ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

Ερευνητική ομάδα «Εξόρυξης Γνώσης από Βάσεις Δεδομένων και τον Παγκόσμιο Ιστό», Οικονομικό Πανεπιστήμιο Αθηνών

Αξιολόγηση ενθέσεων των ελληνικών λέξεων

Σ. Ούτσιος, Χ. Καράτσαλος, Κ. Σκιάνης, Μ. Βαζιργιάννης

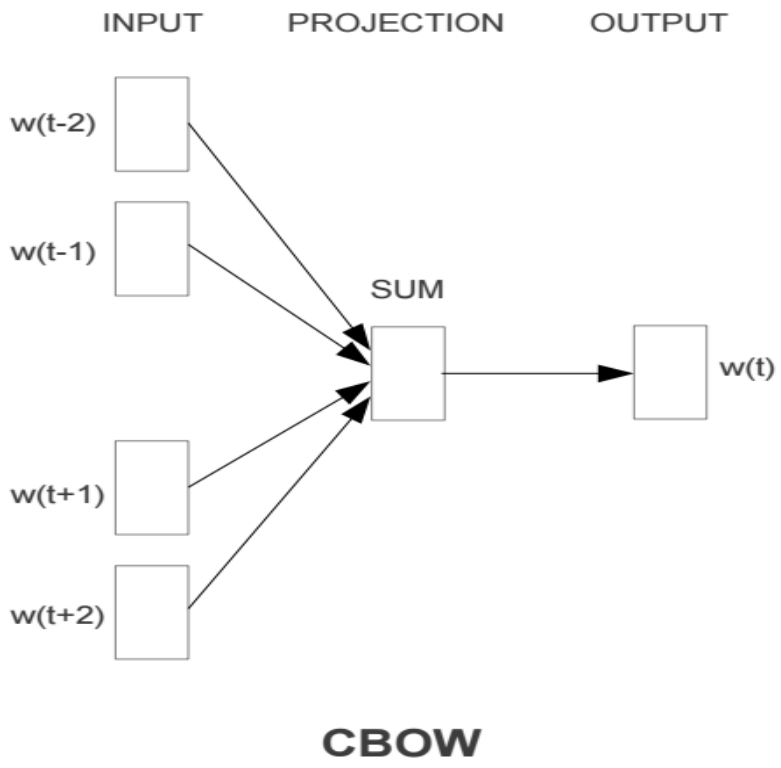
Περιγραφή

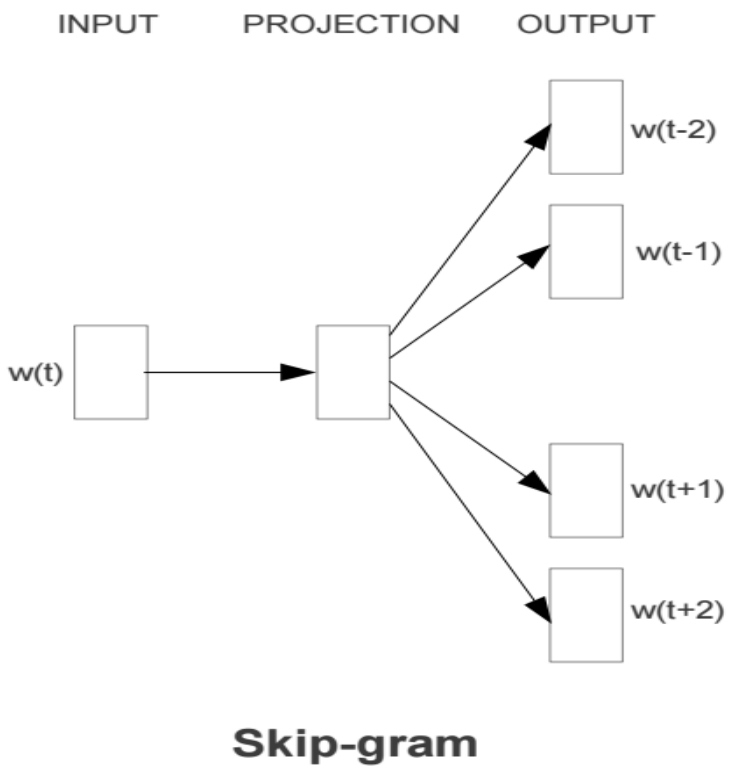
- Ενθέσεις λέξεων
- Μέθοδοι εκπαίδευσης μοντέλων CBOW & Skip-gram
- Αξιολόγηση μοντέλων ελληνικών λέξεων – Συνεισφορά
- Ερωτήματα αναλογιών
- Αποτελέσματα
- Νεοφυείς λέξεις της ελληνικής γλώσσας
- Μελλοντικοί στόχοι

