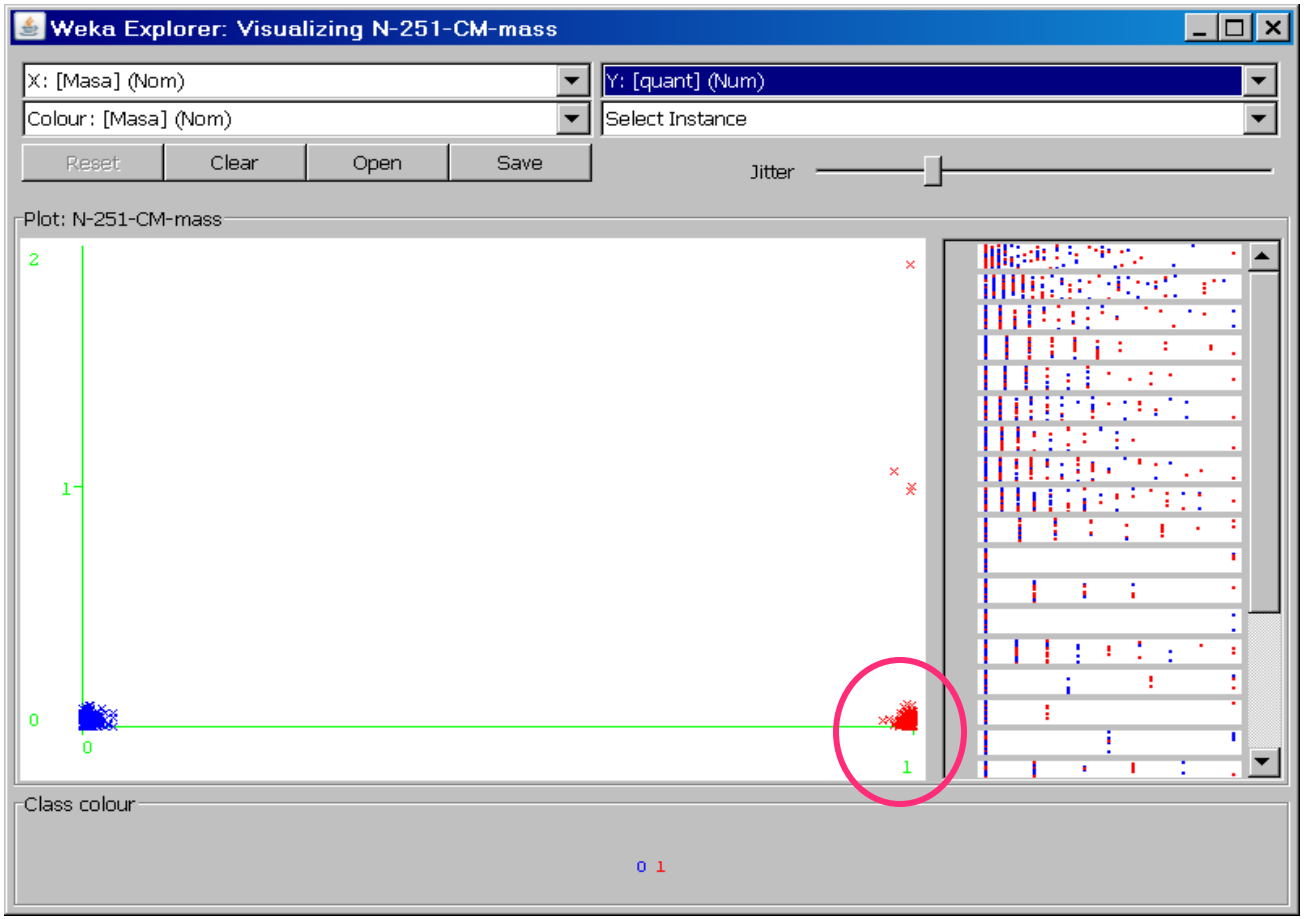
- More Frequent based Baseline = 74.81 (calculated in a 35.000 lemmas MT dictionary). Dataset of 250 Spanish Nouns, handcoded gold-standard.
- Difficult to compare to other's experiments but Baldwin and Bond (2003) had 89% for English but double classification.
- Our first results:

Correctly Classified Instances	199	79.9197 %
Incorrectly Classified Instances	50	20.0803 %
Kappa statistic	0.5783	
Total Number of Instances	249	

=== Detailed Accuracy By Class ===

TP	FP	Precision	Recall	F-Measure	Class
0.858	0.287	0.814	0.858	0.836	0
0.713	0.142	0.774	0.713	0.742	1

The problem of silence



Sparse data

- Joanis and Stevenson, 2003; Joanis et al. 2007; Korhonen et al. 2008 also mention that they have to face the problem of sparse data, many of the types/words are low in frequency and show up very little information.
- Most of the words (according to Zipff law) will appear very little and will show few cues.
- Yallop et al. (2005) calculated that in the 100M-word British National Corpus, from a total of 124,120 distinct adjectives, 70,246 occur only once. The cues we can use as information are mutually exclusive, i.e. an adjective can be prenominal and postnominal, but if it only occurs once, it will only show one cue, the other being a zero value.
- The optional nature and variety of the contexts of occurrence are the origin of missing values also for those types that occur more than once. For instance, in our previous example the word ‘mud’ will also appear with other less informative determiners “the mud”.

Cero values and learning

- For those approaches to lexical acquisition that count on the positive (or number of positives) vs. the negative observation of defined cues, their variety and optional nature create **not only a problem of enough information to decide, but a further uncertainty when learning from the data.**
- The uncertainty is that a zero value could be indeed a negative value, i.e. the cue is that it has not been observed, but it could be that the cue was just not observed in the examined corpus because of its optional nature.
- Note that although C4.5 DT handles missing values by assigning a probability to each of the possible values that is calculated based on the frequencies of the various values of A among the examples at node n, when there are many empty values, **the cue loses its predictive power because of the mentioned uncertainty.**
- Katz (1987) and Baayen and Sproat (1996), among others, acknowledged the importance of preprocessing low frequency events for Markovian methods. And Joanis et al. (2007) also decided to smooth the data, even working with more than 1000 occurrences per verb in the BNC.

Some results for lexical classes

• Event	EN	76.6%	167
• Abstract	EN	58.6%	
• Human	ES	76.2%	3789

Cues that we have used

- Co-occurrence with prepositions
- Co-occurrence with frequent verbs, relying on selectional restrictions
- Arguments or dependents