

# Research topics



German Rigau i Claramunt  
<http://adimen.si.ehu.es/~rigau>

IXA group

Departamento de Lenguajes y Sistemas Informáticos  
UPV/EHU

# Research Topics

- RT1: Domain tagging
- RT2: Domain tagging of WordNet
- RT3: Train-O-Matic
- RT4: VisualGenome
- RT5: New KnowNets
- RT6: Synset embeddings
- RT7: Mapping new multilingual words to synsets
- ...

# RT1: Domain Tagging

Mariano Rajoy announced yesterday in Madrid that the budget cuts will continue next year.

budget\_cuts => ECONOMY? POLITICS?

Mariano\_Rajoy => ECONOMY? POLITICS?

Madrid => ECONOMY? POLITICS? ...

- Where is that knowledge?
- How to use that knowledge to perform the task?
- How to evaluate this task?

# RT1: Domain Tagging

- Where is that knowledge?
  - WN Domains :
    - <http://wndomains.fbk.eu/> (WN1.6)
    - <http://adimen.si.ehu.es/web/MCR> (WN3.0)
  - BabelDomains :
    - <http://lcl.uniroma1.it/babeldomains/> (Babelsynsets => synsets)
- How to use that knowledge to perform the task?
  - Lemmatization and POS tagging (ixa-pipes-tok and ixa-pipes-pos)
  - Personalized PageRank Vector (PPV) (UKB)
  - Generate PPV per sentence
  - Select the top K synsets and its associated Domains
- How to evaluate this task?
  - <http://adimen.si.ehu.es/web/WSD-WN-Glosses>
  - SemEval 2013: <https://www.cs.york.ac.uk/semeval-2013/task12.html>
  - SemEval 2015: <http://alt.qcri.org/semeval2015/task13/>
  - Reuters : <http://www.daviddlewis.com/resources/testcollections/rcv1/>
  - ...

# RT2: Domain Tagging of WordNet

hospital#n#1 : health facility where patients receive treatment

hospital => MEDICINE?

health\_facility => MEDICINE?

patients => MEDICINE?

treatment => MEDICINE?

- Where is that knowledge?
- How to use that knowledge to perform the task?
- How to evaluate this task?

# RT3 Train-O-Matic

- Reimplement PPV WSD using UKB
  - <http://ixa2.si.ehu.es/ukb>
  - <http://trainomatic.org/>
  - <http://www.aclweb.org/anthology/D17-1008> (sec. 2.1 and 2.2)
- Evaluation:
  - Datasets, metrics, etc.
  - SensEval/SemEval : <https://en.wikipedia.org/wiki/SemEval>
  - <http://lcl.uniroma1.it/wsdeval>
  - ...

# RT4 VisualGenome

- <http://visualgenome.org/>
- Create a Gold Standard dataset
  - Ambiguous contexts per picture/region
  - we already know the answers ...
- Test different WSD methods
  - UKB <http://ixa2.si.ehu.es/ukb/> (using WN+Gloss)
- Use the new knowledge to enrich the graph
  - Evaluate using SenseEval/SemEval
  - ...

# RT5 New KnowNets

- Current KnowNets:
- <http://adimen.si.ehu.es/web/KnowNet>
  - Topic Signatures (word vectors)
  - Knowledge-based WSD (SSI-Dijkstra+)
  - => KnowNets
  - Evaluation: lexical sample WSD

# RT5 New KnowNets

Topic Signatures - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Reload Home Search Favorites

Address C:\Documents and Settings\agirre\Desktop\Topic Signatures.htm Go Links

Topic Signatures Browser (all WN 1.6 polysemous nouns)

Type any noun:

horse (definitions in WordNet 1.6)

**1. sense:** horse, Equus\_caballus "solid-hoofed herbivorous quadruped domesticated since prehistoric times "

**2. sense:** horse "a padded gymnastic apparatus on legs "

**3. sense:** cavalry, horse\_cavalry, horse "troops trained to fight on horseback: "

**4. sense:** sawhorse, horse, sawbuck, buck "a framework for holding wood that is being sawed "

**5. sense:** knight, horse "a chessman in the shape of a horse's head; can move two squares horizontally and one vertically (or vice versa) "