Ενθέσεις λέξεων

- Διανυσματικές αναπαραστάσεις λέξεων
- Παράδειγμα:
  - Γάτα: (1.2, -0.4, 3, 8) - Διάνυσμα διάστασης 4
  - Προκύπτει μετά από εκπαίδευση μοντέλων με μεθόδους που δέχονται ως είσοδο μεγάλο όγκο κειμένων.
- *«Πες μου τη χρήση μιας λέξης να σου πω τι σημαίνει.»* **Βιτγκενστάιν**

Μέθοδοι εκπαίδευσης μοντέλων CBOW & Skip-gram (1)





Μέθοδοι εκπαίδευσης μοντέλων CBOW & Skip-gram (2)

- Στο CBOW προβλέπουμε την λέξη σύμφωνα με το περιεχόμενο
- Στο Skip-Gram προβλέπουμε το περιεχόμενο σύμφωνα με τη λέξη εισόδου
- Μέθοδοι που βασίζονται στα Νευρωνικά Δίκτυα
- Μη επιτηρούμενη μάθηση (unsupervised learning)

Αξιολόγηση μοντέλων ελληνικών λέξεων

- Στη βιβλιογραφία ασχολούνται κυρίως με την Αγγλική γλώσσα
- Δεν υπάρχει παρόμοια δουλειά για την Ελληνική γλώσσα
- Σύγκριση 7 μοντέλων, εκ των οποίων τα 5 έχουν εκπαιδευτεί από δεδομένα μεγάλης κλίμακας τα οποία έχει συλλέξει το εργαστήριο «Εξόρυξης Γνώσης από Βάσεις Δεδομένων και τον Παγκόσμιο Ιστό»

- Μέθοδοι Αξιολόγησης:
  - Ερωτήματα Αναλογιών
  - Συγκρίσεις ομοιότητας κλπ..

Ερωτήματα αναλογιών (1)

- Σημασιολογικές και Συντακτικές κατηγορίες
  - Περιλαμβάνουν υποκατηγορίες, όπως
    - : common_capital_country

      λισαβόνα πορτογαλία βιέννη αυστρία

      κοπεγχάγη δανία βιέννη αυστρία
    - : man_woman_family

      αγόρι κορίτσι μπαμπάς μαμά

      ξάδερφος ξαδέρφη πατέρας μητέρα

- Παράδειγμα:
  - Ελλάδα-Αθήνα, Γερμανία-? => Βερολίνο!

Ερωτήματα αναλογιών (2)

Relation	#pairs	#tuples
Semantic: (13650 tuples)		
common_capital_country	42	1722
all_capital_country	78	6006
eu_city_country	50	2366
city_in_region	40	1536
currency_country	24	552
man_woman_family	18	306
profession_placeof_work	16	240
performer_action	24	552
politician_country	20	370
Syntactic: (25524 tuples)		
man_woman_job	26	650
adjective_adverb	28	756
opposite	35	1190
comparative	36	1260
superlative	25	600
present_participle_active	48	2256
present_participle_passive	44	1892
nationality_adjective_man	56	3080
nationality_adjective_woman	42	1722
past_tense	34	1122
plural_nouns	72	5112
plural_verbs	37	1332
adjectives_antonyms	50	2450
verbs_antonyms	20	380
verbs_i_you	42	1722

Ερωτήματα αναλογιών (3) - Παραδείγματα

Analogy

Πράξεις μεταξύ των διανυσμάτων των λέξεων

Εδώ βλέπουμε πώς απλές γραμμικές πράξεις μεταξύ των διανυσμάτων των λέξεων μπορούν να παράγουν αποτελέσματα τα οποία έχουν νόημα.

ελλάδα - αθήνα + βερολίνο →

γερμανία,0.756761908531189
ευρώπη,0.6723400354385376
ηγερμανία,0.6530500650405884
πολωνία,0.6224887371063232
γερμανίαΰ,0.6130441427230835

Ερώτημα

Analogy

Πράξεις μεταξύ των διανυσμάτων των λέξεων

Εδώ βλέπουμε πώς απλές γραμμικές πράξεις μεταξύ των διανυσμάτων των λέξεων μπορούν να παράγουν αποτελέσματα τα οποία έχουν νόημα.

