Word Sense Alignment as a Classification Problem

Sina Ahmadi, John P. McCrae
National University of Ireland Galway
Insight Centre for Data Analytics

The 11th Global Wordnet Conference 2021
18-22 January 2021

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Overview

- Context
- Objectives
- Approach
- Experiments
- Results

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Context

"Dictionaries are treasure houses of data on the uses of words. They are also our best starting point for all questions regarding word sense distinctions, in NLP, the humanities or lexicography. But to reveal the dictionary's treasures in a systematic way is no simple."

— Adam Kilgarriff (*Dictionary Word Sense Distinctions: An Enquiry into Their Nature*)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Context: Word senses

- Words have different meanings, i.e. senses
- Senses may be differently described in various resources:
  - dictionaries
  - encyclopedia
  - thesauri
  - WordNets

Word sense alignment (WSA) ➡️ linking senses across resources

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Context: entire (adj)

Adjective [edit]

entire (*not comparable*)

1. (*sometimes postpositive*) Whole; complete. [quotations ▼]

   *We had the **entire** building to ourselves for the evening.*

2. (*botany*) Having a smooth margin without any indentation.
3. (*botany*) Consisting of a single piece, as a corolla.
4. (*complex analysis*, *of a complex function*) Complex-differentiable on all of ℂ.
5. (*of a male animal*) Not gelded.
6. morally whole; pure; sheer [quotations ▼]
7. Internal; interior.

Adjective

- S: (adj) **entire**, full, total (constituting the full quantity or extent; complete) *"an entire town devastated by an earthquake"; "gave full attention"; "a total failure"*
- S: (adj) integral, **entire**, intact (constituting the undiminished entirety; lacking nothing essential especially not damaged) *"a local motion keepeth bodies integral"- Bacon; "was able to keep the collection entire during his lifetime"; "fought to keep the union intact"*
- S: (adj) **entire** ((of leaves or petals) having a smooth edge; not broken up into teeth or lobes)
- S: (adj) **entire**, intact ((used of domestic animals) sexually competent) *"an entire horse"*

Global WordNet Association

entire

Pronunciation (?) /ɪnˈtʌɪə/ /ɛnˈtʌɪə/

Translate entire into Spanish

ADJECTIVE

1 *[attributive]* With no part left out; whole.

'my plans are to travel the entire world'

+ More example sentences + Synonyms

1.1 Without qualification or reservations; absolute.

'an ideological system with which he is in entire agreement'

+ More example sentences + Synonyms

2 Not broken, damaged, or decayed.

+ More example sentences + Synonyms

3 (of a male horse) not castrated.

4 *Botany*

(of a leaf) without indentations or division into leaflets.

Powered by OXFORD LEXICO

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

elexis european lexicographic infrastructure

Objectives

1. What are the relevant features that can be used for WSA?
   - ➡️ Manual extraction of features (outdated? 🤔)
2. How representation learning can help in improving the results?
   - ➡️ Restricted Boltzmann machine
3. Use the features for classifying semantic relationship between two given senses
   - ➡️ `"exact"`, `"narrower"`, `"broader"`, `"related"`, `"none"`
4. How our observations can pave the way for a more robust approach?

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

elexis european lexicographic infrastructure

Approach

1. Data description
2. Feature extraction
3. Data augmentation:
   a. symmetric relationships: `exact`, `related`
   b. asymmetric relationships: `narrower`, `broader`
4. Representation learning
5. Classification

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Approach: Data description

Monolingual Word Sense alignment (MWSA)* datasets which contain:

- Lexicographic data in 15 languages
- 17 datasets from various resources
  - expert-made lexicons, such as the Danish Dictionary
  - collaboratively-curated resources such as Wiktionary
- Semantically-annotated senses based on SKOS**: `exact`, `broader`, `narrower`, `related` and `none`

```
{
    "lemma": "splenetic",
    "POS_tag": "adjective",
    "gender": "",
    "meta_ID": "",
    "resource_1_senses": [
        {
            "#text": "of or relating to the spleen",
            "external_ID": "splenic.a.01"},
        {
            "#text": "very irritable",
            "external_ID": "bristly.s.01"}
    ],
    "resource_2_senses": [
        {
            "#text": "affected with spleen; malicious;
            ↪  spiteful; peevish; fretful.",
            "external_ID": ""}
    ],
    "alignment": [
        {
            "sense_source": "very irritable",
            "sense_target": "affected with spleen;
            ↪  malicious; spiteful; peevish;
            ↪  fretful.",
            "semantic_relationship": "exact"}
    ]
}
```