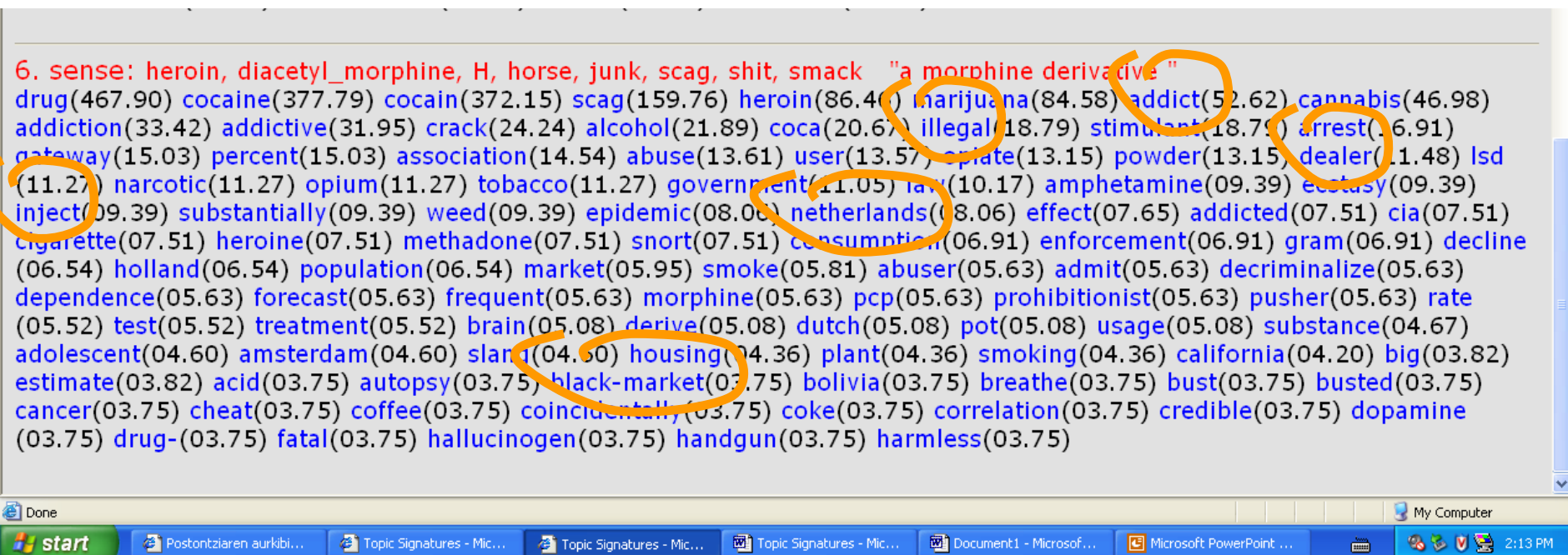
**6. sense:** heroin, diacetyl\_morphine, H, horse, junk, scag, shit, smack "a morphine derivative "

**1. sense:** horse, Equus\_caballus "solid-hoofed herbivorous quadruped domesticated since prehistoric times "

polo(112.40) equus(102.66) zebra(101.61) **ehinnus**(86.65) quagga(83.87) horse(79.18) pony(78.52) hinny(67.16) stablemate(54.63) racehorse(53.24) **donkey**(47.32) **liver**(34.45) mare(34.35) mussel(31.66) race(28.98) pinto(26.67) **bangtail**(26.10) **workhorse**(25.75) palomino(24.75) **saddle**(14.36) stallion(24.36) dawn(23.68) mesohippus(22.27) equid(19.18) riding(19.20) companion(18.57) harness(18.30) **specie**(17.71) extinct(15.66) offspring(15.66) chestnut(15.61) female(15.47) hyracotherium(15.31) foal(14.61) ass(13.92) ancestor(13.72) **hybrid**(13.22) stable(12.67) filly(11.30) trainer(10.66) fossil(10.09) mule(10.08) thoroughbred(09.74) dreissena(08.70) **breed**(08.50) burn(08.35) ride(07.50) breeding(06.96) age(06.77) wild(06.62) racing(06.61) modern(06.22) champion(06.18) ago(06.05) male(05.70) broodmare(05.56) finch(05.56) mammal(05.56) dog(05.38) printer(05.38) colt(05.33) equine(05.12) owner(05.04) derby(04.87) **midget**(04.87) oligocene(04.87) sterile(04.87) arabian(04.69) ownership(04.69) genus(04.48) rescue(04.48) domestic(04.44) trail(04.30) eocene(04.17) mustang(04.17) subspecies(04.17) animal(03.85) bean(03.84) stud(03.84) gelding(03.82) **sheep**(03.82) evolution(03.63) tail(03.50) breeder(03.48) protohippus(03.48) dressage(03.41) prehistoric(03.41) rider(03.36) toe(03.23) creature(03.20) equidae(03.13) feral(03.13) sorrel(03.13) sire(03.09) mane(02.98) native(02.98) retire(02.98) evolve(02.96) tooth(02.96) cave(02.78)

