Natural Language Engineering

Pattern Recognition and Artificial Intelligence

Technical University of Valencia, Spain

Paolo Rosso

Research topics

1. **Knowledge-based lexical disambiguation**
2. **Question Answering** (QA)
   1. Multilingual QA
   2. Cross-language QA: multi-translator integration
3. **The web as lexical resource**
   1. Web-based lexical disambiguation
   2. Web-based lexical pattern extraction
   3. Web-based QA
4. **Information Retrieval** (IR) **and categorization**
   1. Semantic (geographical) IR
   2. Semantic text categorization
   3. Transition point indexing reduction technique
5. **Text clustering**
   1. Clustering of very short narrow-domain texts
   2. Cluster analysis of transcribed spoken dialogues

The Natural Language Engineering subgroup

Ph.D. students:

Davide Buscaldi

David Pinto

Yassine Benajiba

Rafel Guzmán

e- Natalia Ponomareva

In collaboration with:

other colleagues and Ph.D. students of the main group

University of Genova, Italy (Stefano Rovetta)

INAOE (Manuel Montes) and NPI (Mikhail Alexandrov), Mexico

Research topics

1. **Knowledge-based lexical disambiguation**
2. Question Answering (QA)
   1. Multilingual QA
   2. Cross-language QA: multi-translator integration
3. The web as lexical resource
   1. Web-based lexical disambiguation
   2. Web-based lexical pattern extraction
   3. Web-based QA
4. Information Retrieval (IR) and categorization
   1. Semantic (geographical) IR
   2. Semantic text categorization
   3. Transition point indexing reduction technique
5. Text clustering
   1. Clustering of very short narrow-domain texts
   2. Cluster analysis of transcribed spoken dialogues

Knowledge-based lexical disambiguation

- **Word Sense Disambiguation (WSD)** consists in examining word tokens and specifying exactly which **sense** of each **word** is being used;
- A **word** is usually disambiguated along with a portion of the text in which it is embedded (its **context**);
- External **lexical resources** are often used in WSD

Knowledge-based lexical disambiguation

- The **WordNet ontology** as external lexical resource; developed at Princeton University: http://www.cogsci.princeton.edu/~wn/
- It is based on **synsets** (set of synonyms defining a lexical concept), connected by various semantic relations such as:
  - *Synonymy*
  - *Hypernymy (is_a); Hyponymy (vice versa)*
  - *Meronymy (part_of)*
  - *...*
- A **polysemic lexeme** belongs to more synsets

Knowledge-based lexical disambiguation

- **Corpus-based** lexical disambiguation systems
- **Knowledge-based** lexical disambiguation systems

Problem: not always a corpus is available

Aim: to use knowledge to disambiguate anyway

e.g. ***Noun*** *Sense Disambiguation* using mainly:

Conceptual Density and

WordNet sense frequency

Adjective, ***Verb*** *and* ***Adverb*** Sense Disambiguation using: *WordNet Domains*

Knowledge-based lexical disambiguation

1. Select the **nouns** in the context

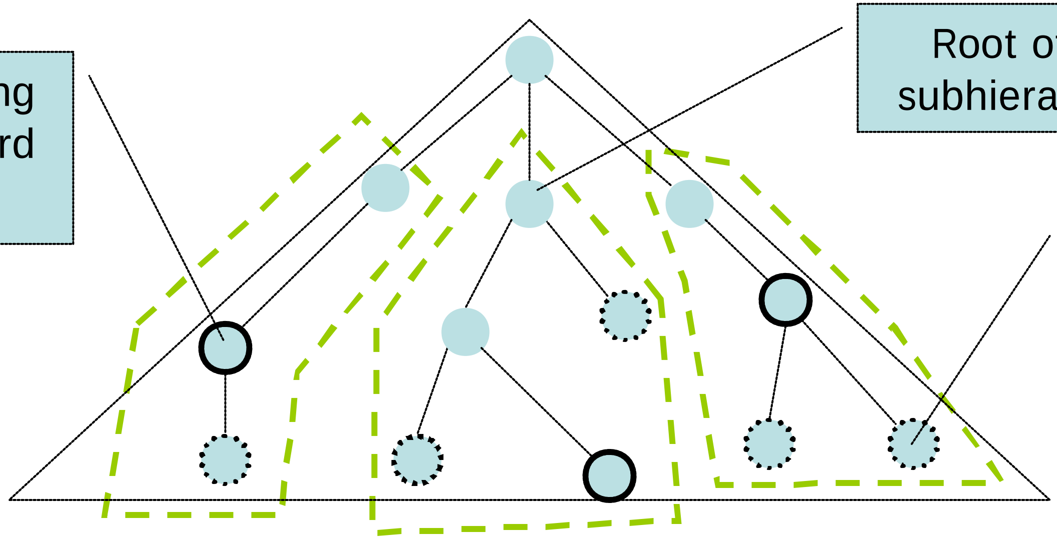
   E.g. *Senseval-3* competition: www.senseval.org

   “**Brakes** howled and a *horn* blared furiously, but the *man* would have been hit if Phil hadn’t called out to him a *second* before”

2. Build subhierarchies
3. Compute densities
4. Assign the sense with highest CD to the noun (when possible)

Knowledge-based lexical disambiguation

1. We start *building subhierarchies* by considering the *word's senses* and the paths connecting those senses to the root synset
2. Then we find the *roots of subhierarchies*: nodes from which only one sense of the word can be reached
3. Finally, we add the *context words*' paths, if they fall within the subhierarchies