* https://github.com/elexis-eu/mwsa

** SKOS Simple Knowledge Organization System Primer: https://www.w3.org/TR/skos-primer/

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Approach: Data description

Language	# Entries	# SKOS	# SKOS+none
Basque (eu)	256	813	3661
Bulgarian (bg)	1000	1976	3708
Danish (da)	587	1644	16520
Dutch (nl)	161	622	20144
English (en)	684	1682	9269
Estonian (et)	684	1142	2316
German (de)	537	1211	4975
Hungarian (hu)	143	949	15774
Irish (ga)	680	975	2816
Italian (it)	207	592	2173
Serbian (sr)	301	736	5808
Slovenian (sl)	152	244	1100
Spanish (es)	351	1071	4898
Portuguese (pt-pt)	147	275	2062
Russian (rs)	213	483	3376

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Approach: Data augmentation

➡️ Collecting new data instances by inverting sense order and semantic relation

Example: `observer`

	Data instance 1	**Data instance 2**
Sense 1	an expert who observes and comments on something (WordNet)	an annotator (Webster)
Sense 2	an annotator (Webster)	an expert who observes and comments on something (WordNet)
Sem. rel.	**narrower**	**broader**

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Approach: Data augmentation

Language	# Entries	# SKOS	# SKOS+none	# All
Basque (eu)	256	813	3661	4382
Bulgarian (bg)	1000	1976	3708	5656
Danish (da)	587	1644	16520	18164
Dutch (nl)	161	622	20144	20766
English (en)	684	1682	9269	10951
Estonian (et)	684	1142	2316	3426
German (de)	537	1211	4975	6185
Hungarian (hu)	143	949	15774	16716
Irish (ga)	680	975	2816	3763
Italian (it)	207	592	2173	2758
Serbian (sr)	301	736	5808	6542
Slovenian (sl)	152	244	1100	1343
Spanish (es)	351	1071	4898	5919
Portuguese (pt-pt)	147	275	2062	2337
Russian (rs)	213	483	3376	3845

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Approach: Features

Representing various textual and semantic relations by incorporating word embeddings and a semantic network

#	feature	definition	possible values
1	POS_tag	part of speech of the headword	a one-hot vector of {N, V, ADJ, ADV, OTHER}
2	s_1_len_no_func	number of space-separated tokens in s_1	$\mathbb{N}$
3	s_2_len_no_func	number of space-separated tokens in s_2	$\mathbb{N}$
4	s_1_len	number of space-separated tokens in s_1 without function words	$\mathbb{N}$
5	s_2_len	number of space-separated tokens in s_2 without function word	$\mathbb{N}$

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Approach: Features

6	hypernymy	hypernymy score between tokens	sum of weights in CONCEPTNET
7	hyponymy	hyponymy score between tokens	sum of weights in CONCEPTNET
8	relatedness	relatedness score between tokens	sum of weights in CONCEPTNET
9	synonymy	synonymy score between tokens	sum of weights in CONCEPTNET
10	antonymy	antonymy score between tokens	sum of weights in CONCEPTNET
11	meronymy	meronymy score between tokens	sum of weights in CONCEPTNET
12	similarity	similarity score between tokens	sum of weights in CONCEPTNET
13	sem_sim	semantic similarity score between senses	averaging word vectors and cosine similarity
14	sem_sim_no_func	semantic similarity score between senses without function words	averaging word vectors and cosine similarity excluding function words
15	sem_bin_rel	target class	0 for unrelated, 1 for related
16	sem_rel_with_none	target class	{exact, narrower, broader, related, none}
17	sem_rel	target class	{exact, narrower, broader, related}

