Lexical Semantic Change: Models, Data and Evaluation

LREC 2022 - Tutorial - 20 June 2022

Pierpaolo Basile[1], Annalina Caputo[2], Pierluigi Cassotti[1] and Rossella Varvara[3]
University of Bari[1], Dublin City University[2], Université de Fribourg[3]

Evaluation

Outline

- Tasks
  - Binary
  - Ranking
  - Temporal Analogies
  - LSC Discovery
- Results

Evaluation Tasks

Binary Task

- Binary classification
  - Given a target word t, decide if it **lost** or **gained senses** from T1 to T2
- SemEval 2020 Task 1 Subtask 1
- Diacr-ITA 2021
- Evaluation on accuracy

SemEval 2020 Task 1.1 : Unsupervised Lexical Semantic Change Detection

- Four languages
  - English
    - Clean Corpus of Historical American English (CCOHA) [1810-2000]
  - German
    - DTA (different genre) [16th-20th century]
    - BZ + ND (newspapers) [1945-1993]
  - Latin
    - LatinISE [2nd century B.C. - 21st century A.D.]
  - Swedish
    - Kubhist corpus (newspaper) [18th-20th century]

SemEval 2020 Task 1.1 : Unsupervised Lexical Semantic Change Detection

- Four languages
  - English
    - Clean Corpus of Historical American English (CCOHA) [1810-2000]
  - German
    - DTA (different genre) [16th-20th century]
    - BZ + ND (newspapers) [1945-1993]
  - Latin
    - LatinISE [2nd century B.C. - 21st century A.D.]
  - Swedish
    - Kubhist corpus (newspaper) [18th-20th century]

Lemmatized & POS-tagged

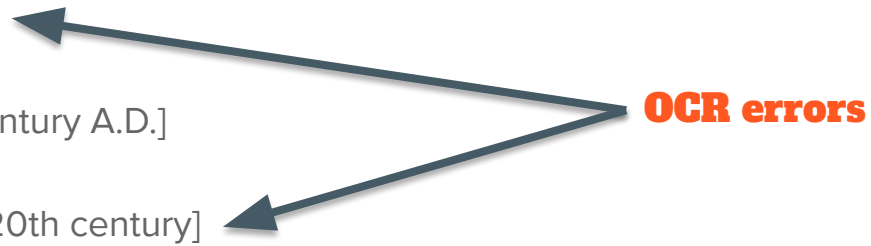
SemEval 2020 Task 1.1 : Unsupervised Lexical Semantic Change Detection

- Four languages
  - English
    - Clean Corpus of Historical American English (CCOHA) [1810-2000]
  - German
    - DTA (different genre) [16th-20th century]
    - BZ + ND (newspapers) [1945-1993]
  - Latin
    - LatinISE [2nd century B.C. - 21st century A.D.]
  - Swedish
    - Kubhist corpus (newspaper) [18th-20th century]

Spelling Normalised

SemEval 2020 Task 1.1 : Unsupervised Lexical Semantic Change Detection

- Four languages
  - English
    - Clean Corpus of Historical American English (CCOHA) [1810-2000]
  - German
    - DTA (different genre) [16th-20th century]
    - BZ + ND (newspapers) [1945-1993]
  - Latin
    - LatinISE [2nd century B.C. - 21st century A.D.]
  - Swedish
    - Kubhist corpus (newspaper) [18th-20th century]



SemEval 2020 Task 1.1 : Unsupervised Lexical Semantic Change Detection

- Two time periods: C_1 and C_2
- In-domain task

	C_1					C_2				
	corpus	**period**	**tokens**	**types**	**TTR**	**corpus**	**period**	**tokens**	**types**	**TTR**
English	CCOHA	1810–1860	6.5M	87k	13.38	CCOHA	1960–2010	6.7M	150k	22.38
German	DTA	1800–1899	70.2M	1.0M	14.25	BZ+ND	1946–1990	72.3M	2.3M	31.81
Latin	LatinISE	-200–0	1.7M	65k	38.24	LatinISE	0–2000	9.4M	253k	26.91
Swedish	Kubhist	1790–1830	71.0M	1.9M	47.88	Kubhist	1895–1903	110.0M	3.4M	17.27

