Wordnet-based Evaluation of Large Distributional Models for Polish

Maciej Piasecki, Gabriela Czachor, Arkadiusz Janz, Dominik Kaszewski, Paweł Kędzia

G4.19 Research Group, Department of Computational Intelligence
Wrocław University of Science and Technology, Wrocław, Poland

& CLARIN-PL `clarin-pl.eu`

Politechnika
Wrocławska

Agenda

- Wordnet-based tests for Distributional Semantics
- Synonymy tests
- Cut-off rendering tests
- Experiments
  - Corpora and preprocessing
  - Word embedding models tested
  - Tests based on *plWordNet*
  - Analogy tests (not wordnet-based)
- Results
- Conclusions and suggestions

Wordnet-based tests for Distributional Semantics

- Background
  - Distributional Semantics (DS) is focused on describing semantic associations between words on the basis of their distributional patterns in corpora
  - Corpora $\rightarrow$ *Measures of Semantic Relatedness* $\rightsquigarrow$? *Measure of Semantic Similarity*
  - *Word embeddings* are based on predicting a word occurrence in a context (mostly a sequence) of other words
  - A large wordnet is built on knowledge originating from humans
- Goals
  - to construct large scale test datasets for word embeddings on the basis of a large wordnet
  - to evaluate and compare different word embeddings extracted from a very large corpus of Polish
  - to publish: tests and word embeddings

Wordnet-based DS Evaluation

- Wordnet-based Similarity Measures $\longrightarrow$ Correlation of similarity rankings
  - but this comparison depends on a particular wordnet-based similarity measure applied
- *Wordnet-based Synonymy Test* (WBST)
- *Wordnet-based Cut-off Rendering Test* (WBCR)

Wordnet-based Synonymy Tests

- Proposed by (Freitag et al., 2005) following TOEFL synonymy tests:
  - for a *question word* x
  - an n-tuple is automatically generated:
    $\mathbf{D} = \langle d_1, \ldots d_n \rangle$,
    such that one the elements:
    - d_i is the correct *answer* – synonymous with x
    - all other $d_j \neq d_i$ are *detractors*, i.e. false answers, not synonymous with x
  - Elements of **D** and the position of the correct answer are randomly selected
- *Pros*: very large tests can be generated enabling very intensive testing
- *and cons*: numerous singleton synsets, too easy detractors

Wordnet-based Synonymy Tests

Hypernymy-expanded WBST (HWBST)

- Answers for singleton synsets are selected from their hypernym synsets
- Such hypernyms are excluded from possible detractors
- Examples of QA tuples:
    - *majątek* ‘property’:
      ⟨*okręt* ‘ship’, *uszanowanie* ‘respect’, ***mienie*** ‘property, assets’, *żywot* ‘≈life’⟩
    - *student* ‘student’:
      ⟨*momencik* ‘≈an indefinitely short time’, *łysina* ‘bald spot’, *skażenie* ‘contamination’, ***żak*** ‘≈student’⟩

Wordnet-based Synonymy Tests

Extended WBST (EWBST) (1)

- Idea: higher probability for the selection of detractors from synsets semantically similar to the question words
- EWBST consists of pairs: $\langle x_I, \mathbf{D}_I \rangle$, where
  - x_I is a question word,
  - $\mathbf{D}_I = \langle d_1, \ldots d_n \rangle$ such that
    - d_i is the correct answer, i.e. a synonym or hypernym of x_I, as in HWBST,
    - $d_j \in \mathbf{D}_I \wedge d_j \neq d_i$ are selected randomly from the whole wordnet but with the probability correlated to the *wordnet-based similarity measure* $WSM(d_j, x_I)$.