Incorporate ConceptNet*

Word embeddings

Target features

* https://conceptnet.io/

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Approach: Features

An example: `entire (adj)`

- Sense 1: constituting the full quantity or extent
- Sense 2: complete in all parts
- Relation: exact

* en complete (a, wn) — SimilarTo → en full (a, wn)
Weight: 2.0

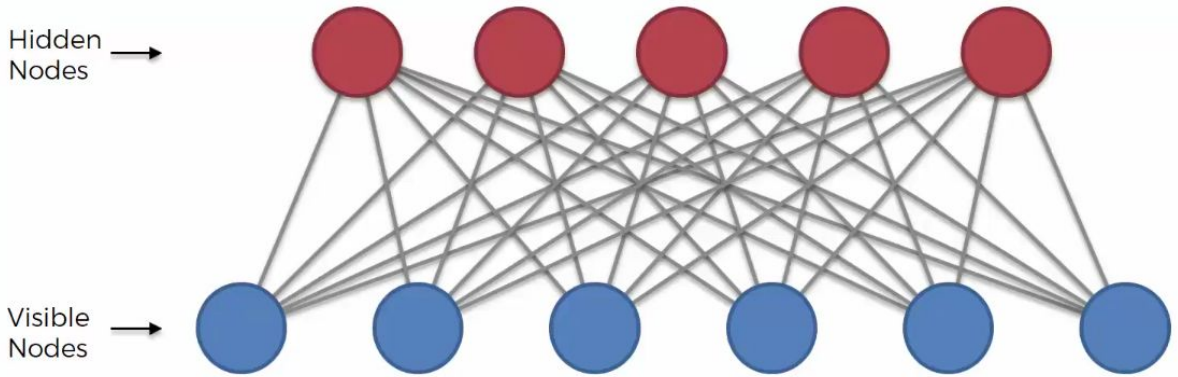
IsA	hypernymy	RelatedTo	Synonym	Antonym	PartOf	**SimilarTo**
0.01	0.01	0.10	0.04	0.01	0.01	**0.80**

* https://www.conceptnet.io/c/en/complete

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Approach: Representation learning

- Using a Restricted Boltzmann machine, map the extracted features (visible) to a set of hidden ones
  - A generative stochastic artificial neural network
  - **Latent features**: Find hidden parameters within the set of input features
  - Many applications in dimensionality reduction, recommender systems and topic modelling



$$p(v) = \sum_h p(v, h)$$

where v ∈ V, h ∈ H

* Hinton, G. E. (2012). A practical guide to training restricted Boltzmann machines. In *Neural networks: Tricks of the trade* (pp. 599-619). Springer, Berlin, Heidelberg.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Approach: Classification

Three classification problems using support vector machines (SVMs):

1. **Binary classification** (`alignable`, `non-alignable`)

   predicts if two senses can possibly be aligned together

2. **SKOS classification** (`exact`, `broader`, `narrower` **and** `related`)

   predicts semantic relationship type of SKOS

3. **SKOS+none classification** similar to the previous classifier, with none as a target class

   predicts semantic relationship types of SKOS + none

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Experiments

- Features normalized and scaled
- Create a pipeline using scikit-learn (https://scikit-learn.org/)
- Ideal parameters are found using an exhaustive search function (`GridSearchCV`)
- SVM
  - Regularization parameter `C=1`
  - Kernel coefficient `gamma= [0.7, 0.8, 0.9, "scale"]`
  - Kernel function: `['linear', 'poly', 'rbf']`
  - Other parameters used as the default ones*
- RBM
  - Bernoulli RBM based, i.e. visible unit between 0 and 1
  - learning rate within [0.05-0.2]
  - hidden unit number within the range of 400 and 600

* see https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Results

- In all cases and languages, our approach performs better than the baseline
- Our models perform best on the binary classification problem
- The most difficult cases belong to the SKOS classification
- Only in 5/17 datasets, using RBM improves the F-measure