SemEval 2020 Task 1.1 : Unsupervised Lexical Semantic Change Detection

- Annotation
- Small set of words
- 4 annotators per language
- Sampled 100 uses per each word from T1 and Tw
- Pairs of word uses from both periods annotated on a four-point scale
- Annotation based on graph usage
    - Edge is the median of annotator judgment
    - Clusters based on senses
    - Compare clusters with sense frequency distribution (SFD)

DIACR-Ita 2020

- L'Unità (newspaper) [1945-1970 / 1990-2014]
- Tokenized, POS-tagged, Lemmatized
- Gold Standard Creation
  - Selection of target words
    - KRONOS-IT
  - Filtering candidates
  - Annotation

DIACR-Ita 2020

- Annotation
- Selected 100 words for each target
  - Total 2,336 occurrences
  - 2 annotators for each sentence
- Valid instances of LSC
  - Targets that have acquired meaning only in T2 and never in T1
  - 18 target words: 6 changes - 12 stable

Ranking Task

- Ranking target words according to the degree of lexical semantic change between T1 and T2
  - SemEval 2020 Task 1 Subtask 2
  - RuShiftEval 2021

SemEval 2020 Task 1.2: Unsupervised Lexical Semantic Change Detection

- Same languages as task 1.1
- Sense Frequency Distribution (SFD) computed for each time period and cluster of sense
- Jensen-Shannon distance between normalised SFD used for ranking language change
- Spearman’s rank-order correlation with the gold rank

RuShiftEval

- Follows similar approach to SemEval 2020 Task1
- Russian National Corpus [pre-Soviet-post-Soviet]
- participants to provide 3 grades of LSC for each target word (99)
- Spearman rank correlation

Lexical Semantic Sense Discovery

- Differently from SemEval-like evaluation, list of target words comes from the corpora itself
  - Discovery of previously unknown semantic changes
- Challenges
  - Large number of predictions
- Filter only nouns, verbs and adjectives
- After selection, annotation phase
  - Based on clustering of word usage graph
  - 25 sentences annotated per each time period

Temporal Analogies

- Categorise words by meaning based on word representations
- New York Times articles [1990-2016]
- 27 Time beans – 1 for each year
- Vocabulary size: 20.936 (after removing words < 200 occurrences & stopwords)
- Articles associated with metadata: Title, Author, Release Date, Section Label

Temporal Analogies Task

- 59 Section labels: Business, Sports, Technology, etc
- Section label used to determine the word meaning
  - e.g. Amazon
  - 1995: 41% occurrences in ***World***
  - 2012: 50% occurrences in ***Technology***
- Dataset construction
  - Identify words in years that are particularly frequent into a section
  - 11 main sections retained: Arts, Business, Fashion & Style, Health, Home & Garden, Real Estate, Science, Sports, Technology, U.S., World
  - For each section s, word w, and time t, computes p = percentage of occurrence of w in s
  - For each word w, and section s, retain only the triplet <w, t, s> with highest percentage p
  - Filter triplet where p < 35%
  - For each section s, retain only the top-200 words ranked by percentage p

Temporal Analogies Task

Ground Truth

- 1888 <w, t, s> triples across 11 sections
- For each word-year pair, the associated category is the ground truth: <w, t>: s

Clustering Task

- Every pair of words as a series of decisions
- Pick any two (w, t) pairs:
  - If they are clustered together and have the same section label, this is a correct decision; otherwise
  - Clustering performed a wrong decision

