MULTIPLE FEATURES FOR MULTIWORD EXTRACTION: A LEARNING-TO-RANK APPROACH

Elena Tutubalina [1] Pavel Braslavski [2]

June 2, 2016

[1]Kazan Federal University, Kazan, Russia

[2]Ural Federal University, Ekaterinburg, Russia

Ural Federal University
named after the first President of Russia B.N.Yeltsin

KAZAN (VOLGA REGION) FEDERAL UNIVERSITY KFU 1804

INTRO: MULTIWORD EXPRESSION EXTRACTION

OVERVIEW

- Goal: to extract and rank MWEs according to their fitness for a lexical resource (a unified approach to all MWE types);
- Data: Russian National Corpus and Russian Wikipedia
- Method: Learning to Rank (LETOR) + multiple features + limited training data

EXTRACTION OF MULTIWORD EXPRESSIONS

- Extraction of MWEs:
  - Traditional approaches used statistical association measures (AMs);
  - Machine learning models are trained to classify expressions;
  - Recently, Wikipedia-based approaches and topic models are applied to extract terminology (e.g., using Wikipedia categories and topics).
- *Challenge*:
  - a ranked list of MWEs to be included in the resource with minimal manual intervention;
  - minimal manual labeling.

MULTIWORD EXPRESSIONS

- *Definition* (Baldwin and Kim, 2010): Multiword expressions (MWEs) are lexical items that:
  - can be decomposed into multiple simplex words;
  - display lexical, syntactic, semantic, pragmatic and/or statistical idiomaticity.
- We focus on nominal bigrams (the most common MWE type).
- Unified apprach to different types of MWEs (collocations, idioms, set phrases, etc.).

Examples: новый год [New Year], перекись водорода [peroxide], стиральная машина [washing machine], письменный стол [writing desk], белый дом [the White House], информационные технологии [information technology]

DATA

- Russian National Corpus (RNC) (364M tokens)
- Russian Wikipedia (1.2M articles, 318M tokens)
- candidate MWEs – morpho-syntactic patterns:
  - Adjective + Noun, Noun + Adjective
    - солёная вода [brackish water], дух нечистый [daemon], лишний вопрос [unwanted question], вино красное [red wine]
  - Participle + Noun, Noun + Participle
    - вода дистиллированная [distilled water], вращающийся диск [rotating disk], полынь поникающая [Artemisia nutans]
  - Noun + Noun (genitive), and Noun + Noun (instrumental)
    - период колебаний [period of oscillation], губы сердечком [heart-shaped lips]
- 10+ in the RNC or Wikipedia title

DATA

We collected nominal bigram entries from three dictionaries:

- Wiktionary (3,155)
  - зубной врач [dentist], детская площадка [playground], белый телефон [toilet, literally - white phone], артериовенозная мальформация [arteriovenous malformation]
- Small Academic Dictionary (2,955):
  - дробные числительные [fraction numeral], холодная война [cold war], темная лошадка [dark house], каскад гидроэлектростанций [cascade of hydroelectric station], желтый дом [nut hospital, literally - yellow house]
- Ushakov's Dictionary (2,506):
  - внематочная беременность [ectopic pregnancy], коломенская верста [lanky person], зажиточный колхозник [middle class collective farmer], кисейная барышня [prim young lady], домашняя наставница [home-dame]

DATA SUMMARY

# of unique words in labeled examples	5,871
# of positive examples	3,981
# of negative examples	3,770
# of positive examples in test set	1,322
# of candidate MWEs from the RNC	190,416
# of candidate MWEs from Wikipedia	157,748
# of unique MWEs from both corpora	329,866
# of MWEs overlapping with labeled set (RNC)	82,456
# of MWEs overlapping with labeled set (Wiki)	117,837
# of unique overlapping MWEs	188,441

