Web mining for Natural Language Engineering tasks

Pattern Recognition & Artificial Intelligence group

Dept. of Computation and Information System

Polytechnic University of Valencia, Spain

Paolo Rosso

prosso@dsic.upv.es

1. Web-based Word Sense Disambiguation
2. Web-based Passage Retrieval
3. Web-based Translation

1. Web-based Word Sense Disambiguation

Davide Buscaldi, Ph.D. student, RFIA, UPV

2. Web-based Passage Retrieval

3. Web-based Translation

---

1. Web-based Word Sense Disambiguation

- Problem: WSD and the knowledge acquisition bottleneck (sample size is too small)
- Aim: to use the Web as a lexical resource for the disambiguation of nouns using modifier adjectives (Web hits)
- Background:
  - study of noun-verb relationships (Mihalcea,99)
  - study of the redundancy of the Web (Brill,03): "data, data, more data"

Preliminary definitions

- *w* word (noun)
- |*w*| senses
- *a* adjective
- *n* synonyms s_{ik} of w_k
- *m* words h_{jk} in the direct hypernym synset of w_k
- $f_S(x,y)$: function returning the # of pages containing "*x y*" (according to the search engine *S*)
- $f_S(x)$: function returning the # of pages containing *x*

1. Web-based Word Sense Disambiguation

Web-based algorithm

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

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

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

d. Select the w_k with the highest weight

1. Web-based Word Sense Disambiguation

Senseval-3 English all-word corpus:
e.g. 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}

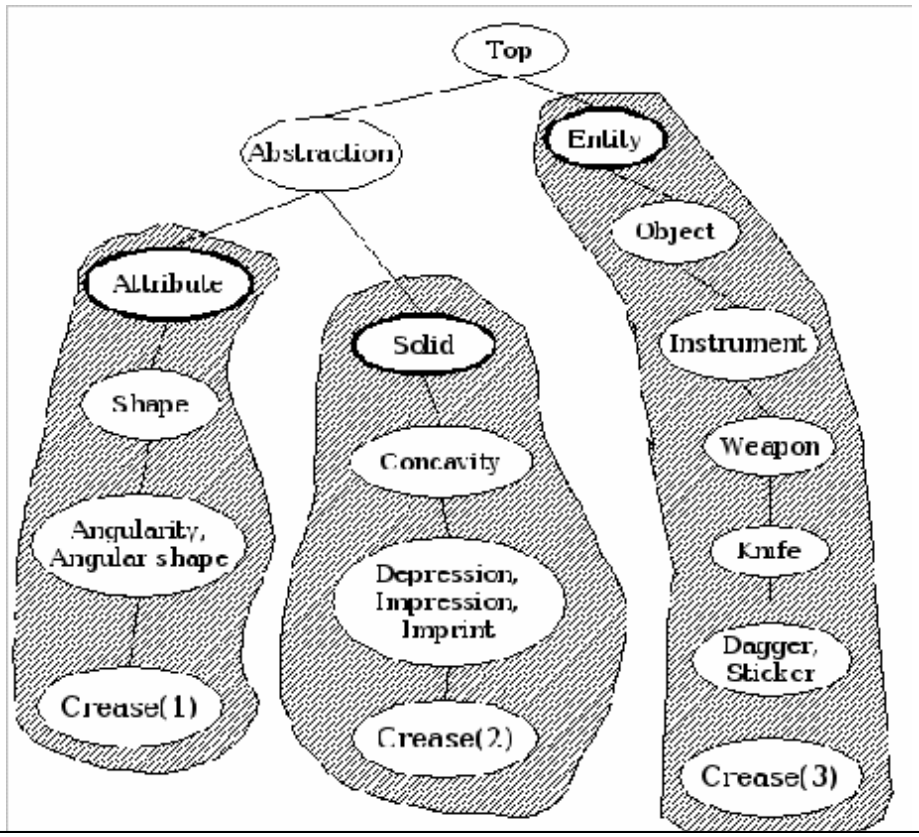
1. Web-based Word Sense Disambiguation

searching the Web for the patterns:

(faint, *x*) where *x* is a word in the synset of the related sense or the hyperonyms of that sense

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)