Wordnet-based Synonymy Tests

Extended WBST (EWBST) (2)

- WSM based on the normalised length of a shortest path in the wordnet graph (Agirre and Edmonds, 2006)

$$WSM(w_1, w_2) = -\log \frac{path(w_1, w_2)}{2D_m} \tag{1}$$

  - w_1 and w_2 are lemmas,
  - $path(w_1, w_2)$ is the shortest path in the extended hypernymy graph between two synsets including w_1 and w_2,
  - D_m is the maximum depth of the extended hypernymy graph.
- Modified

$$WSM_a(w_1, w_2) = max(-\log \frac{path(w_1, w_2)}{2D_a}, 0) \tag{2}$$

  - D_a is an average depth – promotes closer synsets

Wordnet-based Synonymy Tests

Extended WBST – Examples of QA tuples

- *majątek* 'property':
  ⟨***mienie*** 'property, assets', *banknot* 'banknote', *bon* 'voucher', *wyrównanie* 'compensation'⟩
- *student* 'student':
  ⟨*aspirant* '≈candidate', *licencjat* 'bachelor's degree', ***żak*** '≈student', *lektor* 'lector'⟩

Wordnet-based Cut-off Rendering Test

- Idea: to expand tests on other relations than synonymy and hyper/hyponymy
- For each question word x a bag-of-words of words is generated in which they come from:
  - the synset S_x of x
  - and synsets S_i connected directly and also indirectly to S_x by selected wordnet relations.
- Different *path definitions* can be used
- Evaluated MSR is used to reconstruct the extracted bag-of-words
  1. for a word x the *k-nearest neighbours* list $k\text{-}NNL(x)$ of the words most related to x according to MSR
  2. for the assumed k, the top k words from the list are collected as a reconstructed bag-of-words,
  3. and compared with the wordnet-based bag-of-words

Experiments

Corpora and preprocessing

- Wordnet
  - plWordNet 3.1 – a very large wordnet of Polish
  - 190,853 lemmas, 284,925 lexical units, 219,380 synsets and ≈650,000 relations
  - expresses very good coverage of words in large corpora
- plWordNet Corpus 10.0 (plWNC) of Polish
  - more than 4 billion words: several corpora supplemented with text acquired from the Web, only text in Polish, automated elimination of duplicates
- Corpora created from the Polish Wikipedia (of ≈ 600M words)

plWNC-lem
: morphosyntactically tagged, strings: "lemma:grammatical class" were in the input to *word2vec* (Mikolov et al., 2013)

plWNC-multi
: Proper Names and multiword expressions (60k) from plWordNet 3.1 merged as single tokens

Experiments

Word embedding models tested (1)

- Models generated by *word2vec* – *Gensim* library implementation
  1. vector size: 100, 300 and 1000,
  2. algorithm type: *Skip-gram*, *CBOW ns* (negative subsampling) and *CBOW hs* (with hierarchical softmax).
  3. tested models:
     Skip-gram 100, Skip-gram 300, Skip-gram 1000,
     CBOW ns 100, CBOW ns 300, CBOW ns 1000,
     CBOW hs 100, CBOW hs 300 and CBOW hs 1000
  4. minimal frequency of tokens: $\geqslant 8$ (min_count=8)
  5. freely available:
     `https://clarin-pl.eu/dspace/handle/11321/442`

Experiments

Word embedding models tested (2)

- *word2vec* models from literature
  - (Rogalski & Szczepaniak, 2016) built on Wikipedia
    - CBOW and Skip-gram models with negative sampling and the vector size: 300
    - text to lower case, numbers were divided into separate digits, and some non-text elements were deleted
  - (Mykowiecka et al., 2017)) on National Corpus of Polish
    - '*ncp*-lemmas' or '*ncp*-forms' – full data
    - "restricted data sets": only nouns, adjectives, adverbs, verb forms, and abbreviations
    - Skip-gram and CBOW architectures and the vector size of 100 and 300
- *fastText* models (words represented as n-grams)
  - (Bojanowski et al., 2016) Skip-gram models, vector size the vector size 300 for many languages on the basis of Wikipedia
  - *fastText.plWNC*: Skip-gram 300 models with min. word frequencies of 5, 20 and 50 built on the plWNC 10.0 Corpus

Experiments

Tests based on *plWordNet* (1)