RANKING

IDEA

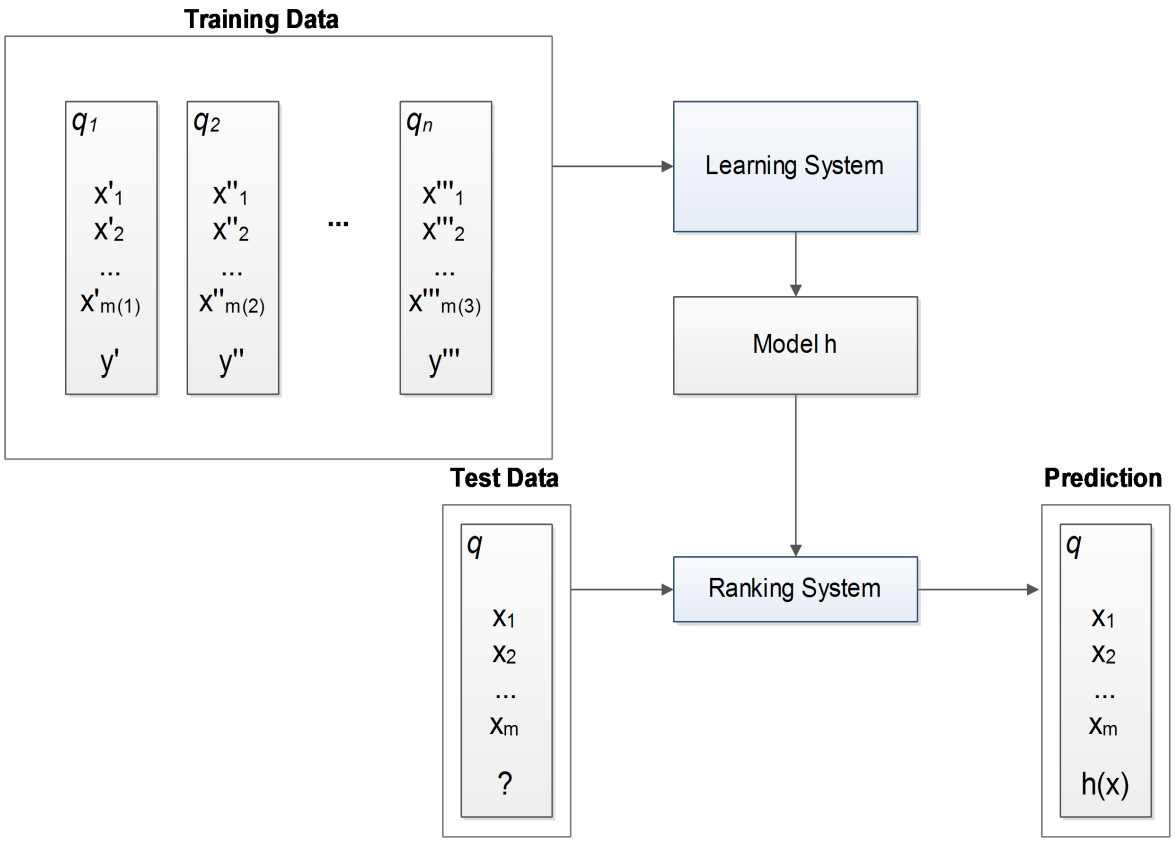
- unified approach to different MWE types;
- ranking instead of classification;
- rich set of features.

