RUSSIAN COLLOCATION EXTRACTION BASED ON WORD EMBEDDINGS

Enikeeva E. V. (protoev@yandex.ru)
Mitrofanova O. A. (o.mitrofanova@spbu.ru)
Saint Petersburg State University

Collocation (extraction)

- Unformalized definition
- No classification according to meaning or semantic structure
- Simple frequency-based methods

---

raw word combination lists

Collocation

- «A collocation is a succession of two or more words that must be learnt as an integral whole and not pieced together from its component parts». (Palmer 1933)
- distinguished from idiom:
  - еловая, сосновая, кедровая, etc. шишка ‘fir, pine, cedar, etc. cone’ vs. важная шишка ‘boss’
  - бить тревогу, рекорд, поклоны, etc. ‘sound the alarm, beat the record, beat bows, etc.’ vs. бить баклуши ‘twiddle’

Collocation

«A collocation **AB** of language **L** is a semantic phraseme of **L** such that its signified ‘X’ is constructed out of the signified of the one of its two constituent lexemes — say, of **A** — and a signified ‘C’ [‘X’ = ‘A ⊕ C’] such that the lexeme **B** expresses ‘C’ contingent on **A**» (Mel’čuk 1998)

⇒ Lexical Function

Lexical function

- «A lexical function **f** associates with a word **w**, called its argument or key word, the set of words and phrases, which express - contingent on **w** - the meaning of role which corresponds to **f**.» (Mel'čuk 1984)
- lexical relation
- both paradigmatic and syntagmatic

Lexical Function

Examples:

- **Magn** - high degree

высокое напряжение ‘high voltage’, значительная высота ‘considerable height’

- **Bon** - standard praise for **w**

аккуратно резать ‘cut neatly’, комфортабельное судно ‘comfortable ship’

- **Oper$_i$** - an action for which the the i-th participant of situation is expressed by the subject

Oper$_1$: оказывать сопротивление ‘show resistance’

Oper$_2$: встречать сопротивление ‘meet resistance’

Method

Rodríguez-Fernández, S., Anke, L., Carlini, R., Wanner, L. (2016) Semantics-driven recognition of collocations using word embeddings.

training and test samples drawn from **Macmillan Collocations Dictionary**

assumption: headword and collocate embeddings should be trained on different corpora

Method

Argument matrix: $$A_T = \left[a_{t_1}, \ldots, a_{t_n} \right]$$

Collocate matrix: $$C_T = \left[c_{t_1}, \ldots, c_{t_n} \right]$$

A linear transformation matrix from training set .

$$A_T \Psi_T = C_T$$

approximated using SVD to minimize the sum:

$$\sum_{i=1}^{|T|} \left\| \Psi_T a_{t_i} - c_{t_i} \right\|^2$$

Test settings

Training and test collocations extracted from

- SynTagRus Treebank (http://www.ruscorpora.ru/instruction-syntax.html)

Reference data:

- Verbal collocations of Russian abstract nouns dictionary (http://dict.ruslang.ru/abstr_noun.php)

Vector models:

- RusVectōrēs project (http://rusvectores.org, version 3)

Test settings

LF	**argument**	**value**
OPER1	цель ‘aim’	иметь ‘have’
MAGN	каблук ‘heel’	высокий ‘high’
CAUSFUNC0	соревнование ‘competition’	проводить ‘hold’
FUNC0	открытие ‘opening’	состояться ‘be held’
INCEPOPER1	работа ‘work’	приступать ‘start’
OPER2	правка 'correction'	подвергаться ‘undergo’
REAL1-M	ракета ‘rocket’	запускать ‘launch’
REAL1	средства ‘means’	расходовать ‘spend’
INCEPFUNC0	речь ‘conversation’	заходить ‘turn to’

Test settings

9 lexical functions

10 headwords for each LF

10 top-ranked collocates for each headword

Metrics: precision(10), recall, MRR

Filtering: UD POS-tags, NPMI score

Test settings

- M1 – baseline: cos(c_i , ai – c_i + a_j), where (a_i, c_i) is an example collocation for a given LF and a_j is a test headword;
- M2 –baseline filtered by POS tags and NPMI scores;
- M3 – proposed model, same vector spaces for headwords and collocates trained on RNC;
- M4 – M3 filtered by POS tags and NPMI scores;
- M5 –M3, collocate vectors from Russian Wikipedia;
- M6 – M5 filtered by POS tags and NPMI scores.