- Wordnet-based Synonymy Tests
  - WBST, HWBST and EWBST
  - three versions corresponding to the minimal frequency of words in plWNC 10.0: 30, 200 and 1000
  - e.g.
    - EWBST(min. 1000) includes 19,996 question – answers pairs,
    - HWBST (min. 30) includes 48,263 pairs,
    - and WBST(min. 1000) includes 9,100 pairs (singleton synsets omitted)
  - freely available: `https://clarin-pl.eu/dspace/handle/11321/446`

Experiments

Tests based on *plWordNet* (2)

- Wordnet-based Cut-off Rendering Tests
  - three versions for to the minimal frequency of words in plWNC 10.0: 30, 200 and 1000
  - smaller numbers of bag of words, but still large data sets
  - types of paths for indirect links to the problem lemma x

    Cnt – only direct relation links (synset or lexical), including synonymy

    CntH – **Cnt** expanded with all indirect hyponyms and hypernyms of x up to the path length 3.

    CntHC – **CntH** expanded with all $k = m + n$ *cousins* of x with $k = 3$
  - freely available: `https://clarin-pl.eu/dspace/handle/11321/446`

Experiments

Analogy tests

- Most popular technique of the evaluation of word embeddings
  - testing MSR ability of reflecting word analogies
  - analogy consists of 2 pairs of words in a similar relation
  - MSR is used to find the best fitting lemma d in $(\vec{b} + \vec{c}) - \vec{a} = \vec{d}$
- Limitations
  - small size of a dataset – typically 200-300
  - potential polysemy of lemmas in pairs
- dataset of (Mykowiecka et al., 2017): ?200? analogy pairs from (?????) manually translated to Polish (but out of context)

Results – selected

Wordnet-based Synonymy Tests

Vector size	Min freq.	Model	WBST	HWBST	EWBST
1000	1000	w2w-*plWNC-multi*-skipg-ns	**92.43**	89.00	**63.97**
		w2w-*plWNC-multi*-cbow-hs	91.54	**89.34**	63.21
		w2w-*plWNC-multi*-cbow-ns	91.68	89.31	62.99
	200	w2w-*plWNC-multi*-skipg-ns	92.52	89.80	**62.51**
		w2w-*plWNC-multi*-cbow-hs	**92.71**	90.11	60.94
		w2w-*plWNC-multi*-cbow-ns	92.58	**90.11**	60.97
	30	w2w-*plWNC-multi*-skipg-ns	90.43	88.84	**58.92**
		w2w-*plWNC-multi*-cbow-hs	**92.56**	90.05	57.35
		w2w-*plWNC-multi*-cbow-ns	92.51	**90.07**	57.30
	1000	pl-embeddings-cbow	71.63	69.36	43.71
		pl-embeddings-skip	76.30	74.54	47.16
		fastText.wiki.pl	**80.01**	**78.17**	**52.42**
	200	pl-embeddings-cbow	71.79	69.46	42.31
		pl-embeddings-skip	76.89	74.65	45.53
		fastText.wiki.pl	**80.11**	**79.16**	**51.40**
	30	pl-embeddings-cbow	71.49	70.35	41.85
		pl-embeddings-skip	77.41	75.69	45.28
		fastText.wiki.pl	**81.44**	**80.27**	**51.39**

Results – selected

Wordnet-based Cut-off Rendering Test

Cut-off Precision

	k NN	10			100		
Model	Min. f.	Cnt	CntH	CntHC	Cnt	CntH	CntHC
w2w-*plWNC-multi*-cbow-hs	1000	13.42	15.12	**35.67**	**3.31**	**4.29**	**17.04**
w2w-*plWNC-multi*-cbow-ns	1000	**13.62**	**15.16**	34.25	3.30	4.22	15.96
w2w-*plWNC-multi*-skipg	1000	12.35	13.47	28.07	2.66	3.18	10.12
ft-*plWNC-multi*-skipg	1000	8.74	9.24	15.72	2.59	3.00	8.14
w2w-*plWNC-lem*-cbow-hs	1000	12.86	14.26	33.38	3.11	3.93	15.75
w2w-*plWNC-lem*-cbow-ns	1000	9.65	10.58	25.40	2.17	2.60	9.71
w2w-*plWNC-lem*-skipg	1000	11.61	12.61	27.15	2.47	2.92	9.82
ft-*plWNC-lem*-skipg	1000	7.39	7.72	13.31	2.25	2.54	7.25