IDEA

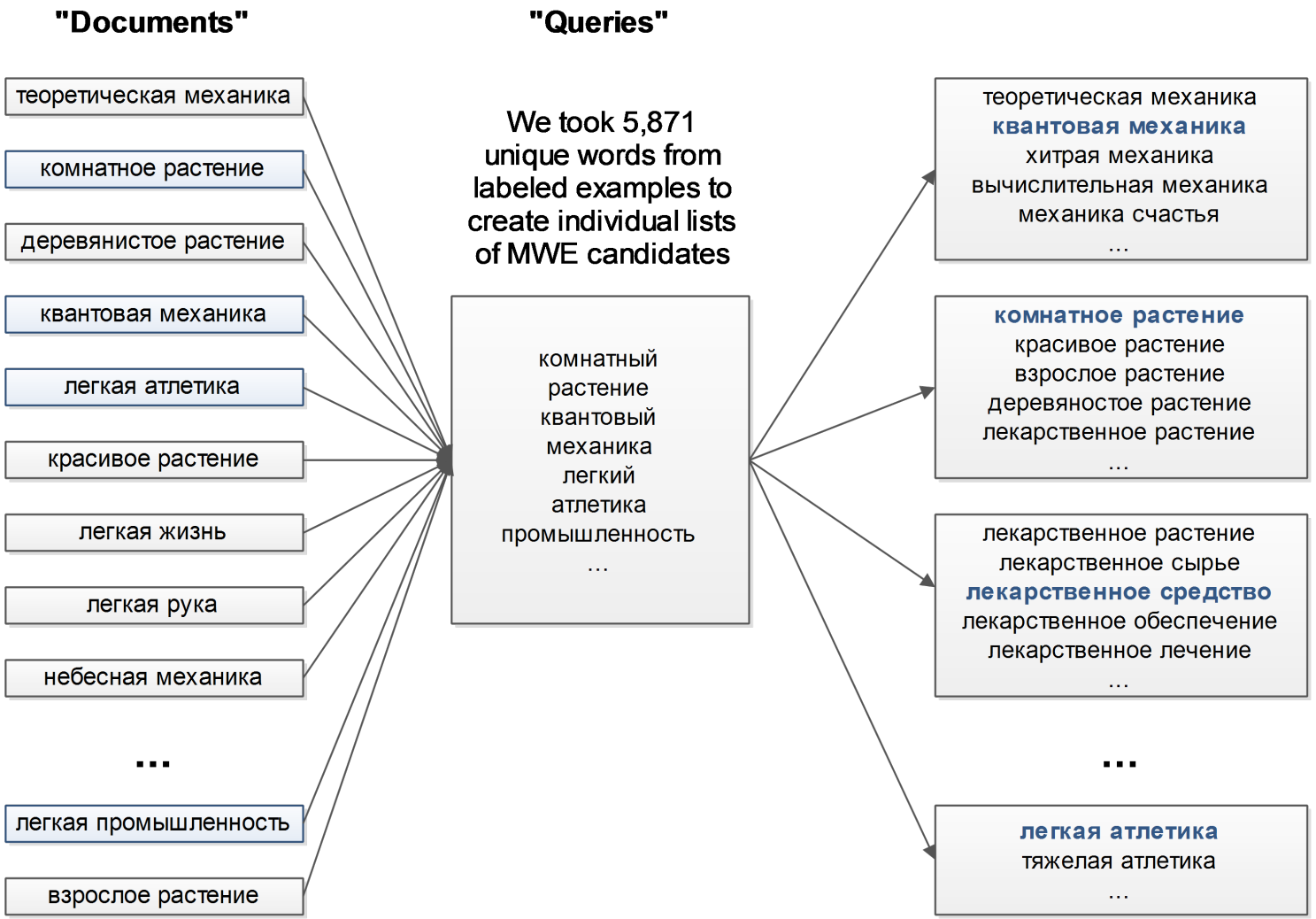
- Ranking has become a central research problem for informational retrieval (IR);
- The task of ranking in IR:
  - given a query, the ranking function measures the relevance of each document;
  - sorts all documents based on their relevance scores;
  - presents a list of top-ranked ones;
- *Learning to rank* for IR is a task to automatically construct a ranking model using training data.

LEARNING TO RANK METHODS

- $q_i (i = 1, \cdots, n)$ – the set of n queries for the training step;
- $x^{(i)} = \{x_j^{(i)}\}^{m(i)} (j = 1, \cdots, m)$ – the feature vectors associated to each query; m – the number of documents in q_i;
- $y^i (i = 1, \cdots, n)$ are the set of relevance judgements.



We represent the data as a set of “queries” and “documents”.



EXAMPLES OF ‘QUERIES’

#	MWEs
1	неправильная установка (wrong installation), неправильная постановка (wrong statement), неправильная музыка (wrong music), неправильная галактика (wrong galaxy), неправильная переменная (wrong variable), **неправильная дробь** (improper fraction)
2	слабая струна (weak string), натянутая струна (tense string), гетеротическая струна (heterotic string), бозонная струна (bosonic string), квантовая струна (quantum string), золотая струна (gold string), космическая струна (cosmic string), **спинная струна** (primitive backbone)
3	белая корова (white cow), старая корова (old cow), черная корова (black cow), синяя корова (blue cow), священная корова (sacred cow), **дойная корова** (milk cow), **морская корова** (sea cow)

FEATURES (1)

We use the following feature set (42 features in total):