Results (precision)

LF	*M1*	*M2*	*M3*	*M4*	*M5*	*M6*
OPER1	0.11	0.31	0.10	0.14	0.37	**0.63**
MAGN	0.23	0.28	0.24	0.28	0.63	**0.84**
CAUSFUNC0	0.10	0.33	0.22	0.23	0.54	**0.64**
FUNC0	0.21	0.40	0.29	0.33	**0.42**	**0.42**
INCEPOPER1	0.10	0.38	**0.64**	**0.64**	0.15	0.15
OPER2	0.17	0.28	0.12	0.11	0.29	**0.39**
REAL1-M	0.20	**0.66**	0.24	0.26	0.40	0.52
REAL1	0.15	0.37	0.32	0.33	**0.66**	**0.66**
INCEPFUNC0	0.13	0.28	0.24	0.23	0.35	**0.43**

Results (recall)

LF	*M1*	*M2*	*M3*	*M4*	*M5*	*M6*
OPER1	0.50	0.50	0.48	**0.65**	0.57	**0.65**
MAGN	0.68	0.70	0.70	0.70	**0.83**	0.78
CAUSFUNC0	0.33	0.45	0.87	**0.9**	0.80	0.80
FUNC0	0.70	0.80	**0.90**	0.81	0.58	0.58
INCEPOPER1	0.40	**0.55**	0.50	0.50	0.50	0.50
OPER2	0.55	0.60	0.53	0.52	**0.75**	0.70
REAL1-M	0.55	0.70	**0.87**	**0.87**	0.70	0.75
REAL1	0.40	0.45	**0.73**	0.70	0.60	0.60
INCEPFUNC0	0.50	0.70	0.72	0.67	**0.82**	0.77

Results (MRR)

LF	*M1*	*M2*	*M3*	*M4*	*M5*	*M6*
OPER1	0.22	0.55	0.11	0.48	0.34	**0.73**
MAGN	0.30	0.68	0.30	0.68	0.50	**0.90**
CAUSFUNC0	0.11	0.43	0.37	0.76	0.41	**0.82**
FUNC0	0.30	0.82	0.47	**0.89**	0.36	0.64
INCEPOPER1	0.11	0.60	0.01	**0.64**	0.15	0.48
OPER2	0.19	0.64	0.23	0.48	0.37	**0.66**
REAL1-M	0.08	0.77	0.36	0.70	0.34	**0.86**
REAL1	0.23	0.50	0.37	**0.70**	0.30	0.59
INCEPFUNC0	0.10	0.64	0.37	**0.73**	0.42	0.72

Results (discussion-1)

- NPMI filtering discards relevant examples

MAGN(*медаль*) = *золотой* …

- training on different corpora — lower scores for specific LFs

- frequent LFs — more correct collocates in top-10

Results (discussion-2)

MAGN(довод)= *решительный,* ***убедительный,*** *основательный, веский, главный, бесспорный, достаточный…*

OPER1(домино) = ***играть,*** *поиграть, стучать, резаться, сыграть, игра, бильярд, футбол…*

INCEPFUNC0 (день) = *наступать,* ***наставать,*** *начинаться, приходить, прийти, намечаться, длиться, заканчиваться…*

Results (discussion-3)

INCEPOPER1(азарт) = входить	*<u>приходить,</u> игра, увлекаться…*
OPER2(арест) = сидеть	*<u>подвергаться,</u> <u>находиться,</u> подвергать, брать, миновать, попадать…*
FUNC0 (дорога) = проходить	*<u>идти,</u> <u>пойти,</u> <u>тянуться,</u> <u>лежать,</u> плестись, тащиться …*

Thank you!