1. Web-based Word Sense Disambiguation



1. Web-based Word Sense Disambiguation

Some of the tested formulae

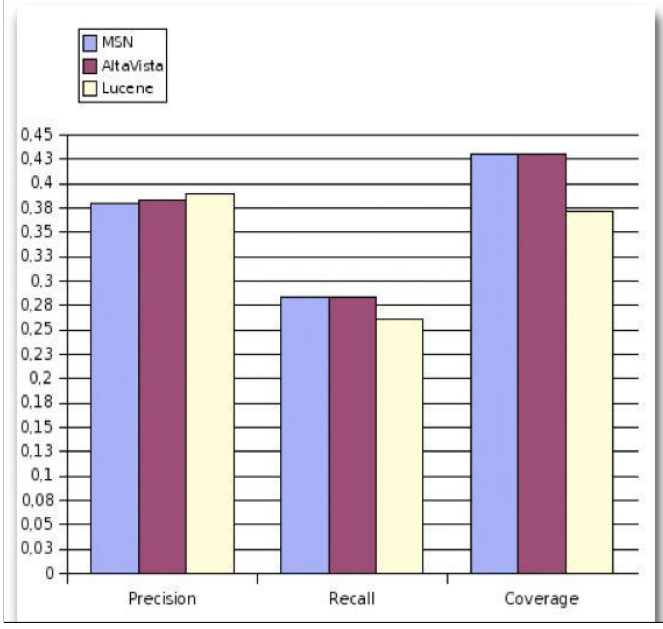
- Based on weight average:
  - F_I: $1/2*(\sum f_S(a,s_{ik}) / n + \sum f_S(a,h_{jk}) / m)$
  - F_{II}: F_I with hyponyms
- Based on weight maximum:
  - F_{III}: $\max (f_S(a,s_{ik}) , f_S(a,h_{jk}))$
- Based on similarity measures:
  - F_{IV}: $\max (f_S(a,s_{ik}) \log (f_S(a,s_{ik}) / f_S(s_{ik})), f_S(a,h_{jk}) \log (f_S(a,h_{jk}) / f_S(h_{jk})))$

1. Web-based Word Sense Disambiguation

Formula	Precision	Recall	Coverage	$Prec_{na}$
MFS	68.9	68.9	100	62.3
I	62.7	27.1	43.3	31.8
II	66.1	28.6	43.3	39.2
III	66.0	27.8	42.0	37.3
IV	**77.7**	33.7	43.3	63.4

1. Web-based Word Sense Disambiguation

Search Engines comparison



- MSN, AltaVista + Lucene with the TREC-8 Document Collection
- no significative differences were detected between the Web SE
- Lucene (offline):
  - + precision, - coverage
  - less data but of better quality

Remarks

- Frequency (calculated over SemCor) corrected formulae gives better results in precision (and recall : 4% gain)
- Importance of the adjectives' polysemy for nouns sense disambiguation: the less polysemic is the adjective, the higher is the probability of selecting the right sense
- Only one adjective is not enough to understand the meaning of a noun (e.g. pair still ambiguous: *cold fire ~ cold passion*): a greater context should be taken into account
- The same approach for the disambiguation of adjectives (i.e., searching for $f_S(a_{ik}, w)$) obtained a poor precision: 21.3%

1. Web-based Word Sense Disambiguation

Conclusions

- Lower quality of the Web as lexical resource vs. a large static data set
- Better to integrate Web-based approaches and not to use them standalone