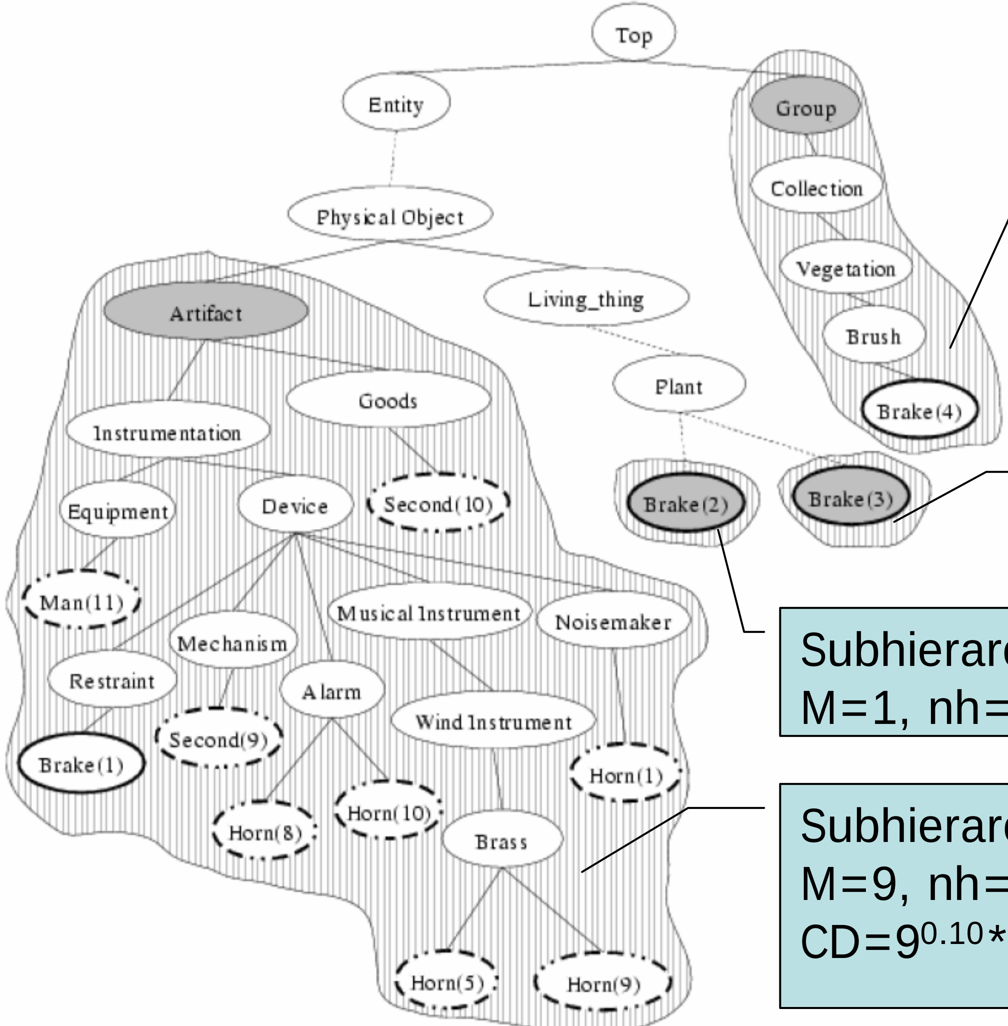
Synset representing a sense of the word to disambiguate

Root of a subhierarchy

Synset representing a sense of a context word

Knowledge-based lexical disambiguation

E.g. *brake* (4 senses) with *context words*: {horn, man, second}



Subhierarchy for sense 4:
M=1, nh=5
CD=$1^{0.10}$*$(1/5)^{\log 4}$=0.07

Subhierarchy for sense 3:
M=1, nh=1, CD=1

Subhierarchy for sense 2:
M=1, nh=1, CD=1

Subhierarchy for sense 1:
M=9, nh=21
CD=$9^{0.10}$*$(9/21)^{\log 1}$=1.25

Knowledge-based lexical disambiguation

- Some results:

precision (nouns):	~82% (SemCor corpus)
	~74% (Senseval corpus)
recall (nouns):	~60% (SemCor corpus)
	~51% (Senseval corpus)
precision (adjectives):	~73% (SemCor corpus)
	~66% (Senseval corpus)
recall (adjectives):	~57% (SemCor corpus)
	~51% (Senseval corpus)

- Integration with corpus-based WSD systems

Research topics

1. Knowledge-based lexical disambiguation
2. Question Answering (QA)
   1. Multilingual QA
   2. Cross-language QA: multi-translator integration
3. **The web as lexical resource**
   1. **Web-based lexical disambiguation**
   2. Web-based lexical pattern extraction
   3. Web-based QA
4. Information Retrieval (IR) and categorization
   1. Semantic (geographical) IR
   2. Semantic text categorization
   3. Transition point indexing reduction technique
5. Text clustering
   1. Clustering of very short narrow-domain texts
   2. Cluster analysis of transcribed spoken dialogues

Web-based lexical disambiguation

- *Problem*: knowledge acquisition bottleneck (sample size is too small) for WSD

- *Aim*: to use web redundancy to disambiguate ***nouns*** using modifier *adjectives* (web hits)

Web-based **algorithm** for adjective-noun lexical patterns

1. Select the adjective *a* before *w*

2. For each w_k , synonym s_{ik} , hypernym (or hyponym) h_{jk} compute: $f_S(a,s_{ik})$ and $f_S(a,h_{jk})$

3. Assign a weight to each w_k (combining the results of 2.) using a given formula *F*

4. Select the w_k with the highest weight

Web-based lexical disambiguation

E.g. Senseval-3:

“A *faint crease* appeared between the man’s eyebrows”

