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: Winograd Schema Challenge
- RT2: Domain tagging of WordNet (or text)
- RT3: NMT of WordNet definitions
- RT4: Automatic generation of definitions in context
- RT5: Acquiring new WordNet semantic relations
- ...

Winograd Schema Challenge

- The trophy would not fit in the brown suitcase because **it** was too big (small). What was too big (small)?

- Answer 0: the trophy

- Answer 1: the suitcase

Winograd Schema Challenge

- The bee landed on the flower because **it** had pollen.

- Answer: the flower

- The bee landed on the flower because **it** wanted pollen.

- Answer: the bee

Winograd Schema Challenge

- Levesque, Hector J., Ernest Davis, and Leora Morgenstern. *The Winograd schema challenge*. Knowledge representation. 2012.

- Winograd, T. *Understanding Natural Language*. NewYork: Academic Press. 1972.

- https://www.cs.nyu.edu/davise/papers/WS.html

- http://www.hlt.utdallas.edu/~vince/data/emnlp12/

- SOTA

RT2: Domain Tagging of WordNet (or text)

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?
- Multilingual? ES-CA-EUS

RT2: Domain Tagging of WordNet (or text)

- 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?
  - As a text classification problem?
  - Which method is best? Knowledge-based, statistical, ML, DL?
  - Which system is best? UKB, Simple Transformers, ...
- 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 (or text)

Pedro Sanchez 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?
- Multilingual? ES-CA-EUS

RT3: NMT of WordNet definitions

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

ES hospital#n#1 : centro de salud donde los pacientes reciben tratamiento

- Where is that knowledge?
- How to use that knowledge to perform the task?
    - Which method? Rule-base, statistical, DL? Which system?
- How to evaluate this task?
- Multilingual: ES-CA-EUS-GLG

RT4: Automatic generation of definitions in context

Word	Context	Definition	Usage
order	Every person is hereby ordered to immediately evacuate the city of New Orleans.	make a formal or authoritative request to someone	He was ordered to leave the room.
order	The order of the knights templar was formed during the crusades when many knights and squires set out for the holy land.	a formal rule of law, especially in the roman catholic church	The order of the church.
skirt	She was wearing a knee-length dark blue jean skirt with a front slit and a blue backless top.	a garment worn in the upper part of the body	She wore a silk skirt.
skirt	Along the scenic route skirting the rim we stopped at every lookout to gaze at the fantastic scenery.	go over a wide area	The road was not to be skirted.
agreement	No provision was made in the agreements for time off, sick pay, or holidays.	the action of establishing a relationship between two or more parties	The agreement between the two countries.
gaze	I took my chair to the open corridor and sat there with my book, gazing at the sunset.	look at a place steadily	She gaze at the room.

Table 6: Some examples of definitions and usages generated by our multi-task models from Oxford-2019 test set.

- Where is that knowledge?
- How to use that knowledge to perform the task?
- How to evaluate this task?
- Multilingual: ES-CA-EUS-GLG
- See (Zhang et al. 2019)

RT4: Acquiring new WordNet semantic relations

- No semantic relation connecting "tree" and "leaf" in WN

- Where is that knowledge?
  - There is a plethora of common sense resources
- How to use that knowledge to perform the task?

  - Applying a WSD algorithm to disambiguate the senses?

  - How to acquire the relation?
- How to evaluate this task?

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: ukb vs. w2v on similarity/relatedness
- RT2: New KnowNets
- RT3: Mapping WordNet to Wikipedia
- RT4: New eXtended WordNet domains
- RT5: Synset embeddings
- RT6: Mapping new multilingual words to synsets
- ...

ukb vs. w2v on similarity/relatedness

- 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.cs.cmu.edu/~mfaruqui/suite.html

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

- 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

New KnowNets

- New KnowNets:
    - word2vec (word embeddings)
    - Knowledge-based WSD (SSI-Dijkstra+, UKB)
    - => New KnowNets
    - Evaluation: lexical sample WSD

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

RT5 New KnowNets

Topic Signatures - Microsoft Internet Explorer

File   Edit   View   Favorites   Tools   Help

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

Topic Signatures Browser (all WN 1.6 *polysemous* nouns)