1. Web-based Word Sense Disambiguation

2. Web-based Passage Retrieval

José Manuel Gómez, Ph.D. student, RFIA, UPV

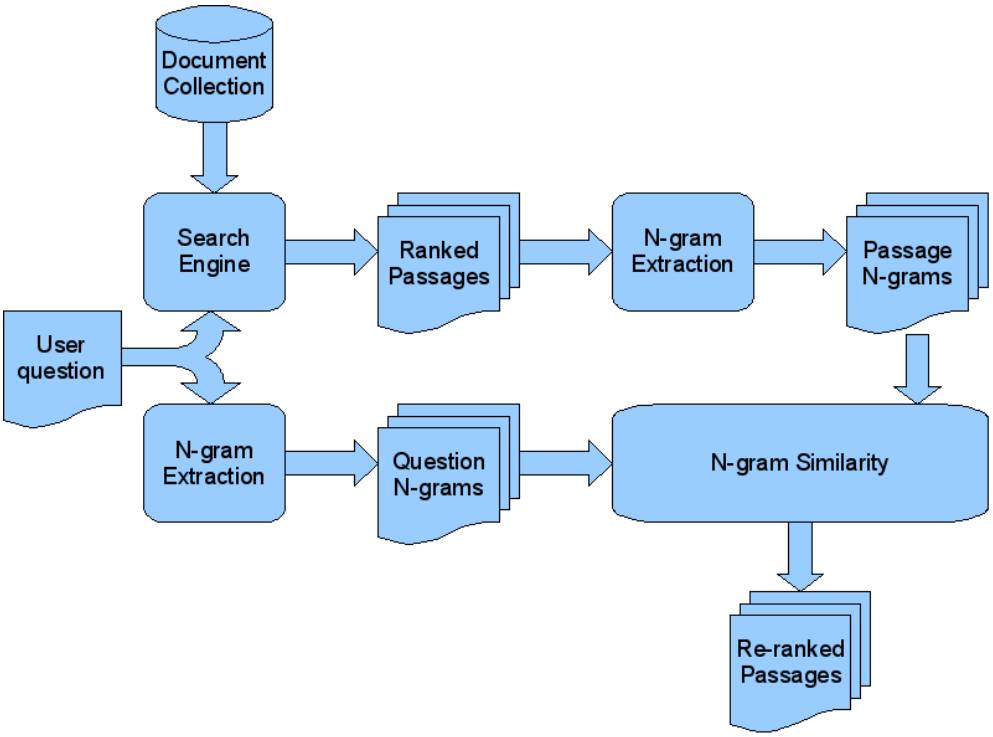
3. Web-based Translation

---

2. Web-based Passage Retrieval

- Problem: given a question, to retrieve the relevant passages containing the correct answer
- Aim: to use the Web as a lexical resource to retrieve the snippets containing the correct answer

2. Web-based Passage Retrieval



JIRS (Java Information Retrieval System)
language-independent PR: http://jirs.dsic.upv.es

2. Web-based Passage Retrieval

Query N-Grams

e.g. ¿Quién es la viuda de John Lennon?

(Who is the widow of John Lennon?)

es la viuda de John Lennon	1x6-gram
es la viuda de John, la viuda de John Lennon	2x5-gram
es la viuda de, la viuda de John, viuda de John Lennon	3x4-gram
es la viuda, la viuda de, viuda de John, de John Lennon	4x3-gram
es la, la viuda, viuda de, de John, John Lennon	5x2-gram
es, la, viuda, de, John, Lennon	6x1-gram

2. Web-based Passage Retrieval

Passage N-Grams

Passage 1

... 4, 2004.- La viuda de John Lennon, Yoko Ono, se ha movilizado ... .

La viuda de John Lennon 1 x 5-gram

Passage 2

... inicio de la carrera de John fuera de los Beatles ... musicales que incluían a John Lennon, Eric Clapton, Keith Richards ... .

John Lennon 2 x 2-gram
de la

Both passages contain the most relevant words (John and Lennon) but the passage 1 has one 5-gram whereas the passage 2 has two 2-grams

2. Web-based Passage Retrieval

Term weight

$$w_k = 1 - \frac{\ln(n_k)}{1 + \ln(N)}$$

n_k: # of passages in which t_k occurs

N: the total # of passages

2. Web-based Passage Retrieval

Density Distance N-Gram Model

Passage 1

... 4, 2004.- La viuda de John Lennon, Yoko Ono, se ha movilizado ... .

$$h(x,P_j)=\begin{cases}\sum_{k=1}^{|x|} w_k\, d(x,x_{\max}) & if\ x\in P_j\\ 0 & otherwise\end{cases}$$

$$d(x,x_{\max})=\frac{1}{1+\ln(1+L)}$$

Passage 2

... inicio de la carrera de John fuera de los Beatles ... musicales que incluían a John Lennon, Eric Clapton, Keith Richards ... .

L (Length): # of words between *x* and *xmax* ngrams

2. Web-based Passage Retrieval

Density Distance N-Gram Model

	0.061	0.061	0.389	0.061	0.359	0.395
¿Quién	**es**	**la**	**viuda**	**de**	**John**	**Lennon?**
es la viuda de John Lennon				1.326		

Passage 1

... 4, 2004.- La viuda de John Lennon, Yoko Ono, se ha movilizado ... .