crease$_1$={fold, crease,bend,…}

crease$_2$={wrinkle,crease,line,…}

crease$_3$={kris,crease,creese}

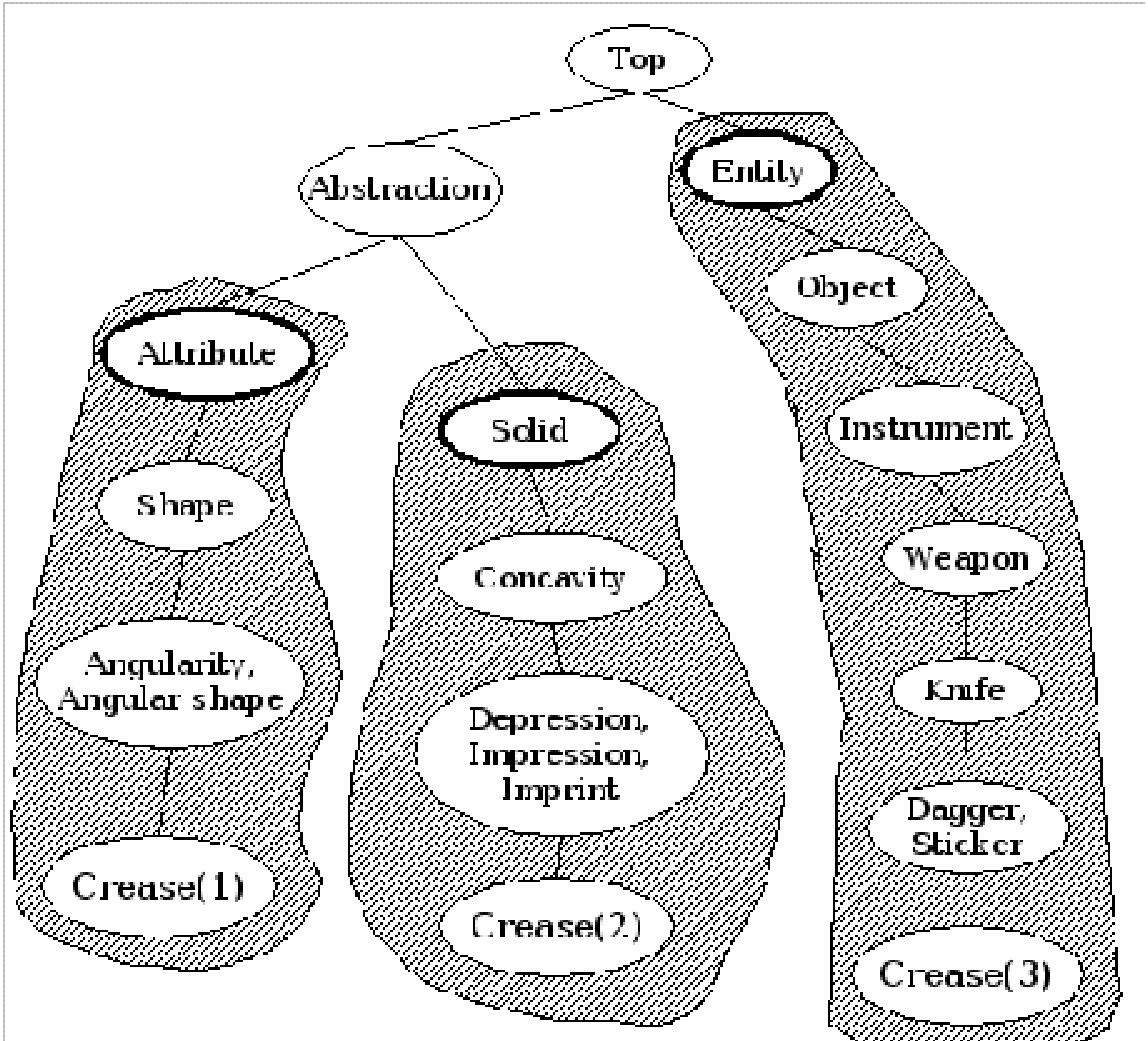
hypernyms:

h$_1$={angular shape,angularity}

h$_2$={depression,impression,imprint}

h$_3$={dagger,sticker}

Web-based lexical disambiguation



Web-based lexical disambiguation

Searching on the Web for the lexical patterns

sense 1:

(faint,fold), (faint,bend), …

(faint, angular shape), (faint,angularity)

sense 2:

(faint,wrinkle), (faint,line), …

(faint, depression), (faint,impression), (faint,imprint)

sense 3:

(faint,kris), (faint,creese)

(faint, dagger), (faint,sticker)

Web-based lexical disambiguation

- Formulae based on:

  weight average

  weight maximum

  similarity measures (mutual information, relative entropy, …

- Some *results*:

  4% gain in recall

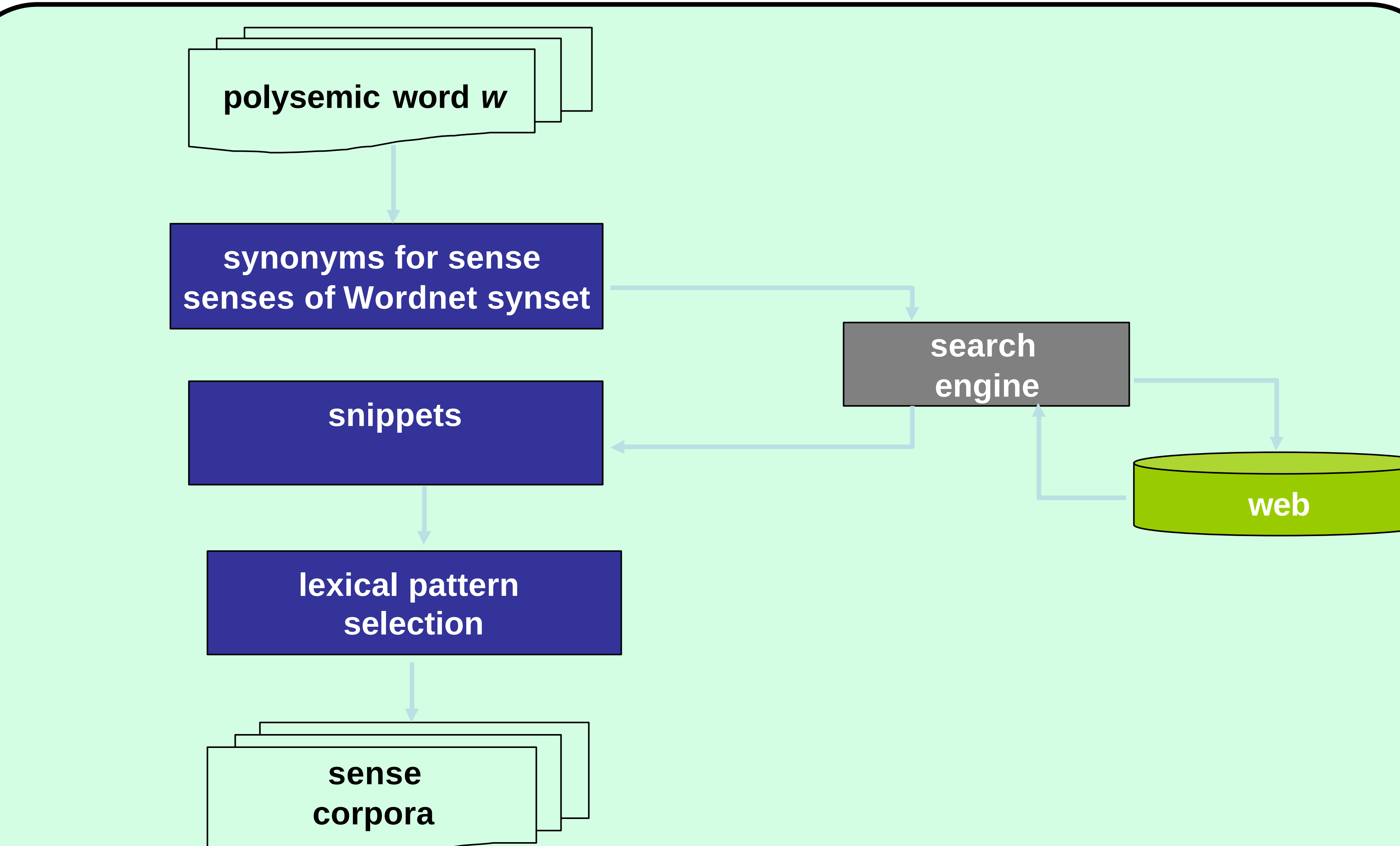
  16% gain in precision (over the words not disambiguated)