- **RNC features** (14):
  - RNC global frequency;
  - ten frequencies in genre subcorpora (e.g., scientific texts, classical literature, legal and official documents, religious texts, children's literature, nonfiction, news, etc.);
  - first and second words' frequencies;
  - the presence of the candidate in the corpus's texts;
- **Structural features** (7):
  - binary features corresponding to extraction patterns;
  - bigram length in characters;

FEATURES (2)

- **Wikipedia-based features** (20):
  - frequency in the Wikipedia corpus;
  - the presence of a redirect with the given MWE, match with a Wikipedia title;
  - the number of in- and out-links;
  - the number of categories assigned to the page, the presence of an infobox;
  - eleven binary features corresponding to the infobox' type[1];
  - capitalization.
- **Web-based feature** (1):
  - the number of documents returned to MWE as a phrase query by a search engine (SE) through an Yandex API[2];

[1] http://en.wikipedia.org/wiki/Wikipedia:List_of_infoboxes
[2] https://xml.yandex.ru/

FEATURE SELECTION

- A feature selection for ranking: find a set of features with maximum importance and minimum similarity (Geng et. al., 2009).
  - It provides a greedy search algorithm to solve the optimization problem.
  - *Mean reciprocal rank* (MRR) is used to measure the importance of features.
  - Kendall's τ is used to measure the similarities between features' ranked lists.
- We held out 20% of the training set as a validation set.

EXPERIMENTS

- We evaluated rankings using two measures:
  - *mean reciprocal rank* (MRR):

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i} \tag{1}$$

  - *bpref* :

$$bpref = \frac{1}{R} \sum_{r} 1 - \frac{|n \text{ ranked higher than } r|}{R} \tag{2}$$

RANKING

- 3 RankLib library algorithms:
  - MART;
  - RankBoost;
  - LambdaMART;
- compared to:
  - Association measures (t-score, log-likelihood, and MI);
  - RNC and Wikipedia frequencies;

- We randomly sampled 80% of the 'queries' for training and held out 20% for testing.
- Both measures were averaged over 1,449 lists in the test set.