Type any noun: horse   submit

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) eohippus(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(24.36) stallion(24.36) dawn(23.68) mesohippus(22.27) equid(19.78) riding (19.20) companion(18.57) harness(18.30) species(17.71) extinct(15.66) offspring(15.66) chestnut(15.61) female(15.47) hyracotherium(15.31) foal(14.61) ass(13.92) ancestor(13.22) 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) burro(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
  - Word vectors (acquired from the web) associated to synsets

6. sense: heroin, diacetyl_morphine, H, horse, junk, scag, shit, smack   "a morphine derivative "
drug(467.90) cocaine(377.79) cocain(372.15) scag(159.76) heroin(86.46) marijuana(84.58) addict(52.62) cannabis(46.98) addiction(33.42) addictive(31.95) crack(24.24) alcohol(21.89) coca(20.67) illegal(18.79) stimulant(18.79) arrest(16.91) gateway(15.03) percent(15.03) association(14.54) abuse(13.61) user(13.57) opiate(13.15) powder(13.15) dealer(11.48) lsd (11.27) narcotic(11.27) opium(11.27) tobacco(11.27) government(11.05) law(10.17) amphetamine(09.39) ecstasy(09.39) inject(09.39) substantially(09.39) weed(09.39) epidemic(08.06) netherlands(08.06) effect(07.65) addicted(07.51) cia(07.51) cigarette(07.51) heroine(07.51) methadone(07.51) snort(07.51) consumption(06.91) enforcement(06.91) gram(06.91) decline (06.54) holland(06.54) population(06.54) market(05.95) smoke(05.81) abuser(05.63) admit(05.63) decriminalize(05.63) dependence(05.63) forecast(05.63) frequent(05.63) morphine(05.63) pcp(05.63) prohibitionist(05.63) pusher(05.63) rate (05.52) test(05.52) treatment(05.52) brain(05.08) derive(05.08) dutch(05.08) pot(05.08) usage(05.08) substance(04.67) adolescent(04.60) amsterdam(04.60) slang(04.60) housing(04.36) plant(04.36) smoking(04.36) california(04.20) big(03.82) estimate(03.82) acid(03.75) autopsy(03.75) black-market(03.75) bolivia(03.75) breathe(03.75) bust(03.75) busted(03.75) cancer(03.75) cheat(03.75) coffee(03.75) coincidentally(03.75) coke(03.75) correlation(03.75) credible(03.75) dopamine (03.75) drug-(03.75) fatal(03.75) hallucinogen(03.75) handgun(03.75) harmless(03.75)

Done   My Computer

start   Postontziaren aurkibi...   Topic Signatures - Mic...   Topic Signatures - Mic...   Topic Signatures - Mic...   Document1 - Microsof...   Microsoft PowerPoint ...   2:13 PM

New KnowNets

- New KnowNets:
  - word2vec (word embeddings)
    - https://code.google.com/p/word2vec
    - http://deeplearning4j.org/word2vec.html
    - https://radimrehurek.com/gensim/models/word2vec.html

      ...
  - Knowledge-based WSD (SSI-Dijkstra+, UKB)
    - http://ixa2.si.ehu.es/ukb
  - => New KnowNets (a very dense graph!)
  - Evaluation: lexical sample WSD
    - http://adimen.si.ehu.es/web/WSD-WN-Glosses

Mapping WordNet to Wikipedia

- Using UKB to disambiguate Wikipedia entries

- Create a sample gold-standard dataset

- Have a look at the disambiguation pages ...

  - For example:

  - https://en.wikipedia.org/wiki/Party_%28disambiguation%29

  - 6 senses as a common noun in Wikipedia

  - 5 sense as a common noun in WN ...

  - ...

New eXtended WordNet domains

- Using WN Domains from MCR

- Use the disambiguated senses from the gloss corpus

- Propagate using ukb_ppr

- Obtain the top k synsets from the resulting vector

- Select the WN Domains from the top k synsets

- Evaluate

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
-

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?

Research topics

German Rigau i Claramunt

http://adimen.si.ehu.es/~rigau

IXA group

Departamento de Lenguajes y Sistemas Informáticos

UPV/EHU