Research topics

1. Knowledge-based lexical disambiguation
2. Question Answering (QA)
   1. Multilingual QA
   2. Cross-language QA: multi-translator integration
3. **The web as lexical resource**
   1. Web-based lexical disambiguation
   2. **Web-based lexical pattern extraction**
   3. Web-based QA
4. Information Retrieval (IR) and categorization
   1. Semantic (geographical) IR
   2. Semantic text categorization
   3. Transition point indexing reduction technique
5. Text clustering
   1. Clustering of very short narrow-domain texts
   2. Cluster analysis of transcribed spoken dialogues

Web-based lexical pattern extraction

(mining the web for sense discrimination patterns



Web-based lexical pattern extraction

(mining the web for sense discrimination patterns)

1. **Strength** of the lexical pattern P:

$$S_P = \frac{f_P - f_m}{s}$$

f_P : frequency of *P* in the sense corpus

f_m : average frequency of all lexical patterns in the corpus

s : standard deviation

2. **Internal dispersion** of the lexical pattern *P*:

Does P occur in the **context** of **all** the synonyms of a **sense** of *w*

Sense relevant !

3. **External dispersion** of the lexical pattern *P*:

Does P occur in just **one sense corpus** of *w* ?

Sense relevant !

Research topics

1. Knowledge-based lexical disambiguation
2. **Question Answering (QA)**
   i. Multilingual QA
   ii. Cross-language QA: multi-translator integration
3. The web as lexical resource
   i. Web-based lexical disambiguation
   ii. Web-based lexical pattern extraction
   iii. Web-based QA
4. Information Retrieval (IR) and categorization
   i. Semantic (geographical) IR
   ii. Semantic text categorization
   iii. Transition point indexing reduction technique
5. Text clustering
   i. Clustering of very short narrow-domain texts
   ii. Cluster analysis of transcribed spoken dialogues

Question answering

E.g. CLEF-05 competition: www.clef-campaign.org

"Who is Silvio Berlusconi?"

Possible answers:

Italian Prime Minister

Italian Premier

Business Tycoon

Italy's richest person

Leader of Forza Italia

Milan's president

Mediaset's managing director

... other answer could be added (even if occurring with less redundancy on the web...)

Cross-language QA: multi-translator integration

E.g. CLEF-03 (it-es): “Che cosa significa la sigla CEE?”

(What does the acronym EEC mean?)

Four translators:

1. ¿Qué significa la sigla CEE?
2. ¿Qué cosa significa siglas el EEC?
3. ¿Qué significa la CEE de la abreviación?
4. ¿Qué cosa significa la pone la sigla CEE?

Cross-language QA:
multi-translator integration

1. **Double-Translation** method (it’->es->it”)

Best t: t_i with the greatest similarity (it’, it”)

2. **Word-count** method (exploits the redundancy of terms in all the Ts)

Best t: t_i with the greatest number of words in common

Cross-language QA: multi-translator integration

? **Dice** formula:

$$Sim(t_i, t_j) = \frac{2 * len(t_i \cap t_j)}{len(t_i) + len(t_j)}$$

? **Cosine** formula:

$$f(i, j) \times \log(1 + \frac{n_i}{N})$$

f(i,j)=freq(i,j) / max(freq(i,j))

$$Sim(t_j, t_q) = \frac{(\sum_i t_{ji} * t_{qi})}{\sqrt{\sum_i t_{ji}^2} * \sqrt{\sum_i t_{qi}^2}}$$

T_1: Sim$t_1 t_2$ + Sim$t_1 t_3$ + Sim$t_1 t_4$

T_2: Sim$t_2 t_1$ + Sim$t_2 t_3$ + Sim$t_2 t_4$

...

Cross-language QA: multi-translator integration

	Date	Person	Organization	Location	Measure
WcDice1-G			46%	59%	58%
WcDice2-G					58%
DtDice2-G	61%				
DtDice3-G	61%	64%			
DtCos3-G	61%				
Baseline	70%	64%	42%	72%	40%

Research topics

1. Knowledge-based lexical disambiguation
2. Question Answering (QA)
   i. Multilingual QA
   ii. Cross-language QA: multi-translator integration
3. The web as lexical resource
   i. Web-based lexical disambiguation
   ii. Web-based lexical pattern extraction
   iii. Web-based QA
4. **Information Retrieval (IR) and categorization**
   i. *Semantic (geographical) IR*
   ii. Semantic text categorization
   iii. **Transition point indexing reduction technique**
5. Text clustering
   i. Clustering of very short narrow-domain texts
   ii. Cluster analysis of transcribed spoken dialogues

IR and categorization

Transition point indexing reduction technique

IR system uses a term reduction process based on the **Transition Point technique** (~ **Zip law** of word frequency): *mid term frequency* terms are closely related to the *conceptual content* of a document

$$TP_{SET} = \{t_i | (t_i, f_i) \in V_{TP}, U_1 \leq f_i \leq U_2\}$$

1. $TP = \frac{\sqrt{8 * I_1 + 1} - 1}{2}$ I_1: # words of frequency equal to 1

2. Alternatively, TP = the lowest frequency that is not repeated

IR and categorization

Transition point indexing reduction technique

Corpus		*Size (Kb)*	*% Reduction*	*Mean Reciprocal Rank*	
Full		**117345**	**0%**	**0.0463**	
TP10		**12616**	**89,25%**	**0.0331**	
TP20		**19660**	**83.25%**	**0.0446**	
TP40		**20477**	**82.55%**	**0.0844**	
TP60		**28903**	**75.37%**	**0.0771**	

WebCLEF-05 task results (*TPx*: TP with a *neighbourhood of x%*

Research topics

1. Knowledge-based lexical disambiguation
2. Question Answering (QA)
   1. Multilingual QA
   2. Cross-language QA: multi-translator integration
3. The web as lexical resource
   1. Web-based lexical disambiguation
   2. Web-based lexical pattern extraction
   3. Web-based QA
4. Information Retrieval (IR) and categorization
   1. Semantic (geographical) IR
   2. Semantic text categorization
   3. Transition point indexing reduction technique
5. **Text clustering**
   1. Clustering of very short narrow-domain texts
   2. Cluster analysis of transcribed spoken dialogues