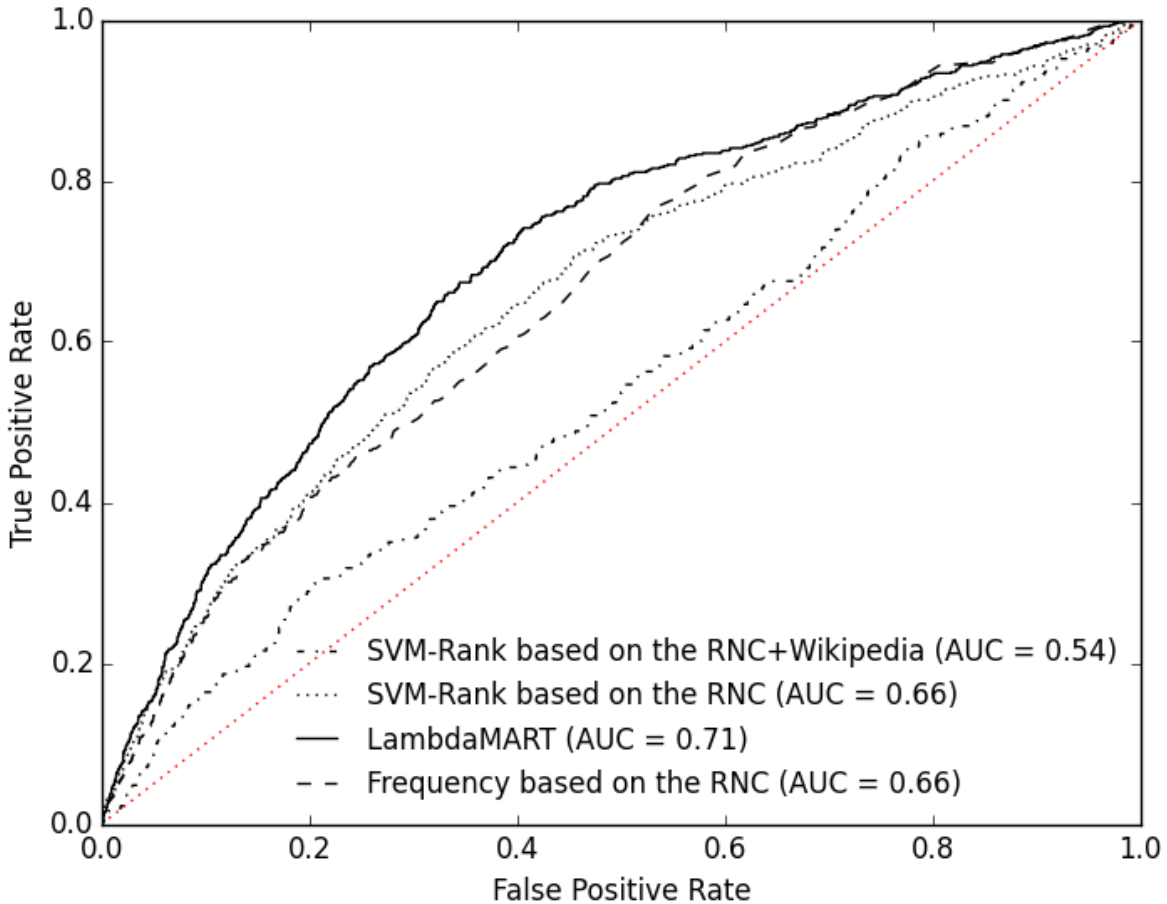
EVALUATION RESULTS

Ranking method	MRR	bpref
MI	0.440	0.353
t-score	0.615	0.321
log-likelihood	0.620	0.353
frequency based on Wikipedia	0.625	0.467
frequency based on the RNC	0.624	0.328
SVM-Rank (RNC)	0.644	0.550
SVM-Rank (Wikipedia)	0.609	0.492
SVM-Rank (RNC+Wikipedia)	0.635	0.483
MART	0.639	0.545
MART + FS	0.639	0.480
LambdaMART	0.679	**0.742**
LambdaMART + FS	0.684	0.546
RankBoost	**0.739**	**0.742**
RankBoost + FS	**0.758**	**0.825**

FEATURE ABLATION EXPERIMENTS

	highest rank	**lowest rank**
all features	0.679	0.598
w/o RNC-based features	0.565	0.497
w/o wiki-based features	0.609	0.543
w/o structural features	0.671	0.592
w/o web frequency	0.678	0.602

ROC CURVES


True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
SVM-Rank based on the RNC+Wikipedia (AUC = 0.54)
SVM-Rank based on the RNC (AUC = 0.66)
LambdaMART (AUC = 0.71)
Frequency based on the RNC (AUC = 0.66)

Examples of ranked MWEs

Cut-off level = 100
земная кора (Earth's crust)
программное обеспечение (software)
основные фонды (basic assets)
биологические науки (bioscience)
общественное мнение (public opinion)
Cut-off level = 2,500
диалектическая логика (dialectical logic)
барионный заряд (baryon charge)
врождённые идеи (innate idea)
гонка вооружений (arms race)
адский огонь (hellfire)

Examples of ranked MWEs

Cut-off level = 10,000
грудная железа (breast gland)
критическая теория (critical theory)
чесменский бой (battle of Chesma)
автоматический огонь (automatic fire)
личное дворянство (personal nobility)
Cut-off level = 30,000
концептуальное искусство (conceptual art)
институциональный инвестор (institutional investor)
земские марки (zhemstvo stamps)
агглютинативные языки (agglutinative language)
ненасыщенный пар (unsaturated steam)

Examples of ranked MWEs

Cut-off level = 100,000
шлиховой анализ (panning)
облеченный тон (invested tone)
дардские народы (dardsky people)
трамвайная археология (tram archeology)
глухой удар (bump)
Cut-off level = 150,000
ноги прохожих (feet of passers-by)
разделенный экран (divided screen)
воркутинская улица (Vorkuta street)
осетинская церковь (Ossetian church)
старый базар (old market)

CONCLUSION

- We have described an experiment on MWE extraction.
- Two data sources in parallel, rich set of features.
- We applied learning-to-rank methods.
- An improvement in ranking/classification task on standard benchmarks.
- Future work:
  - apply the method to verbal MWEs.
  - can we use clustering on the word embeddings for "queries"?