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
  - ...

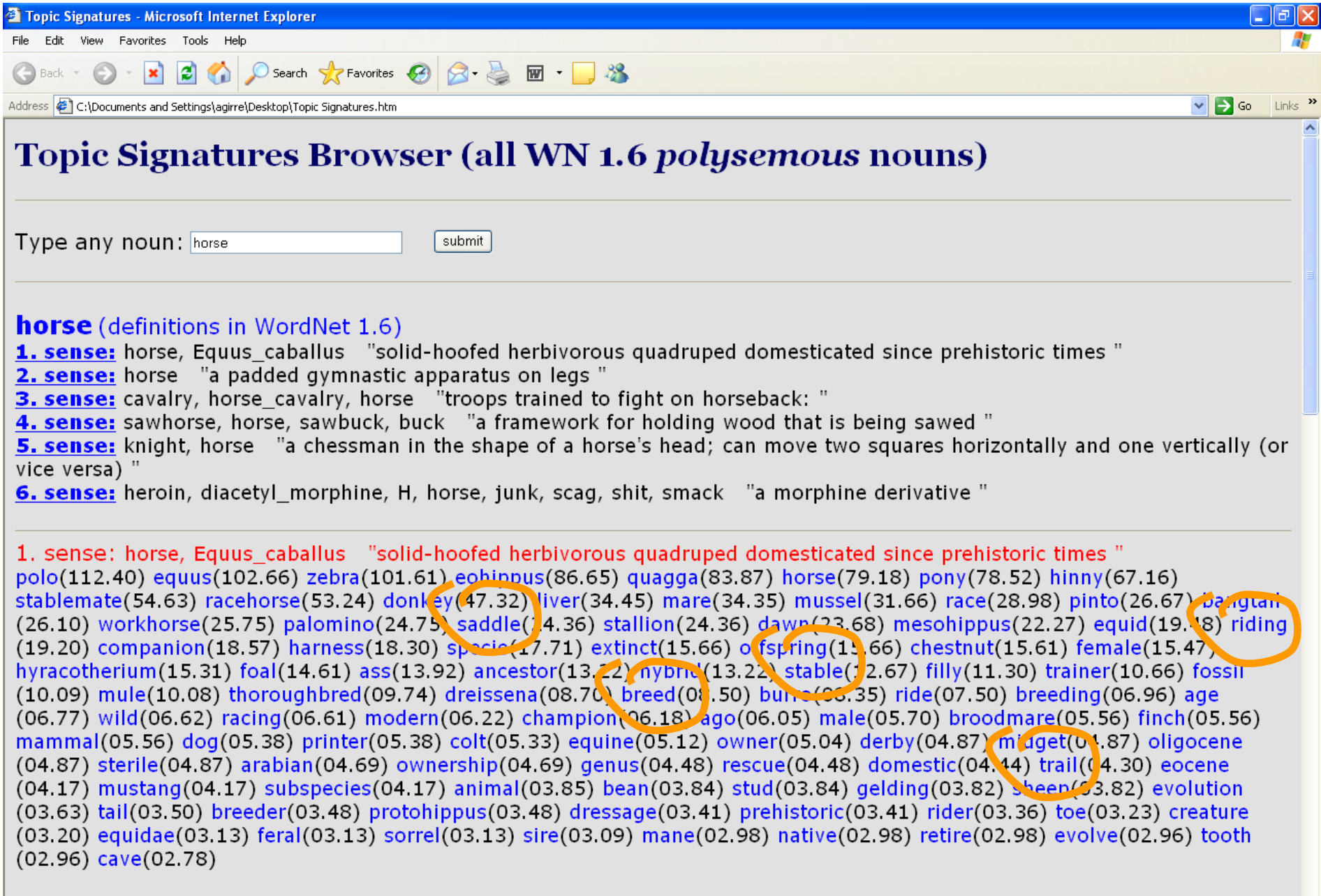
RT4 VisualGenome

- 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 SensEval/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



RT5 New KnowNets

- Topic Signatures:
  - 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/chihming/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?