Text clustering
Clustering of very short narrow-domain texts

Problems

1. Organization of text set => Data structuring
2. Searching interesting texts => Clustering based navigation

Typical situation

1. Free access to full-text scientific papers is limited to only their abstracts consisting of no more than several dozens of words

2. Sometimes the set of full-text scientific papers on a given domain are not available is absent at all and a library has only abstracts

Typical opinion

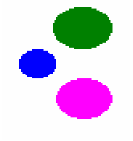
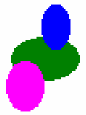
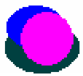
Usual keyword-based methods work well

Text clustering
Clustering of very short narrow-domain texts

Very short texts

1. Texts from different domains

2. Texts from narrow domains

Society	Sciences	Physics
Culture	Physics	Nuclear physics
Economics	Chemistry	Experimental physics
Politics	Biology	Optical physics
………	…………	………
No intersection of vocabularies	Weak intersection of vocabularies	Strong intersection of vocabularies
		

Problem: the stronger the vocabulary intersection is, the more unstable results are

Text clustering
Clustering of very short narrow-domain texts

Very short texts

1. News and other self-contained
2. Abstracts of full scientific texts or technical papers

Abstracts explain the goals of the research reported in the paper the problem), while papers explain the methods used to achieve hese goals (i.e., the algorithms)

Our goal is to shorten the gap between:

1. Automatic abstract clustering vs. manual abstract clustering
2. Automatic abstract clustering vs. manual paper clustering

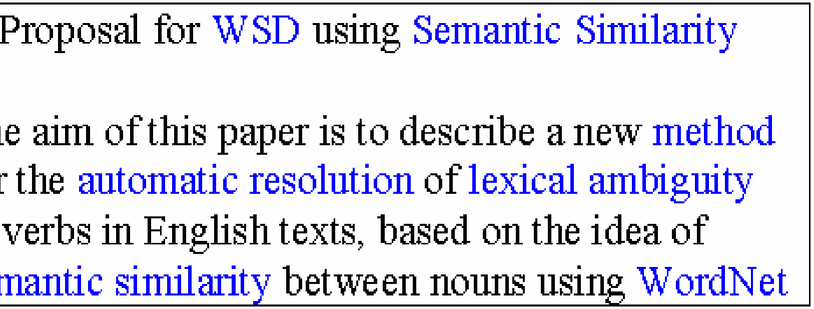
Problem: imprecise results when clustering abstracts

Text clustering
Clustering of very short narrow-domain texts

Very short texts (50-100 words)

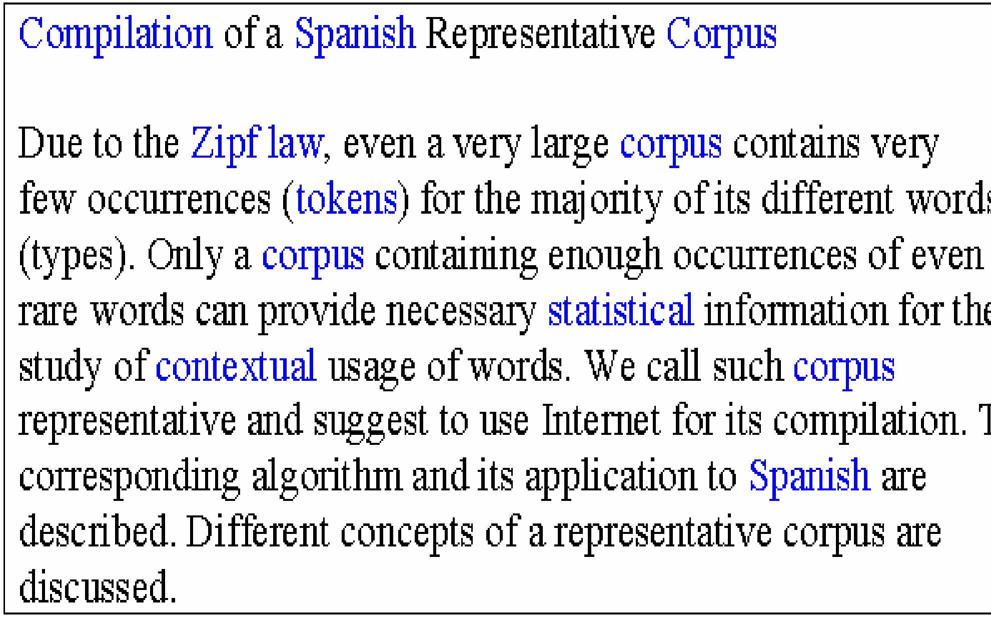
. Absolute frequency of indexes are sometimes 3-4 generally 0-2

. Only 5%-15% of the vocabulary is used in every text



Proposal for WSD using Semantic Similarity

e aim of this paper is to describe a new method r the automatic resolution of lexical ambiguity verbs in English texts, based on the idea of mantic similarity between nouns using WordNet



Compilation of a Spanish Representative Corpus

Due to the Zipf law, even a very large corpus contains very few occurrences (tokens) for the majority of its different words (types). Only a corpus containing enough occurrences of even rare words can provide necessary statistical information for the study of contextual usage of words. We call such corpus representative and suggest to use Internet for its compilation. T corresponding algorithm and its application to Spanish are described. Different concepts of a representative corpus are discussed.

Text clustering
Clustering of very short narrow-domain texts

Traditional approach

1. Constructing word frequency list
   - stop-words are eliminated
   - words having the same base meaning are joined (stemming)

2. Constructing text images according to tf or tf-idf techniques

   $\mathrm{tf}_{\mathrm{i,j}} = f_{i,j} / \max f_{i,j}$ $\mathrm{idf}_{\mathrm{i}} = \mathrm{Log}\,(N/n_i)$ *i*-th word, *j*-th text

3. Clustering using the cosine measure

From (2) : high randomness in text images

Results: not such a big problem when texts are from different domains, but when they are narrow domain…

Text clustering
Clustering of very short narrow-domain texts

Struggling for stability

Using compensative effect

To join indexes (keywords) :

$(w_1 , w_2 , ...w_n)$ => $W_1 = (w_1, w_3, w_{19})$, $W_2 = (w_7, w_{13}, w_{23})$,..

