Анна Зуева
НИУ ВШЭ, ФиКЛ, БКЛ-151
Научный руководитель: С. Ю. Толдова

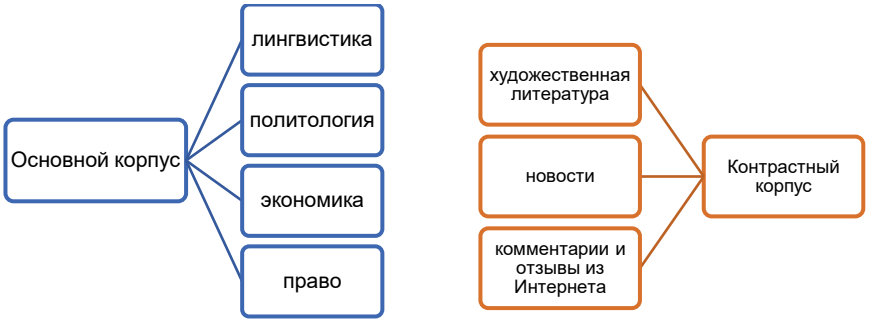
Автоматическое составление словарного минимума для изучения РКИ: извлечение общеупотребительной в научной речи лексики

Цель:

нахождение оптимальных автоматических методов извлечения общеупотребительной в научном стиле лексики (из текстов научной и учебной литературы).

→ Насколько эффективным будет использование методов выделения ключевых слов и терминов для извлечения общенаучной лексики?

Данные:



1 970 426 словоупотреблений

1 165 252 словоупотреблений

Этапы:

1. Предобработка данных
2. Выделение кандидатов из коллекции документов на основе лингвистических фильтров и частотности:
   - 7 морфологических шаблонов: Verb + Noun, Noun + Verb, Prep + Noun, Noun + Prep, Verb + Prep, Adj + Noun, Adv + Verb
   - биграммы встречаются не менее, чем в 6 документах коллекции
3. Вычисление признаков, по которым будет отранжирован список кандидатов:
   - **TF** (Term Frequency): подсчет частотности выражений
   - **TF-IDF** (TF — term frequency, IDF — inverse document frequency) с использованием внешнего корпуса:

$$\mathrm{TF} \cdot \mathrm{IDF(t)} = TF(t) \cdot \log \frac{1}{TF_r(t)},$$

     где $TF_r(t)$ — количество документов внешнего корпуса, в которых содержится кандидат *t*.

   - **t-критерий Стьюдента** (Браславский, Соколов 2006):

$$\text{t-score } (w_1, w_2) = \frac{P(w_1 w_2) - P(w_1)P(w_2)}{\sqrt{\frac{P(w_1 w_2)}{N}}},$$

     где $P(w_1w_2)$ — вероятность появления биграммы, $P(w_1)$ — вероятность появления первого слова из биграммы, $P(w_2)$ — вероятность появления второго слова из биграммы, *N* — общее количество биграмм.

- **C-Value и NC-Value** (Frantzi et al. 2000):

$$\text{C-Value(t)} = \begin{cases} log_2|t| \cdot f(t), \text{ если } \{s : t \subset s\} = \emptyset; \\ log_2|t| \cdot f(t) - \frac{\sum_s f(s)}{|s: t \subset s|}, \text{ иначе,} \end{cases}$$

где *t* — кандидат в термины, |*t*| — количество слов в *t*, *f(t)* — частота встречаемости *t* в коллекции текстов, *s* — множество кандидатов, в состав которых входит *t.*

$$weight(w) = \frac{t(w)}{n},$$

где *w* — контекстное слово, *t(w)* — количество терминов, с которыми встречается w, *n* — общее количество терминов, для которых подсчитывается NC-Value.

$$\text{NC-Value(t)} = 0.8 \cdot \text{C-Value(t)} + 0.2 \cdot \sum_{w \in C(t)} f_t(w) \cdot weight(w),$$

где *t* — рассматриваемый кандидат в термины, *C(t)* — множество контекстных слов для *t*, *w* — контекстное слово из *C(t)*, $f_t(w)$ — количество употреблений *w* в качестве контекстного слова для *t*, *weight(w)* — вес слова *w*.

- **Weirdness** (Ahmad et al. 1999):

$$\text{Weirdness(t)} = \frac{TF_{target}(t) \cdot |Corpus_{reference}|}{TF_{reference}(t) \cdot |Corpus_{target}|},$$

где $TF_{target}(t)$ — частота кандидата *t* в корпусе предметной области, $TF_{reference}(t)$ — частота кандидата *t* во внешнем корпусе, $|Corpus_{target}|$ — число слов в корпусе предметной области и $|Corpus_{reference}|$ — число слов во внешнем корпусе.