# RT5 New KnowNets

- Topic Signatures:
  - [http://ixa3.si.ehu.es/cgi-bin/signatureak/signature\\_lem.cgi](http://ixa3.si.ehu.es/cgi-bin/signatureak/signature_lem.cgi)
  - Word vectors (acquired from the web) associated to synsets



# RT5 New KnowNets

- Similar process in:
  - <http://adimen.si.ehu.es/cgi-bin/WSDbyEvocation.v1/index.php>
  - <http://adimen.si.ehu.es/cgi-bin/WSDbyEvocation.v3/index.php>
- Using LDA on POS tagged BNC to obtain the word vectors
- Using SSSI-Dijkstra+ for WSD
- Try:
  - bank.n + river.n
  - bank.n + money.n

# RT5 New KnowNets

- New KnowNets:
  - word embeddings
    - <https://embeddings.sketchengine.co.uk/>
    - English BNC 100M, Web 20B
    - Characterize queries per synset
    - Generate Topic Signatures
  - knowledge-based WSD (UKB)
  - => New KnowNets
  - Evaluation:
    - SensEval/SemEval, lexical sample WSD
    - <http://adimen.si.ehu.es/web/WSD-WN-Glosses>

# RT6 Synset embeddings

- Existing synset embeddings ...
  - AutoExtend: <http://www.cis.lmu.de/~sascha/AutoExtend/>
  - SensEmbed: <http://lcl.uniroma1.it/senseMBED/>
  - <http://lcl.uniroma1.it/sw2v>
  - <http://www.aclweb.org/anthology/D14-1110>
    - <http://pan.baidu.com/s/1eQcPK8i>
  - <http://www.aclweb.org/anthology/C14-1048>
  - ...
- How to evaluate them?
  - Similarity/relatedness datasets:
    - <http://www.wordvectors.org/>
  - Word vs. Synset similarity

# RT6 Synset embeddings

- New synset embeddings ... with existing software
  - UKB ... <http://ixa2.si.ehu.es/ukb/> instead of word embeddings ...
  - Node2vec ... <https://snap.stanford.edu/node2vec/>
  - DeepWalk: <https://github.com/phanein/deepwalk>
  - LINE: <https://github.com/tangjianpku/LINE>
  - SDNE: <https://github.com/suanrong/SDNE>
  - ...
  - <https://github.com/thunlp/NRLPapers>
  - <https://github.com/chi ming/awesome-network-embedding>
  - Network Representation Learning: A Survey
  - ...
- How to evaluate them?
  - Similarity/relatedness datasets:
    - <http://www.wordvectors.org/>
  - Word vs. Synset similarity
  - ...

# RT6 Synset embeddings

- New synset embeddings ... with existing word embeddings.
- MCR + word embeddings + strategy => synset embeddings.
- <https://embeddings.sketchengine.co.uk/>
  
- Characterise synsets by related words
- Characterise synsets by related (crosslingual) words
  - First rotating vectors?
  - <https://github.com/artetxem/vecmap>
  
- How to evaluate them?
  - Similarity/relatedness datasets:
    - <http://www.wordvectors.org/>
  - Word vs. Synset similarity

# RT7 Mapping new multilingual words to synsets

- Once having synset embeddings
- How to obtain the most similar word vector from another language?
- <https://github.com/spotify/annoy>
-

# Research topics



German Rigau i Claramunt  
<http://adimen.si.ehu.es/~rigau>

IXA group

Departamento de Lenguajes y Sistemas Informáticos  
UPV/EHU

# RT1 Tools: UKB and word2vec

- UKB:
  - <http://ixa2.si.ehu.es/ukb>
- word2vec:
  - <https://code.google.com/archive/p/word2vec>
  - <http://deeplearning4j.org/word2vec>
  - <https://radimrehurek.com/gensim/models/word2vec.html>
- Similarity/relatedness datasets:
  - <http://www.wordvectors.org>

# RT4: Mapping WordNet to EuroVoc / IATE

<http://eurovoc.europa.eu/5650> “greenhouse gas”

- How to perform the task?
- How to evaluate this task?

# RT7 Geo-country Tagging

Mariano Rajoy announced yesterday in Madrid that the budget cuts will continue next year.

Madrid => Spain?

Mariano\_Rajoy => Spain?

- Where is that knowledge?
- How to use that knowledge to perform the task?
- How to evaluate this task?