Temporal Analogies Task

Metrics

- Normalised Mutual Information

$$NMI(L, C) = \frac{I(L; C)}{[H(L) + H(C)]/2},$$

- F-measure

$$F_\beta = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

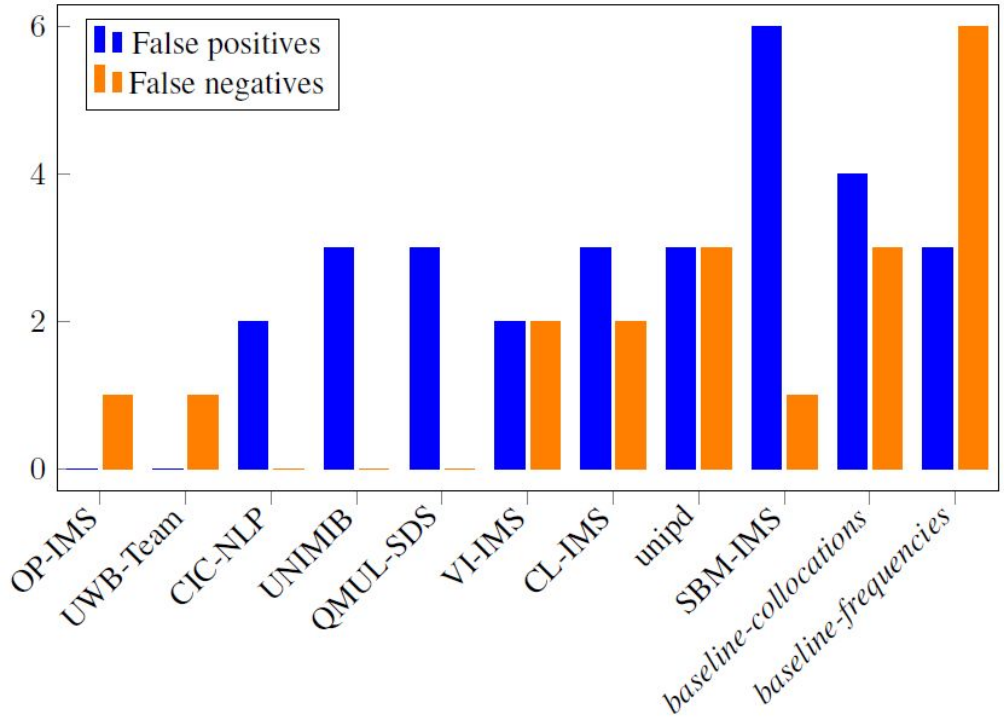
Evaluation Results

SemEval 2020: Task 1.1

Team	Subtask 1 Avg.	EN	DE	LA	SV	System
UWB	.687	.622	.750	.700	.677	type
Life-Language	.686	.703	.750	.550	.742	type
Jiaxin & Jinan	.665	.649	.729	.700	.581	type
RPI-Trust	.660	.649	.750	.500	.742	type
UG_Student_Intern	.639	.568	.729	.550	.710	type
DCC	.637	.649	.667	.525	.710	type
NLP@IDSIA	.637	.622	.625	.625	.677	token
JCT	.636	.649	.688	.500	.710	type
Skurt	.629	.568	.562	.675	.710	token
Discovery_Team	.621	.568	.688	.550	.677	ens.
Count Bas.	.613	.595	.688	.525	.645	-

Team	Subtask 1 Avg.	EN	DE	LA	SV	System
TUE	.612	.568	.583	.650	.645	token
Entity	.599	.676	.667	.475	.581	type
IMS	.598	.541	.688	.550	.613	type
cs2020	.587	.595	.500	.575	.677	token
UiO-UvA	.587	.541	.646	.450	.710	token
NLPCR	.584	.730	.542	.450	.613	token
Maj. Bas.	.576	.568	.646	.350	.742	-
cbk	.554	.568	.625	.475	.548	token
Random	.554	.486	.479	.475	.774	type
UoB	.526	.568	.479	.575	.484	topic
UCD	.521	.622	.500	.350	.613	graph
RIJP	.511	.541	.500	.550	.452	type
Freq. Bas.	.439	.432	.417	.650	.258	-