- **показатель G2 от LogLikelihood**:

$$G2 = 2(a \ln \frac{a}{E_1} + b \ln \frac{b}{E_2}),$$

где *a* — частота кандидата в термины в рассматриваемом корпусе, *b* — частота кандидата в термины в контрастном корпусе, E_1 — ожидаемая частота для кандидата в термины в рассматриваемом корпусе, E_2 — ожидаемая частота для кандидата в термины в контрастном корпусе.

4. Сортировка кандидатов по значению вычисленных признаков и отбор нужного количества кандидатов.

Таблица 1. Топ-10 кандидатов каждого морфологического шаблона с использованием лучшего метода для каждого из них

Adj + Noun (EA)	Prep + Noun (Weirdness)	Noun + Prep (NC-Value)	Verb + Noun (NC-Value)	Noun + Verb (TF)	Verb + Prep (NC-Value)	Adv + Verb (Weirdness)
российский_a федерация_s	в_pr промышленность_s	зависимость_s от_pr	представлять_v себя_spro	речь_s идти_v	привести_v к_pr	можно_adv отметить_v
данный_a случай_s	на_pr продукция_s	право_s на_pr	иметь_v место_s	что_s касаться_v	зависеть_v от_pr	часто_adv использоваться_v
государственный_a власть_s	со_pr ст_s	вопрос_s о_pr	давать_v возможность_s	значение_s иметь_v	приводить_v к_pr	широко_adv использоваться_v
общественный_a отношение_s	на_pr товар_s	язык_s в_pr	обратить_v внимание_s	роль_s играть_v	относиться_v к_pr	непосредственно_adv зависеть_v
федеральный_a закон_s	к_pr осуществление_s	влияние_s на_pr	представлять_partcp себя_spro	место_s занимать_v	основать_partcp на_pr	отдельно_adv взять_partcp
правовой_a акт_s	к_pr рассмотрение_s	изменение_s в_pr	добавить_partcp стоимость_s	внимание_s уделяться_v	состоять_v в_pr	справедливо_adv отмечать_v
второй_a половина_s	от_pr уровень_s	участие_s в_pr	оказывать_v влияние_s	государство_s мочь_v	говорить_v о_pr	можно_adv выделить_v
государственный_a орган_s	в_pr отрасль_s	цена_s на_pr	мочь_v стать_s	государство_s быть_v	заключаться_v в_pr	постоянно_adv проживать_partcp
настоящий_a время_s	на_pr труд_s	роль_s в_pr	осуществлять_partcp функция_s	человек_s мочь_v	исходить_v из_pr	можно_adv отнести_v
составной_a часть_s	на_pr изменение_s	спрос_s на_pr	принимать_v участие_s	государство_s являться_v	вести_v к_pr	можно_adv утверждать_v

Оценка качества

- принадлежат ли извлеченные биграммы к общенаучной лексике? (3 ассесора)
- вычисление Average precision at K (ap@k) по формуле:

$$ap@K = \frac{1}{K}\sum_{k=1}^{K} p@k,$$

где p@k — доля релевантных элементов среди первых k выражений из отранжированного списка, K — количество рассматриваемых элементов.

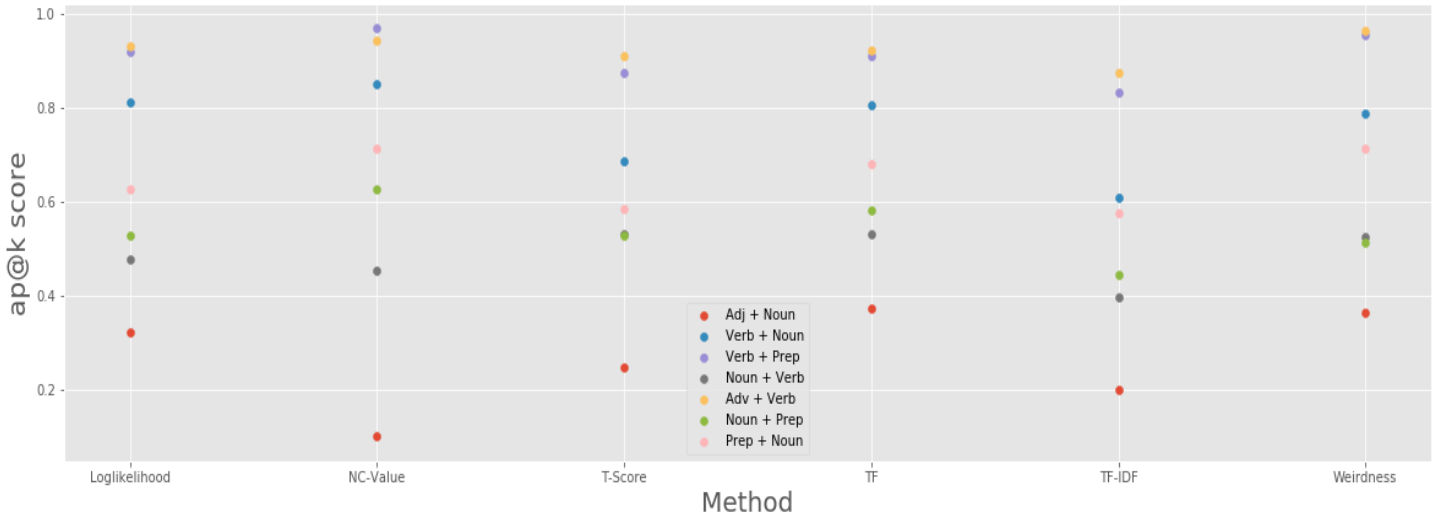
Учитывается как количество релевантных выражений в выбранном интервале списка, так и порядок элементов: чем выше в списке стоит релевантное выражение, тем больший вклад в итоговое значение оно сделает.