Results – selected

Wordnet-based Cut-off Rendering Test

Cut-off Recall

	k NN	10			100		
		Cnt	CntH	CntHC	Cnt	CntH	CntHC
w2w-*plWNC-multi*-cbow-hs	1000	**10.33**	**7.10**	**3.42**	**20.83**	**15.69**	**8.61**
w2w-*plWNC-multi*-cbow-ns	1000	10.09	6.84	3.24	20.27	14.84	8.16
w2w-*plWNC-multi*-skipg	1000	9.24	6.26	2.91	17.22	12.20	6.26
ft-*plWNC-multi*-skipg	1000	7.33	4.87	2.18	17.54	12.22	5.80
w2w-*plWNC-lem*-cbow-hs	1000	8.74	6.05	2.85	17.67	13.03	7.03
w2w-*plWNC-lem*-cbow-ns	1000	6.71	4.61	2.18	13.20	9.46	4.99
w2w-*plWNC-lem*-skipg	1000	8.19	5.64	2.60	15.12	10.82	5.41
ft-*plWNC-lem*-skipg	1000	5.92	4.04	1.82	14.88	10.40	4.85

Results – selected

Wordnet-based Cut-off Rendering Test

F measure

	k NN	10			100		
		Cnt	CntH	CntHC	Cnt	CntH	CntHC
w2w-*plWNC-multi*-cbow-hs	1000	**11.67**	**9.66**	**6.23**	**5.72**	**6.74**	**11.44**
w2w-*plWNC-multi*-cbow-ns	1000	11.59	9.42	5.92	5.68	6.57	10.80
w2w-*plWNC-multi*-skipg	1000	10.57	8.55	5.27	4.61	5.05	7.73
ft-*plWNC-multi*-skipg	1000	7.97	6.38	3.83	4.51	4.82	6.77
w2w-*plWNC-lem*-cbow-hs	1000	10.41	8.49	5.25	5.29	6.04	9.72
w2w-*plWNC-lem*-cbow-ns	1000	7.91	6.42	4.02	3.73	4.08	6.59
w2w-*plWNC-lem*-skipg	1000	9.60	7.80	4.75	4.24	4.60	6.98
ft-*plWNC-lem*-skipg	1000	6.57	5.30	3.20	3.90	4.09	5.81

Results – selected

Wordnet-based Cut-off Rendering Test

F measure

k NN		10			100		
		Cnt	CntH	CntHC	Cnt	CntH	CntHC
ncp-forms-rest-cbow-ns	1000	6.13	5.13	3.26	3.30	3.95	6.88
ncp-lemmas-all-cbow-hs	1000	9.62	**8.08**	**5.07**	**4.91**	**5.78**	**9.46**
ncp-lemmas-all-cbow-ns	1000	**9.72**	7.91	4.80	5.01	5.74	9.04
ncp-lemmas-all-skipg-hs	1000	8.64	7.08	4.36	4.30	4.88	7.84
ncp-lemmas-all-skipg-ns	1000	8.42	6.88	4.18	3.75	4.07	6.20
ncp-lemmas-rest-cbow-hs	1000	9.91	8.27	5.05	4.99	5.82	9.20
ncp-lemmas-rest-cbow-ns	1000	10.03	8.13	4.80	5.06	5.75	8.77
ncp-forms-rest-cbow-ns	200	5.29	4.51	3.35	2.62	3.14	6.67
ncp-lemmas-all-cbow-hs	200	**8.71**	**7.52**	**5.54**	4.00	**4.75**	**9.44**
ncp-lemmas-all-cbow-ns	200	8.55	7.13	5.12	**4.01**	4.61	8.98
ncp-lemmas-all-skipg-hs	200	8.01	6.79	4.96	3.58	4.11	8.13
ncp-lemmas-all-skipg-ns	200	7.30	6.11	4.29	3.00	3.28	6.00
ncp-lemmas-rest-cbow-hs	200	9.02	7.75	5.56	4.07	4.80	9.20
ncp-lemmas-rest-cbow-ns	200	8.87	7.38	5.17	4.07	4.65	8.79