φθηνότερο - φθηνό + ακριβό →

ακριβότερο,0.827736496925354
ακριβότερα,0.7229090929031372
κριβότερο,0.7002600431442261
ακριβότερη,0.6856571435928345
φτηνότερο,0.6727663278579712

Ερώτημα

Αποτελέσματα (1)

Category		gr_def	gr_neg10	cc.el.300	wiki.el	gr_cbow_def	gr_d300_nosub	gr_w2v_sg_n5
Semantic	no oov words	58.42%	59.33%	**68.80%**	27.20%	31.76%	60.79%	52.70%
	with oov words	52.97%	55.33%	**64.34%**	25.73%	28.80%	55.11%	47.82%
Syntactic	no oov words	65.73%	61.02%	**69.35%**	40.90%	64.02%	53.69%	52.60%
	with oov words	**53.95%**	48.69%	49.43%	28.42%	52.54%	44.06%	43.13%
Overall	no oov words	63.02%	59.96%	**68.97%**	36.45%	52.04%	56.30%	52.66%
	with oov words	53.60%	51.00%	**54.60%**	27.50%	44.30%	47.90%	44.80%

Αποτελέσματα (2)

Category	Models						
	gr_def	gr_neg10	cc.el.300	wiki.el	gr_cbow_def	gr_d300_nosub	gr_w2v_sg_n5
all_capital_country	61.78%	65.26%	75.42%	25.03%	26.50%	64.50%	52.40%
city_in_region	53.60%	53.00%	47.70%	10.20%	32.40%	61.30%	62.20%
common_capital_country	69.57%	73.48%	83.72%	43.29%	39.80%	72.10%	61.00%
currency_country	17.00%	19.00%	22.90%	3.20%	7.60%	17.00%	17.80%
eu_city_country	60.40%	64.00%	78.30%	48.60%	33.10%	60.40%	52.20%
man_woman_family	75.83%	76.25%	87.62%	12.08%	75.00%	72.10%	70.80%
performer_action	19.00%	16.49%	22.53%	1.78%	26.10%	21.60%	18.50%
politician_country	66.33%	62.30%	70.80%	21.00%	52.40%	56.00%	58.70%
profession_place_of_work	39.52%	38.57%	59.89%	14.29%	53.30%	47.60%	50.50%
Semantic	58.42%	59.33%	**68.80%**	27.20%	31.76%	60.79%	52.70%
adjective_adverb	30.34%	23.50%	34.90%	8.17%	38.60%	21.40%	25.80%
adjective_antonym	25.40%	23.40%	31.20%	13.10%	20.80%	23.20%	22.70%
comparative	88.18%	75.49%	73.54%	58.50%	84.90%	61.00%	62.40%
verbs_i_you	97.05%	94.49%	98.11%	83.33%	95.30%	90.30%	89.60%
man_woman_job	83.98%	83.33%	73.10%	39.18%	94.60%	77.30%	80.70%
nationality_adjective_man	81.20%	76.20%	83.67%	62.34%	73.10%	62.10%	53.10%
nationality_adjective_woman	61.40%	51.08%	58.40%	38.17%	45.80%	37.50%	35.20%
opposite	36.59%	30.65%	48.15%	24.87%	37.50%	27.60%	25.70%
past_tense	83.71%	76.52%	80.69%	3.16%	89.60%	75.50%	79.00%
plural_nouns	56.22%	50.02%	63.92%	37.10%	46.50%	49.40%	47.90%
plural_verbs	98.50%	97.75%	99.52%	85.54%	98.60%	95.90%	95.60%
present_participle_(active)	82.88%	72.15%	91.96%	67.90%	96.90%	59.20%	63.40%
present_participle_(passive)	44.90%	21.50%	33.46%	43.64%	80.80%	14.60%	16.20%
superlative	61.11%	18.06%	58.33%	50.00%	80.60%	9.70%	5.60%
verbs_antonyms	20.90%	13.20%	19.90%	4.40%	11.50%	14.30%	13.20%
Syntactic	65.73%	61.02%	**69.35%**	40.90%	64.02%	53.69%	52.60%
Overall	63.02%	59.96%	**68.97%**	36.45%	52.04%	56.30%	52.66%
Questions with oov words	14.90%	14.90%	20.80%	24.60%	14.90%	14.90%	14.90%

Νεοφυείς λέξεις της ελληνικής γλώσσας

- Προέκυψαν από την εργασία μας ~80.000 λέξεις εκτός δημοσιευμένων ελληνικών λεξικών
- Αναλύοντας δεδομένα κειμένων μεγάλης κλίμακας από ελληνικές ιστοσελίδες
  - *Μακαρονόδεντρα*
  - *Μπλογκοχώρι*
  - *Γυρεοδοτικά*
  - *Ιντριγκάρω*

Μελλοντικοί στόχοι

- Εκπαίδευση νέων μοντέλων με διαφορετικές παραμέτρους
- Αξιολόγηση σε συγκεκριμένα προβλήματα της Επεξεργασίας Φυσικής Γλώσσας π.χ. ταξινόμηση κειμένων, επισήμανση μερών του λόγου, γλωσσικά μοντέλα