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

elexis european lexicographic infrastructure

Results

Language	Baseline (binary)*	Binary	All	SKOS	RBM-Binary	RBM-all	RBM-SKOS
Basque (eu)	8.10	72.08	58.83	44.80	67.94	55.62	29.46
Bulgarian (bg)	2.00	69.03	65.11	31.79	75.11	40.76	29.69
Danish (da)	4.30	75.02	34.43	28.12	58.02	63.80	43.00
Dutch (nl)	0.00	87.60	78.90	32.22	63.33	52.54	41.88
English (en)	0.00	82.61	74.68	42.66	72.17	70.65	42.57
Estonian (et)	15.90	32.62	62.85	42.35	39.05	54.43	27.83
German (de)	0.00	64.01	62.25	42.79	35.63	39.30	44.63
Hungarian (hu)	2.00	52.12	33.60	34.85	62.05	66.09	27.84
Irish (ga)	29.10	84.44	42.95	37.68	63.06	47.83	27.61
Italian (it)	0.00	58.69	47.32	34.22	56.49	55.32	35.32
Serbian (sr)	26.90	70.83	57.15	33.26	75.43	42.40	30.99
Slovenian (sl)	26.80	77.82	30.62	35.56	54.45	32.51	25.65
Spanish (es)	-	80.07	54.10	36.31	68.47	27.83	39.07
Portuguese (pt-pt)	3.70	42.57	63.74	53.49	64.01	33.78	39.26
Russian (rs)	25.50	77.29	52.39	31.31	70.71	72.04	38.82

*Kernerman, I., Krek, S., McCrae, J. P., Gracia, J., Ahmadi, S., & Kabashi, B. LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