Passage 2

... inicio de la carrera de John fuera de los Beatles ... musicales que incluían a John Lennon, Eric Clapton, Keith Richards ... .

Passage 1

La viuda de John Lennon + 1.265 **1.265** } **0.954**

Passage 2

$d(x,x_{max})$

de la 0.122 x 0.287

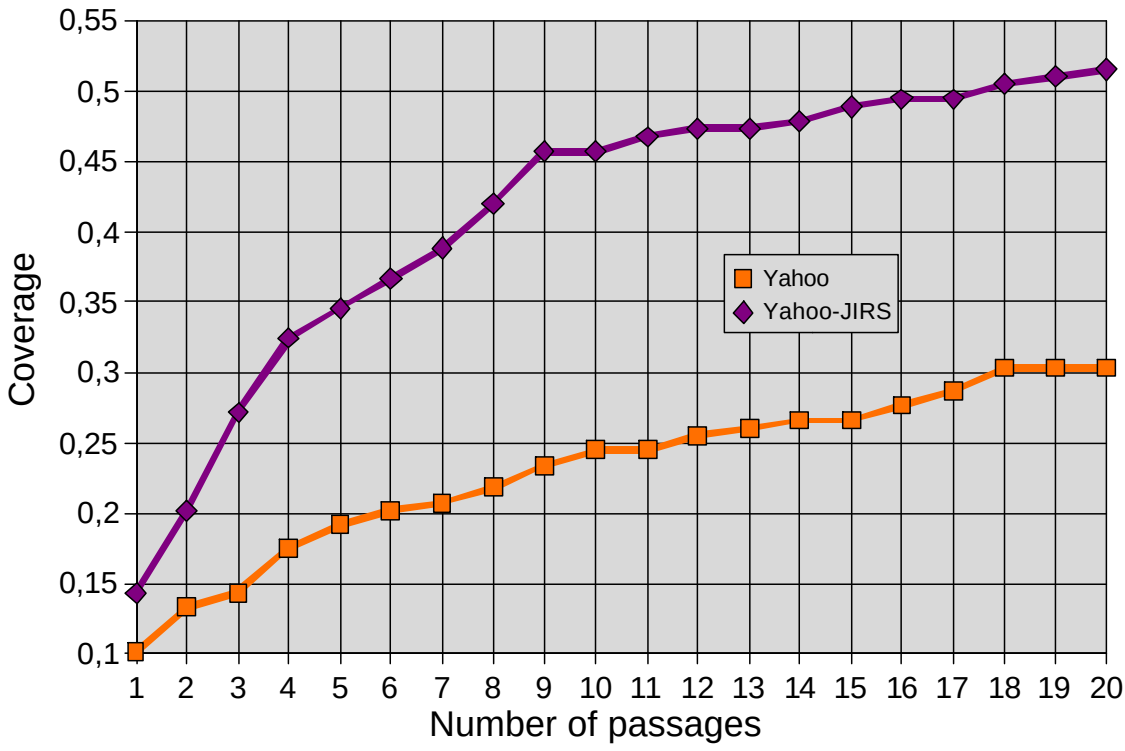
John Lennon + 0,754 **0.789** } **0.595**

2. Web-based Passage Retrieval

Español -> Español	
inao051eses	**42,00%**
tova051eses	**41,00%**
upv051eses	**33,50%**
alia051eses	33,00%
aliv051eses	32,50%
talp051eses	29,00%
mira051eses	25,50%

QA JIRS-based systems @ Spanish CLEF-2005 track
http://clef-qa.itc.it/2005/

2. Web-based Passage Retrieval



Preliminary results on the Web: Yahoo vs. Yahoo-JIRS (CLEF-2005 QA Spanish track)

2. Web-based Passage Retrieval

Remarks and -> Further work

- Snippets are too short, cutting the sentence or not including the previous or following sentences
  -> Search in URLs
- Definition questions
  -> Query reformulation
- Long questions (more difficult to find the correct answer)
  -> Search more than 1000 snippets
  -> Use synonyms
  -> Question feedback
- Weight term function (to be improved)

2. Web-based Passage Retrieval

Conclusions

- Yahoo works very well with current topics

but maybe not so well with topics about history
(some of CLEF questions are about history)

- Yahoo does not return so good passages for JIRS
  – In the first 1000 snippets returned by Yahoo only 70% of answers was found
  – The OR operator