To cluster abstracts in new index space (cluster coordinates):

$(W_1 , W_2 ,....)$

Selection of group of indexes

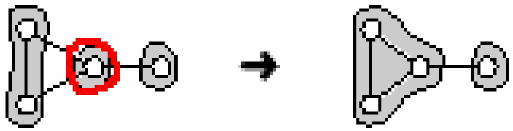
1. Use synsets of an appropriate ontology
2. Use a thesaurus of a given domain
3. Cluster the words in the space of texts <= *our approach* (MajorClust algorithm)

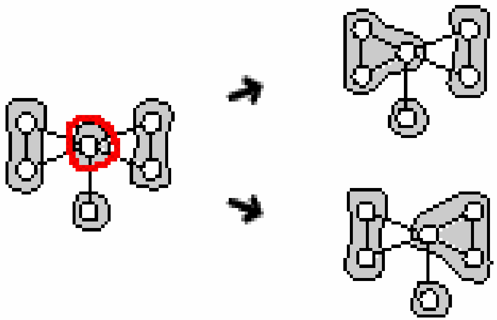
Weighting indexes $W_k = ?\ d_{i,j} / N_k$,

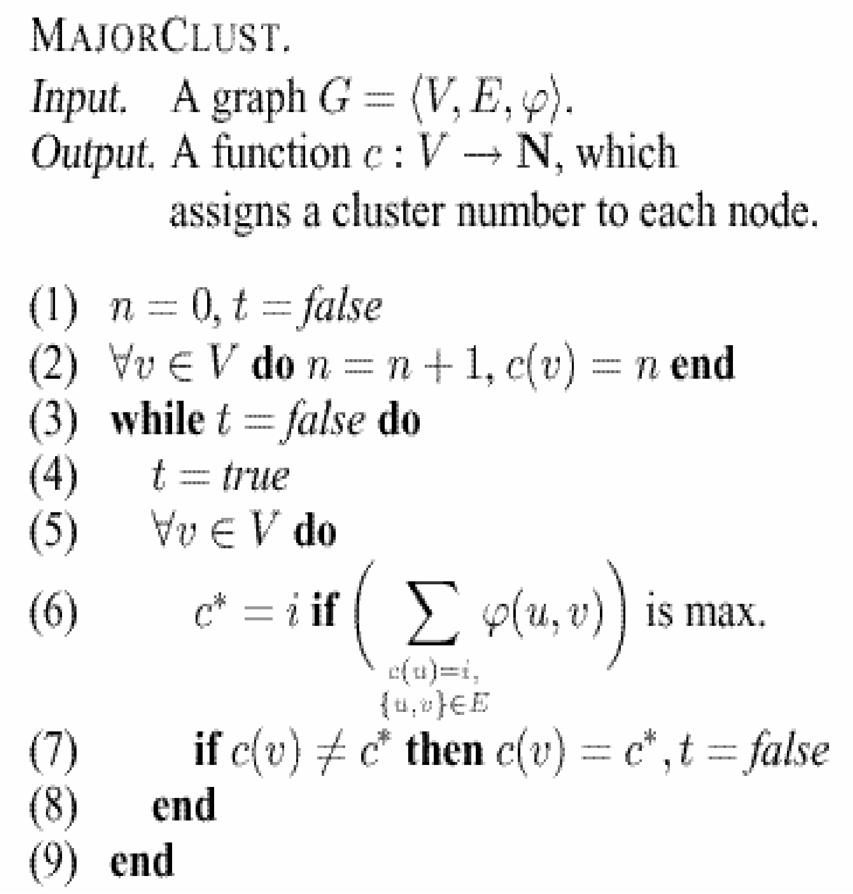
k is the number of the cluster, *i* and *j* are the elements of this clusters (*i* ? *j*),

N_k is the number of links in the cluster *k*.

Text clustering
Clustering of very short narrow-domain texts



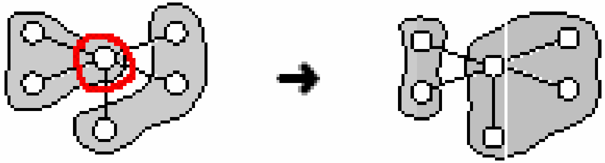




MAJORCLUST.

Input. A graph $G = \langle V, E, \varphi \rangle$.
Output. A function $c : V \to \mathbf{N}$, which
assigns a cluster number to each node.

```
$n = 0, t = false$
$\forall v \in V$ do $n = n + 1, c(v) = n$ end
while $t = false$ do
    $t = true$
    $\forall v \in V$ do
        $c^* = i$ if $\left( \sum_{c(u)=i, \{u,v\} \in E} \varphi(u,v) \right)$ is max.
        if $c(v) \neq c^*$ then $c(v) = c^*, t = false$
    end
end
```



An object belongs to the **cluster whom the majority of its neighbours** belong to

Sub-optimal solution:

only a limited part of neighbours is considered

Text clustering
Clustering of very short narrow-domain texts

Struggling for precision

Using a more adequate measure

We use cosine measure

$$C_{1,2} = \frac{\sum_{k} \left(x_{k1}, x_{k2} \right)}{\left\| x_1 \right\| \left\| x_2 \right\|},$$

where *1, 2* are the numbers of texts

$x_{k,i}$ are the cluster coordinates

Coordinate transformation:

y_{ki} = log (1 + $f_{k,i}$) / log (1 + max(f_i))

Aim: smoothing of high frequencies typical abstract words

e.g. method, experiment, result etc.)

Text clustering
Clustering of very short narrow-domain texts

Clustering indexes

MajorClust method:

number of clusters is defined automatically

Clustering abstracts

NN method (hierarchy-based)

K-means method (example-based)

MajorClust method (density-based)

Abstracts (preliminary results)

Using compensative effect improves results

Using logarithmic measure improves results

Text clustering
Clustering of very short narrow-domain texts

Experiments: Clustering abstracts CICLing-2002

Indexing: 390 keywords

Gold standard: 4 clusters (obtained also with MajorCluster):

Linguistic (semantics, syntax, morphology, parsing)

Ambiguity (word sense disambiguation, anaphora, tagging, spelling)

Lexicon (lexicon and corpus, text generation)

Text processing (information retrieval, summarization, text classification)

Narrow domain: e.g. V_2 n V_4 = 70%

Indexing	**log Scaling**	**F-measure**
tf-idf	No	0.64
tf	No	0.57
Grouping	**Yes**	**0.78**
Grouping	No	0.68

Text clustering
Clustering of very short narrow-domain texts

Digital library and Internet repositories should provide
open access both to abstracts and to document images
of full papers: this does not violate the copyright of authors!