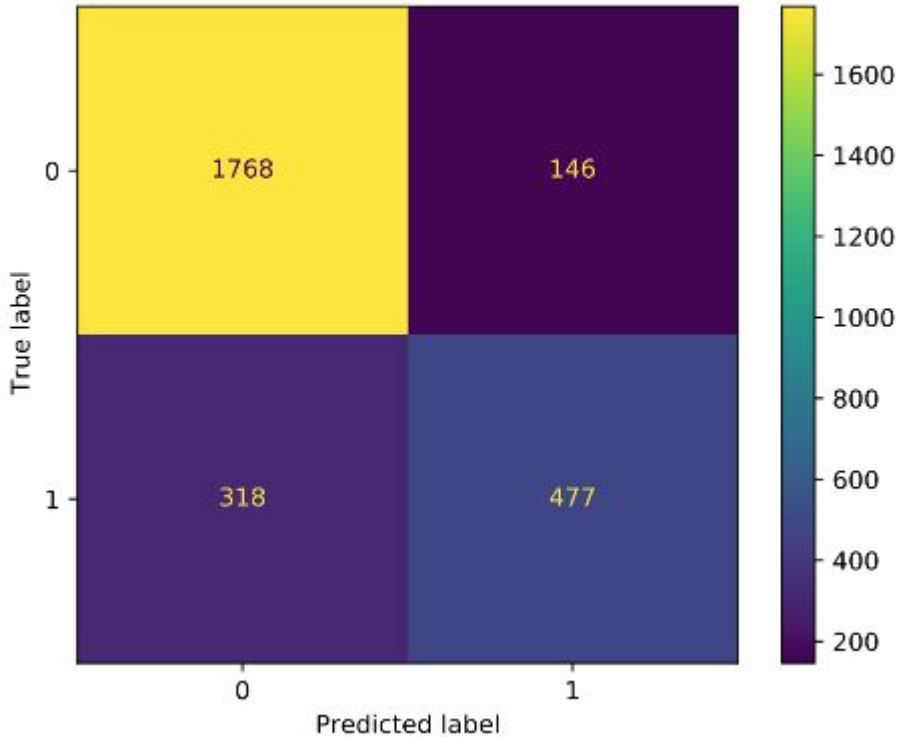
Results

Language	# Entries	# SKOS	# SKOS+none	# All
English (en)	684	1682	9269	10951

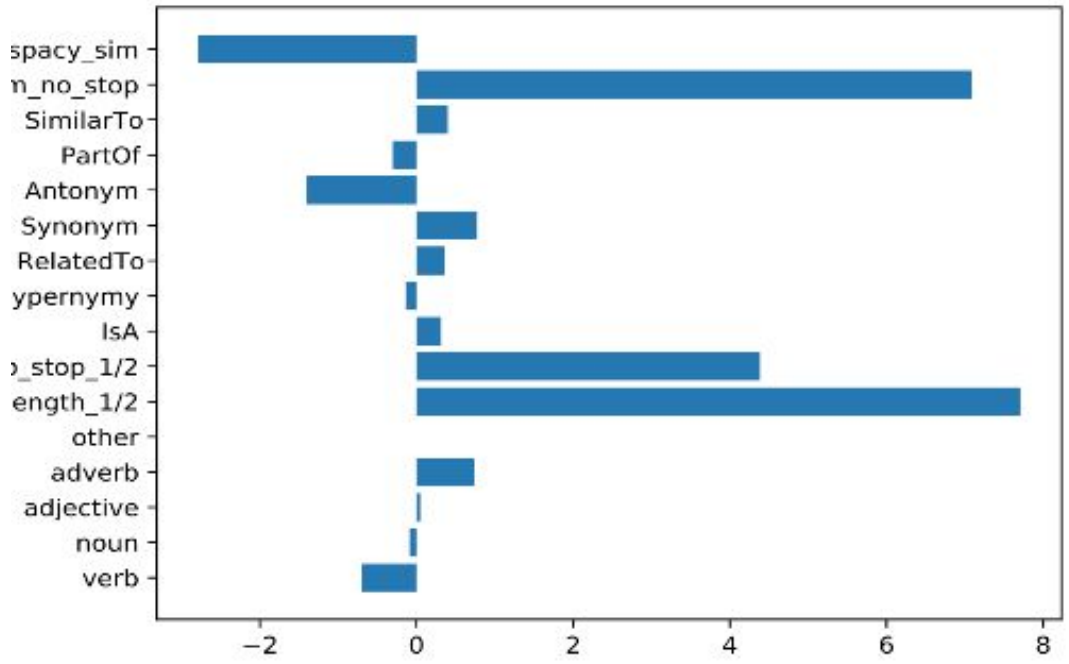
Metric	Binary	All	SKOS	RBM-Binary	RBM-all	RBM-SKOS
Accuracy	89.00	81.00	49.00	80.16	65.03	48.57
Precision	82.35	73.03	39.31	64.36	63.67	55.53
Recall	82.87	76.41	46.63	82.13	79.35	34.51
F-measure	82.61	74.68	42.66	72.17	70.65	42.57

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Results: binary classification