Таблица 2. Значения *ap@k*

	Adj + Noun	**Noun + Prep**	**Prep + N**	**Verb + Prep**
tf	**0.37490831399585967**	0.5816111453757823	0.6816396641378724	0.9112032989054821
td-idf	0.20088753826520656	0.4445182839488938	0.5755724542634985	0.8324139738449475
t-score	0.24783058436991715	0.5288081491353127	0.5846008221014488	0.8744618868199983
loglikelihood	0.3217767965941056	0.5285204868595963	0.6261972895899153	0.9183598325550377
weirdness	0.3654490877676984	0.5147345360123712	**0.7141955927925614**	0.9554349616784723
nc-value	0.10219835058101555	**0.626678117348358**	0.7133991256873199	**0.9694505982118574**

	Adv + Verb	**Noun + Verb**	**Verb + Noun**
tf	0.92204602185819	**0.531840829800309**	0.8067903898581485
td-idf	0.8759060540020487	0.39686903093068293	0.6105339556979353
t-score	0.9092316345194285	0.530519558534079	0.6868396144470746
loglikelihood	0.9314744727315442	0.47688448290958896	0.8131418609713935
weirdness	**0.9650717966155841**	0.5267891562589788	0.7882070558463861
nc-value	0.9430478528666051	0.45468173197694783	**0.8508097912337191**

График 1. Значения *ap@k* в зависимости от метода и морфологического шаблона



Выводы

- Самым лучшим оказался метод NC-Value (лучшие результаты для 3 из 7 морфологических шаблонов), чуть менее эффективными были методы TF (2 из 7) и Weirdness (2 из 7), а TF-IDF, T-Score и Loglikelihood не показали лучших результатов ни для одного из списков.

- Больше всего релевантных ответов было найдено для конструкций с глаголом: Verb + Prep (ap@k ≈ 0.969), Adv + Verb (ap@k ≈ 0.965) и Verb + Noun (ap@k ≈ 0.85). Чуть меньше общенаучных выражений было обнаружено для Prep + N (ap@k ≈ 0.71), Noun + Prep (ap@k ≈ 0.63) и Noun + Verb (ap@k ≈ 0.53). И совсем мало для Adj + Noun (ap@k ≈ 0.37).

- Одно из возможных объяснений – по сравнению с общенаучными конструкциями в качестве именной группы чаще встречаются термины, характерные только для одной предметной области. Тогда как предложная или глагольная группы чаще не являются специальными терминами (Таблица 3).

Таблица 3. Доля общенаучных выражений из отобранных кандидатов для каждого морфологического шаблона

Морфологический шаблон	**Доля общенаучных выражений**
Adj + Noun	≈ 0,43
Noun + Prep	≈ 0,53
Prep + N	≈ 0,66
Adv + Verb	≈ 0,94
Noun + Verb	≈ 0,56
Verb + Noun	≈ 0,82
Verb + Prep	≈ 0,93

Литература

Браславский, Соколов 2006 — П. И. Браславский, Е. А. Соколов. Сравнение четырех методов автоматического извлечения двухсловных терминов из текста // Компьютерная лингвистика и интеллектуальные технологии. Тр. Международной конференции «Диалог» 2006. С. 88–94.

Ahmad et al. 1999 — K. Ahmad, L. Gillam, L. Tostevin. University of surrey participation in trec8: Weirdness indexing for logical document extrapolation and retrieval (wilder) // The Eighth Text REtrieval Conference (TREC-8), 1999.

Frantzi et al. 2000 — K. Frantzi, S. Ananiadou, H. Mima. Automatic recognition of multi-word terms:. the c-value/nc-value method // *International Journal on Digital Libraries*, Vol. 3, No. 2, 2000. P. 115–130.

Rayson, Garside 2000 — P. Rayson, R. Garside. Comparing corpora using frequency profiling // In proceedings of the workshop on Comparing Corpora, held in conjunction with the 38th annual meeting of the Association for Computational Linguistics (ACL 2000), Hong Kong, 2000. P. 1-6.