Proposal by Dr. Pavel Makagonov
Mixteca University of Technology, Mexico

Text clustering
Cluster analysis of transcribed spoken dialogues

Spanish Railway Service

Goal: Designing automatic dialogues systems

Problem: Revealing the typical scenarios of dialog

Data: 100 real dialogues

Difficulties:

Information is fuzzy

Information is absent

Information is in a hidden form

***DI*: Renfe customer service, good morni**

***US*: Good morning**

***DI*: May I help you?**

***US*: Yes, please: I would like to know abo**
a train from Valencia to Barcelona.

***DI*: What day are you interested in?**

***US*: Next Thursday, in the afternoon.**

***DI*: Let’s see. <PAUSE> On Thursday**
there is an EuroMed leaving at 3 P.M
and arriving in Barcelona at 6.45 P.M

***US*: What about the next train?**

***DI*: It leaves at 8 P.M.**

***US*: Too late. Thank you. Bye.**

***US* = User**
***DI* = Directory Inquire Service**
Length = 25% like this

Text clustering
Cluster analysis of transcribed spoken dialogues

Spanish Railway Service

Usual solution:

Manual evaluation
of person-to person dialogs
based on lexical analysis

Example of solution:

Hour of departure, discounts
Hour of departure, price
Return ticket
Type of train

Additional results of lexical analysis:

Why citizens of Tarragona like to travel on Sunday?
Why citizens of Madrid like to ask for discounts?

Text clustering
Cluster analysis of transcribed spoken dialogues

Type of parameters

Reflecting transport service

Reflecting passenger behaviour

List of parameters

Town importance	0, 0.25,…1
Urgency	0, 0.5, 1
Return ticket	1/0
Time of departure	
Time of departure (return)	
Wagon-lit	1/0
Discounts	1/0
Length of talking	0, 0.25,….1
Politeness	0, 0.25, 1
…	

Difficulties

Information is fuzzy

Information is absent

Information is in a hidden form

Nominal scales

Time of departure:

Indifference	1/0
Morning or day	1/0
Evening or night	1/0

=> [(1,0,0) , (0,1,0), (0,0,1)]

Presumption

For absent parameters
it is used:

- a value of indifference
- the cheapest and simplest

Text clustering
Cluster analysis of transcribed spoken dialogues

Problems

nfluence of dominant parameters => real structure will be hidden

nfluence of noise => real structure will be disfigured

arameter analysis => Filtering parameters:

roups of parameters

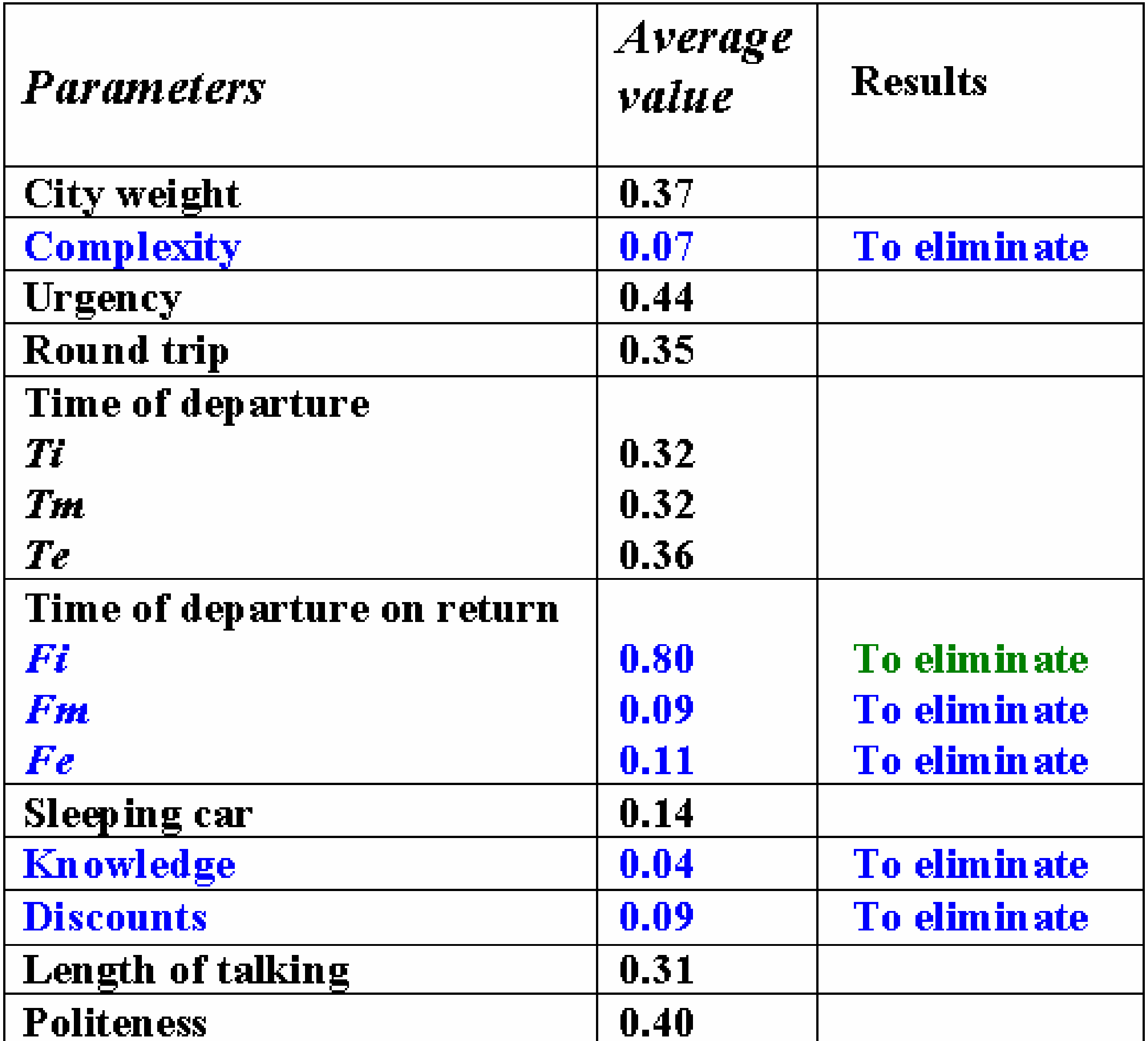
1. Significant value for 90%-95% of objects
2. Significant value for 5%-10% of objects