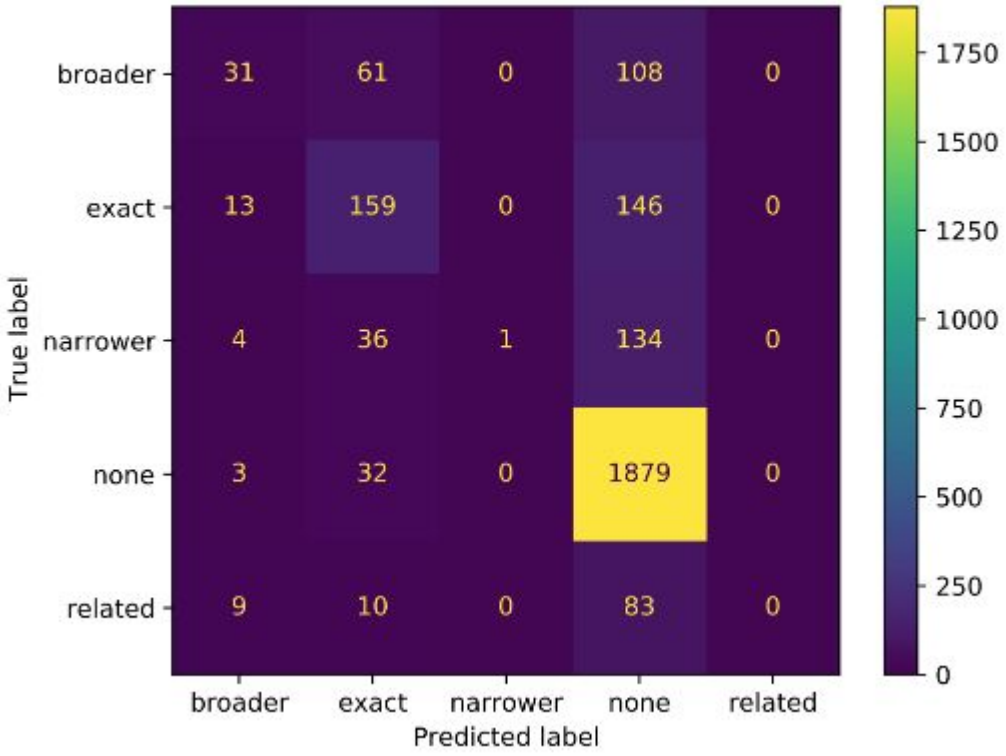
Confusion matrix



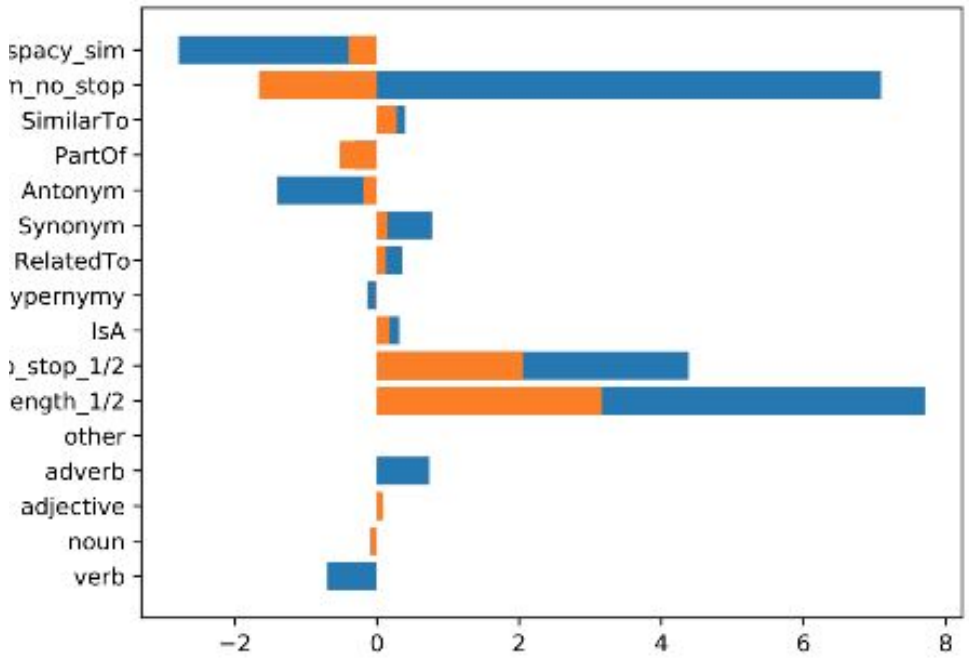
Features importance

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Results: All classification



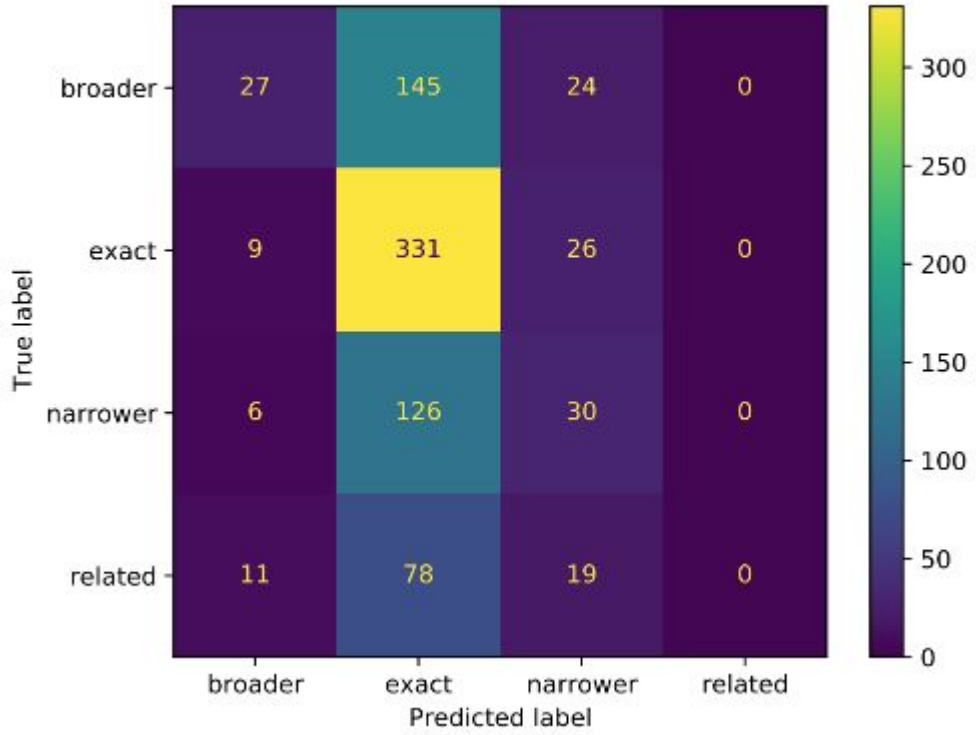
Confusion matrix



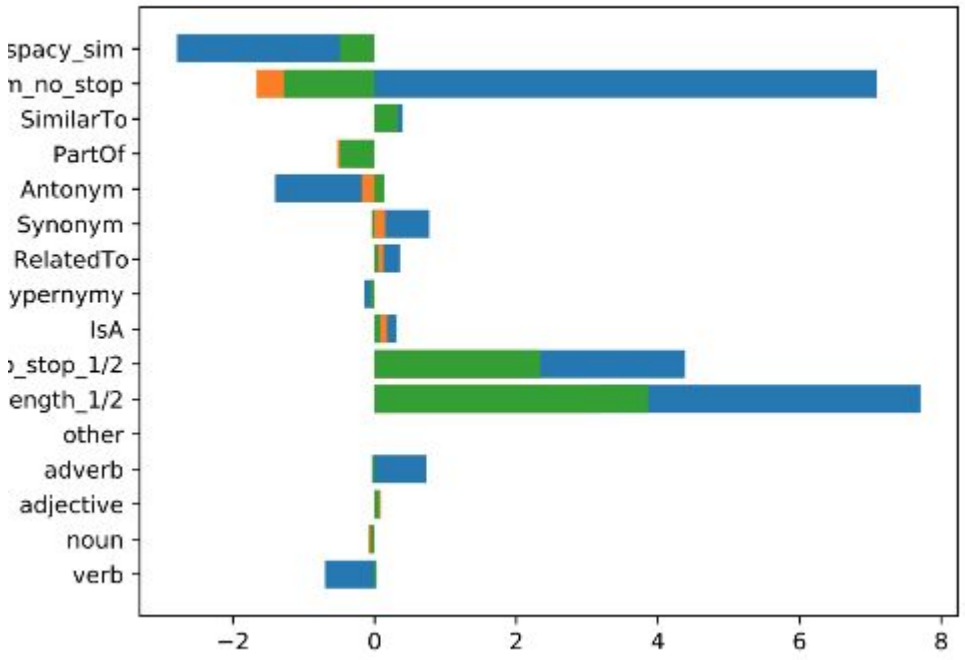
Features importance

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Results: SKOS classification



Confusion matrix



Features importance

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Conclusion

- Our handcrafted features could be efficiently used for binary classification of senses
- Limitations regarding the detection of semantic relations due to imbalanced data
- Using an RBM does not essentially improve the results in all cases
- More robust solutions as future work

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Any question? 🙂

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.