1. Web-based Word Sense Disambiguation

2. Web-based Passage Retrieval

3. Web-based Translation

Davide Buscaldi, Ph.D. student, RFIA, UPV
Matteo Iskra, student of DISI, University of Genova, Italy

---

3. Web-based Translation

- Problem: Cross-Language Question Answering task (better not relying only on just 1 on-line translator…)
- Aim:
  - to use n on-line translators
  - to use the Web as a lexical resource: Web hits of pages with bigram and trigram of each translation
  - to weight each translation (using the Shannon's entropy formula)

3. Web-based Translation

Machine Translators:

- FreeTranslation
- Promt
- Reverso
- Systran e.g.:

ENG What is the name of the European Aeronautical Consortium?

ESP ¿Cuál es el nombre del consorcio aeronáutico europeo?

¿Cuál es el nombre del consorcio aeronáutico europeo?	Yahoo	Google
¿**Cuál es el** nombre del consorcio aeronáutico europeo?	**643**	**5800000**
¿Cuál **es el nombre** del consorcio aeronáutico europeo?	**700000**	**1910000**
¿Cuál es **el nombre del** consorcio aeronáutico europeo?	**2380000**	**8430000**
¿Cuál es el **nombre del consorcio** aeronáutico europeo?	**1370**	**66400**
¿Cuál es el nombre **del consorcio aeronáutico** europeo?	**0**	**396**
¿Cuál es el nombre del **consorcio aeronáutico europeo**?	**1**	**439**

3. Web-based Translation

Original Shannon's entropy formula:

$$H(X) = -K\sum_{i=0}^{n}(p(i)\log p(i)$$

Adapted formula:

$$H(X) = -\frac{1}{n}\sum_{i=0}^{n}\frac{\log c(t_i)}{\log c(b_i)}(\log c(t_i) - \log c(bi))$$

t="w1 w2 w3" # web pages where trigram occurs

b="w1 w2" # web pages where bigram occurs

3. Web-based Translation

Machine Translator	What is the name of the European Aeronautical Consortium?	H(X) Yahoo	H(X) Google
Systran	¿Cuál es el nombre del consorcio aeronáutico europeo?	10.023547	12.538347
Promt	¿Cuál es el nombre del Consorcio Aeronáutico europeo ?	7.310060	8.339859
Reverso	¿Cuál es el nombre del Consorcio europeo Aeronáutico ?	9.995533	13.575086
Free Translator	¿Cómo se llama el Consorcio Aeronáutica europeo?	13.198722	15.172317

weights obtained by the web search engines for each translation
in case of equal weight : the *Prompt* baseline translation is selected

3. Web-based Translation

Evaluation:

- CLEF-2005 English-Spanish cross-language task (200 questions)
- BLEU evaluation model (NIST)

http://www.nist.gov/speech/tests/mt/resources/scoring.htm

3. Web-based Translation

Machine translator	BLEU score
Promt (baseline)	0.4139
Reverso	0.3930
FreeTranslator	0.3686
Systran	0.3462

3. Web-based Translation

Comparison of the Web vs. a static collection (CLEF)

Method	BLEU score
best Yahoo	0.3610
best Google	0.3696
best Lucene	0.3736
"ideal" baseline (manual)	0.4622

- Lucene: free open-source (Java) Information Retrieval System http://lucene.apache.org/
- "ideal baseline": obtained by manually selecting the best translation

3. Web-based Translation

Remarks:

- Simple questions (many of the CLEF questions are short): good translation for all translators
- Not always the Shannon's entropy formula allows to select the best translation
- Some errors are the same for all the translators
- Parts of grams may be written in another language, in between punctuation signs or written with some accent or capital letter)

3. Web-based Translation

Conclusions and further work:

- Not lower quality of the Web as lexical resource vs. a large static data set
- To take into account the possibility to use multigrams
- To be aware of the commonest translation mistakes
- To substitute names, acronyms, abbreviations with wildcards and consider separately the problem of Named Entity translation

(e.g. Who is Silvio Berlusconi? -> Who is #?)