=> 3. Significant value for more ~ 20%-30% of objects

ole of parameters:

1. First group parameters are oriented to uniform object set: eliminate
2. Second group parameters oriented to very granulated object set (in subsystems): eliminated

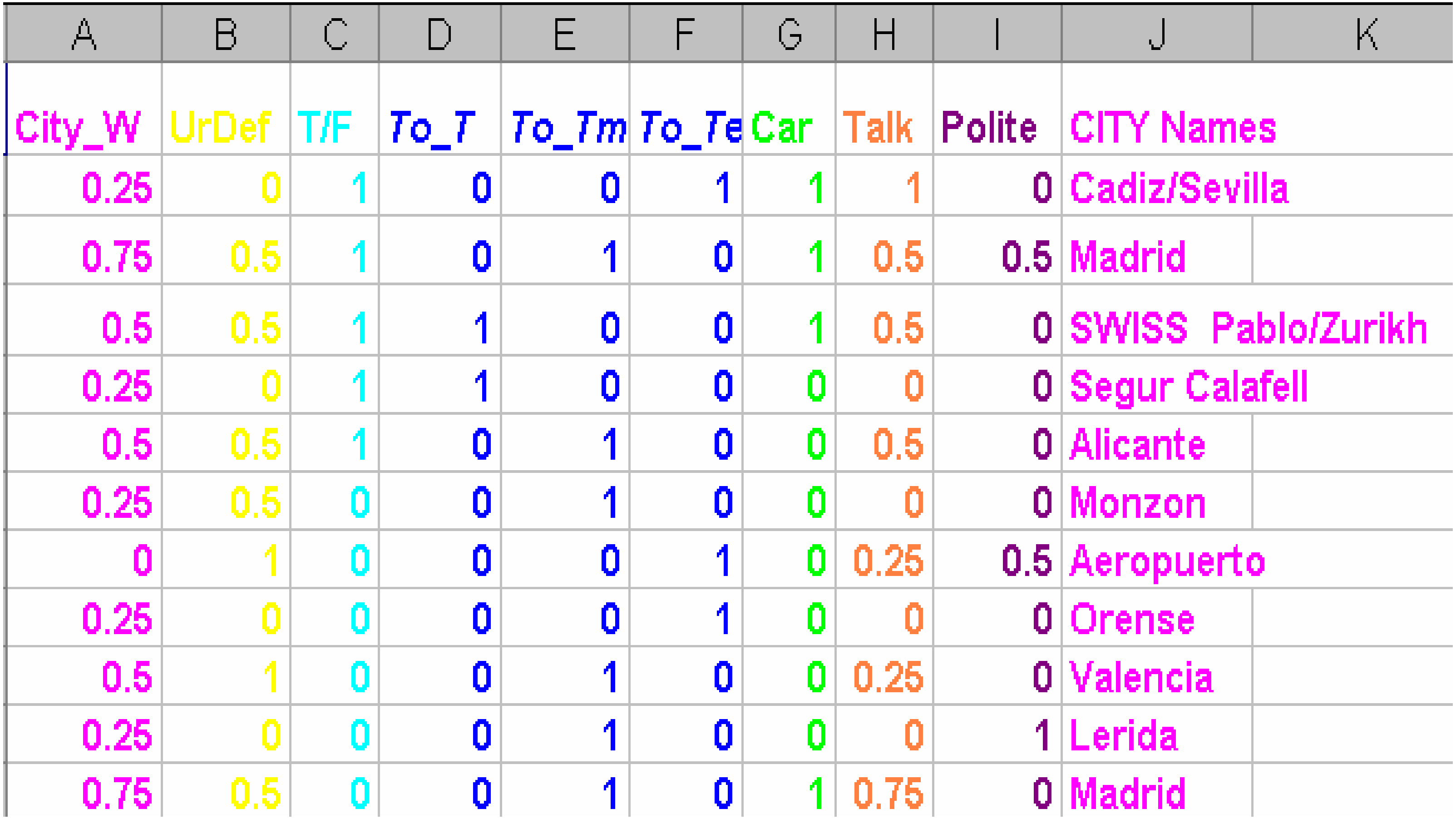
Text clustering
Cluster analysis of transcribed spoken dialogues



Parameters	*Average value*	Results
City weight	0.37	
Complexity	0.07	To eliminate
Urgency	0.44	
Round trip	0.35	
Time of departure		
Ti	0.32	
Tm	0.32	
Te	0.36	
Time of departure on return		
Fi	0.80	To eliminate
Fm	0.09	To eliminate
Fe	0.11	To eliminate
Sleeping car	0.14	
Knowledge	0.04	To eliminate
Discounts	0.09	To eliminate
Length of talking	0.31	
Politeness	0.40	

Text clustering

Cluster analysis of transcribed spoken dialogues



A	B	C	D	E	F	G	H	I	J	K
City_W	UrDef	T/F	To_T	To_Tm	To_Te	Car	Talk	Polite	CITY Names	
0.25	0	1	0	0	1	1	1	0	Cadiz/Sevilla	
0.75	0.5	1	0	1	0	1	0.5	0.5	Madrid	
0.5	0.5	1	1	0	0	1	0.5	0	SWISS Pablo/Zurikh	
0.25	0	1	1	0	0	0	0	0	Segur Calafell	
0.5	0.5	1	0	1	0	0	0.5	0	Alicante	
0.25	0.5	0	0	1	0	0	0	0	Monzon	
0	1	0	0	0	1	0	0.25	0.5	Aeropuerto	
0.25	0	0	0	0	1	0	0	0	Orense	
0.5	1	0	0	1	0	0	0.25	0	Valencia	
0.25	0	0	0	1	0	0	0	1	Lerida	
0.75	0.5	0	0	1	0	1	0.75	0	Madrid	

Objects/Attributes matrix

Clustering method

NN method

K-means method

MajorClust metho

Text clustering
Cluster analysis of transcribed spoken dialogues

Some conclusions:

1. Scenarios of dialogues may be determined by clustering them in the space of parameters defined by an expert
2. Importance of how to parameterize dialogues in orde to compensate incompleteness and fuzziness of source information
3. Procedure of weighting dialogues and parameters allows to obtain information useful for a user
4. The MajorClust method seems to be the one for solving this kind of problems

1. Knowledge-based lexical disambiguation
2. Question Answering (QA)
3. The web as lexical resource
4. Information Retrieval (IR) and categorization
5. Text clustering
6. **Thanks**