DIACR-Ita 2020



Team	Accuracy
OP-IMS	0.944
UWB Team	0.944
CIC-NLP	0.889
UNIMIB	0.833
QMUL-SDS	0.833
VI-IMS	0.778
CL-IMS	0.722
unipd	0.667
SBM-IMS	0.611
baseline-collocations	0.611
baseline-frequencies	0.500

Table 3: Results.

SemEval 2020: Task1.2

Team	Subtask 2 Avg.	EN	DE	LA	SV	System
UG_Student_Intern	.527	.422	.725	.412	.547	type
Jiaxin & Jinan	.518	.325	.717	.440	.588	type
cs2020	.503	.375	.702	.399	.536	type
UWB	.481	.367	.697	.254	.604	type
Discovery_Team	.442	.361	.603	.460	.343	ens.
RPI-Trust	.427	.228	.520	.462	.498	type
Skurt	.374	.209	.656	.399	.234	token
IMS	.372	.301	.659	.098	.432	type
UiO-UvA	.370	.136	.695	.370	.278	token
Entity	.352	.250	.499	.303	.357	type
Random	.296	.211	.337	.253	.385	type

Team	Subtask 2 Avg.	EN	DE	LA	SV	System
NLPCR	.287	.436	.446	.151	.114	token
JCT	.254	.014	.506	.419	.078	type
cbk	.234	.059	.400	.341	.136	token
UCD	.234	.307	.216	.069	.344	graph
Life-Language	.218	.299	.208	-.024	.391	type
NLP@IDSIA	.194	.028	.176	.253	.321	token
Count Bas.	.144	.022	.216	.359	-.022	-
UoB	.100	.105	.220	-.024	.102	topic
RIJP	.087	.157	.099	.065	.028	type
TUE	.087	-.155	.388	.177	-.062	token
DCC	-.083	-.217	.014	.020	-.150	type
Freq. Bas.	-.083	-.217	.014	.020	-.150	-
Maj. Bas.	-	-	-	-	-	-

RuShiftEval

	Team	RuSemShift1	RuSemShift2	RuSemShift3	Mean	Type
1	**GlossReader**	0.781	**0.803**	**0.822**	**0.802**	token
2	**DeepMistake**	**0.798**	0.773	0.803	0.791	token
3	vanyatko	0.678	0.746	0.737	0.720	token
4	**aryzhova**	0.469	0.450	0.453	0.457	token
5	Discovery	0.455	0.410	0.494	0.453	token
6	**UWB**	0.362	0.354	0.533	0.417	type
7	dschlechtweg	0.419	0.373	0.383	0.392	type
8	jenskaiser	0.430	0.310	0.406	0.382	token
9	**SBX-HY**	0.388	0.281	0.439	0.369	type
	Baseline	0.314	0.302	0.381	0.332	type
10	svart	0.163	0.223	0.401	0.262	type
11	**BykovDmitrii**	0.274	0.202	0.307	0.261	token
12	fdzr	0.217	0.251	0.065	0.178	type

Temporal Analogies

Table 4: Normailized Mutual Information (NMI).

Method	10 Clusters	15 Clusters	20 Clusters
SW2V	0.6736	0.6867	0.6713
TW2V	0.5175	0.5221	0.5130
AW2V	0.6580	0.6618	0.6386
DW2V	**0.7175**	**0.7162**	**0.6906**

Table 5: F-measure (F_β).

Method	10 Clusters	15 Clusters	20 Clusters
SW2V	0.6163	0.7147	0.7214
TW2V	0.4584	0.5072	0.5373
AW2V	0.6530	0.7115	0.7187
DW2V	**0.6949**	**0.7515**	**0.7585**

Thanks!