Results – selected

Wordnet-based Cut-off Rendering Test

F measure

	k NN	10			100		
		Cnt	CntH	CntHC	Cnt	CntH	CntHC
pl-embeddings-cbow	1000	**3.79**	**3.31**	**2.15**	2.32	**2.94**	**5.08**
pl-embeddings-skipg	1000	3.35	2.82	1.80	2.15	2.56	4.20
fastText.wiki.pl	1000	3.52	2.83	1.70	**2.63**	2.81	4.12
pl-embeddings-cbow	200	3.26	2.90	2.11	1.86	2.38	**4.79**
pl-embeddings-skipg	200	3.01	2.63	1.89	1.77	2.15	4.17
fastText.wiki.pl	200	**3.82**	**3.14**	**2.16**	**2.31**	**2.48**	4.45
pl-embeddings-cbow	30	2.87	2.58	1.97	1.58	2.03	4.44
pl-embeddings-skipg	30	2.80	2.50	1.93	1.55	1.90	4.11
fastText.wiki.pl	30	**3.89**	**3.24**	**2.41**	**2.06**	**2.22**	**4.53**

Results – selected

Analogy Tests

Model	*VS*	Score	Model	*VS*	Score
w2w-*plWNC-multi*-cbow-hs	100	40.82	w2w-*plWNC-multi*-cbow-ns	300	**57.14**
w2w-*plWNC-lem*-cbow-ns	100	47.96	w2w-*plWNC-lem*-skipg	300	**60.20**
ft-*plWNC-multi*-skipg-mC20	300	**53.30**	ft-*plWNC-lem*-skipg-mC20	300	54.23
ft-*plWNC-multi*-skipg-mC50	300	50.75	ft-*plWNC-lem*-skipg-mC50	300	**59.28**
ncp-lemmas-all-300-cbow-ns	300	**57.95**	*ncp*-forms-all-300-cbow-ns	300	43.18
ncp-lemmas-all-300-skipg-ns	300	54.36	*ncp*-forms-all-300-skipg-ns	300	**46.82**
ncp-lemmas-rest-300-cbow-ns	300	**59.49**	*ncp*-forms-rest-300-cbow-ns	300	43.64

Conclusions and Further Works

- EWBST is the hardest synonymy tests and its difficulty can be tuned with the help of a wordnet-based similarity measure
- Skip-gram model is better than CBOW according to WBST and EWBST
  - only better performance of CBOW-ns in HWBST can be attributed to a kind of generalisation caused by hypernyms in answers
  - also models from literature based on Skip-gram scheme , including *fastText.wiki.pl* express higher results
- CBOW models are superior in all cases in comparison to Skip-gram models in WBCRT
  - models with merged MWEs and PNs are better than those based on lemmas
- Skip-gram models are better in describing meaning differences, while CBOW enable broader exploration of potential lexico-semantic relations

Conclusions and Further Works

- WCBRT can be used also as a diagnostic tool to spot worse described subdomains
- Hierarchical softmax consistently produces better results in all frequency ranges
- Smaller corpora
  - all results are much worse than those obtained on plWNC 10 corpus
  - the models behave in a slightly different way in WBCR tests
  - Skip-gram models express higher recall, especially fastText Skip-gram with sub-word representation
- In analogy tests Skip-gram model built on plWNC 10 is still the best one, but the difference to models built on much smaller NCP is minimal
  - the analogy tests of include mostly general and frequent words
  - the differences are small only for models based on the restricted version of NCP

Conclusions and Further Works

- A large comprehensive wordnet can be successfully used as a basis for two different types of MSR evaluation methods
- The datasets are enough large to conveniently partitioned according to the frequency criteria of semantic criteria
- the datasets and tests are based on human decisions expressed in the wordnet structure
- We plan
  - to develop a wordnet-based test that has properties of contextual tests
  - and tests covering all four PoS

Thank you very much for your attention!

http://clarin-pl.eu

http://nlp.pwr.edu.pl

http://plwordnet.pwr.edu.pl