EVALITA 2009

The Lexical Substitution Task

A. Toral
Istituto di Linguistica Computazionale (CNR)

Outline

- Introduction to the Lexical Substitution Task
- Evaluation
  - Dataset
  - Metrics
  - Participants
  - Results
  - Discussion
- Conclusion and future work

Lexical Substitution

- Given a word in a context → find substitutes that fit (synonyms in context)
  - [...] nel tribunale di Pescara, riprendeva il **processo** contro [...] *([...] in the court of Pescara, the **trial** against [...] resumed)*
    - Giudizio
  - [...] deve avere convinto i vertici della necessità di accelerare il **processo** di ristrutturazione *([...] should have convinced the top brass of the need to accelerate the restructuring **process**)*
    - Sviluppo, procedura

Lexical Substitution

- Alternative to evaluate WSD
- Advantages
  - Absence of pre-defined inventory
    - Allows to evaluate unsupervised systems
    - WSD inventories too fine-grain
- First time for IT
  - EN in Semeval 2007

Dataset

- Words to be substituted chosen from LRs (IWN, PSC) following diverse criteria to guarantee
  - High level of polysemy
  - Representativeness of the language
- Contexts: sentences from ISST where these words appear
- 210 words x 10 contexts (= 2,010) annotated with substitutes by 3 annotators
  - 300 trial
  - 1,710 test

Annotation

- Tool developed for Semeval'07

processo.n 1925 :: sviluppo 1;sistema 1;procedura 1;

processo.n 1926 :: giudizio 1;

Metrics

- Two scoring types:
  - Best → best substitute
  - out-of-ten (oot) → best 10 substitutes
- Metrics
  - Precision (P)
  - Recall (R)
  - F-measure
  - ModeP, ModeR → like P, R but consider only substitute by majority of annotators

Baseline

- No discrimination between contexts
- Semantic relations from LRs
  - Fetch all senses in LRs that correspond to word and extract synonyms and hyperonyms
  - Synonyms weight 3 times hyperonyms
  - If a word extracted more than once → sum weights
  - Lists of substitutes output ordered by weight
  - 3 runs:
    - PSC
    - IWN
    - PSC+IWN

Participants

System	Techniques, tools	Lexicons	Corpora
Basile_a	WSD	ItalWordNet	
Basile_b	n-grams	ItalWordNet	ItWaC
	IR	De Mauro	
De Cao et al.	LSA	ItalWordNet	Repubblica
	SVD	MultiWordNet	
Baroni et al.	WSM	None	Repubblica
	PoS-tagger		ItWac
	Lemmatiser		Italian Wikipedia

Results: best

Run	Precision	Recall	F	mode P	mode R
uniba2	**8.16**	**7.18**	**7.64**	10.58	10.58
baroniCutugnoLenciPucci	6.26	6.01	6.13	**11.28**	**10.84**
uniba1	6.80	5.53	6.10	8.90	8.90
uniba3	6.28	5.46	5.84	8.13	8.13
decao3	3.95	3.21	3.54	6.58	6.58
decao2	3.90	3.17	3.50	6.71	6.71
decao1	3.16	3.16	3.16	6.97	6.97
decao4	3.52	2.80	3.12	5.03	5.03
baseline_psc	**10.86**	**9.06**	**9.88**	**13.94**	**13.94**
baseline_iwn_psc	9.71	8.19	8.89	13.16	13.16
baseline_iwn	2.72	1.78	2.15	2.19	2.19

Results: oot

Run	Precision	Recall	F	mode P	mode R
uniba2	**41.46**	**36.50**	**38.82**	**47.23**	**47.23**
uniba1	37.74	30.69	33.85	34.84	34.84
uniba3	28.54	24.79	26.53	34.58	34.58
decao3	23.48	19.11	21.07	26.58	26.58
decao2	23.00	18.72	20.64	26.32	26.32
decao1	20.09	20.09	20.09	27.74	27.74
decao4	18.62	14.78	16.48	20.52	20.52
baroniCutugnoLenciPucci	16.65	16.00	16.32	24.97	24.00
baseline_iwn_psc	**27.52**	**23.23**	**25.19**	**37.24**	**32.39**
baseline_psc	23.00	19.20	20.93	26.97	26.97
baseline_iwn	14.51	9.51	11.49	12.77	12.77

Results: IT vs EN

Scoring type	Measure	Italian	English	Difference
best	P	10.86	12.90	18.78
	R	9.06	12.90	42.38
	mode P	13.94	20.73	48.70
	mode R	13.94	20.73	48.70
oot	P	41.46	69.03	66.50
	R	36.50	68.90	88.80
	mode P	47.23	66.26	40.30
	mode R	47.23	66.26	40.30

Conclusions, future

- Higher participation than WSD → higher interest?
- Different approaches presented
  - Complementary? Combine?
- Baselines: PSC much better than IWN
  - Some systems exploited IWN, replace by SIMPLE?
- Results far behind english → room for improvement

End

Thanks! Questions?

EVALITA 2009

The Lexical Substitution Task

A. Toral

Istituto di Linguistica Computazionale (CNR)