1. Web-based Word Sense Disambiguation

2. Web-based Passage Retrieval

3. Web-based Translation

4. **Web-based (sense discrimination) lexical pattern extraction**

Rafael Guzmán, Ph.D. student, RFIA, UPV
+ Manuel Montes, INAOE-Puebla, Mexico

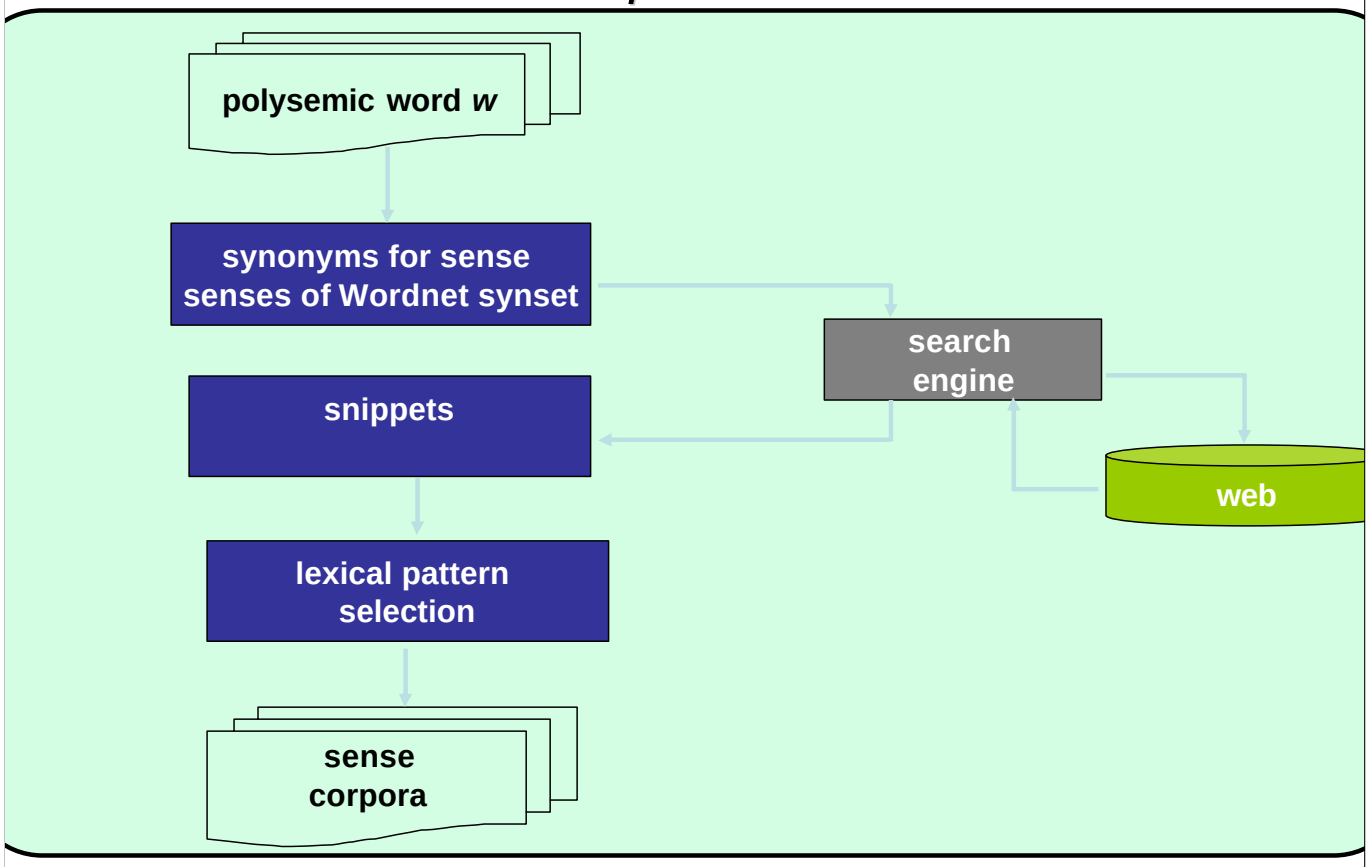
---

4. Web-based (sense discrimination) lexical pattern extraction

Aim:

- to use the Web for enriching a given corpus with lexical patterns
- To use the obtained corpus for training in tasks:
  - Text Categorization (e.g. natural disaster corpus)
  - Named Entity Recognition
  - Word Sense Disambiguation

4. Web-based (sense discrimination) lexical pattern extraction



"under construction"…

5. Web-based…

…Thanks

Gràcies

Gracias

Grazie