Yandex

The impact of different data sources on finding and ranking synonyms for a large-scale vocabulary

Dialogue 2016

Alexandra Antonova, Taras Kobernik, Alexey Misyurev

The impact of different data sources on finding and ranking synonyms for a large-scale vocabulary

(2015) RUSSE: The First Workshop on Russian Semantic Similarity

Human-oriented translation dictionaries

SMT Phrase Table

Quality Evaluation

(2015) RUSSE: The First Workshop on Russian Semantic Similarity.

To go deeper:

Reproducibility of results

Possible areas of practical application

Contradictory results
(word2vec parameters, …)

Human-oriented translation dictionaries

Яндекс Переводчик ТЕКСТ САЙТ

РУССКИЙ

ум

2 / 10000

АНГЛИЙСКИЙ

mind

ПОХОЖИЕ СЛОВА Сообщить об ошибке

ум сущ

разум · рассудок · интеллект · смысл · здравый смысл · гений · разумение · толк

голова · мозг · память · головной мозг

догадка · мысль · мышление · взгляд · мнение · намерение · соображение

сметка · догадливость · остроумие · смекалка · сообразительность · смышленость

рассудительность · трезвость

мыслительные способности · умственные способности

интеллектуальность · интеллигентность

СЛОВАРЬ Скрыть примеры Сообщить об ошибке

ум сущ *м*

1 intelligence, intellect, brain, reason, mind
(интеллект, разум, мозг)
незаурядный ум – extraordinary intelligence
острый ум – keen intellect
пытливый ум – inquisitive mind

2 wit, cleverness
(остроумие, ловкость)
быстрый ум – quick wit

3 head
(голова)

Human-oriented translation dictionaries

ПОХОЖИЕ СЛОВА ^ Сообщить об ошибке

ум сущ

разум · рассудок · интеллект · смысл · здравый смысл · гений · разумение · толк

голова · мозг · память · головной мозг

догадка · мысль · мышление · взгляд · мнение · намерение · соображение

сметка · догадливость · остроумие · смекалка · сообразительность · смышленость

рассудительность · трезвость

мыслительные способности · умственные способности

интеллектуальность · интеллигентность

Human-oriented translation dictionaries

Objectives:

1)

Cheap, fast and dirty

Human-oriented translation dictionaries

Objectives:

2)

- Objectivity
- Relevance
- Ordered by occurrence
- Feedback

Quality Evaluation

Reproducibility of results

Open data

Possible areas of practical application

Formal

Reference synonym pairs from human-built dictionaries

Features

Translation Similarity

Word2vec

Glove

Frequencies (logarithms)

Glove + Syntax

SMT Phrase Table

Parallel texts, Word alignment

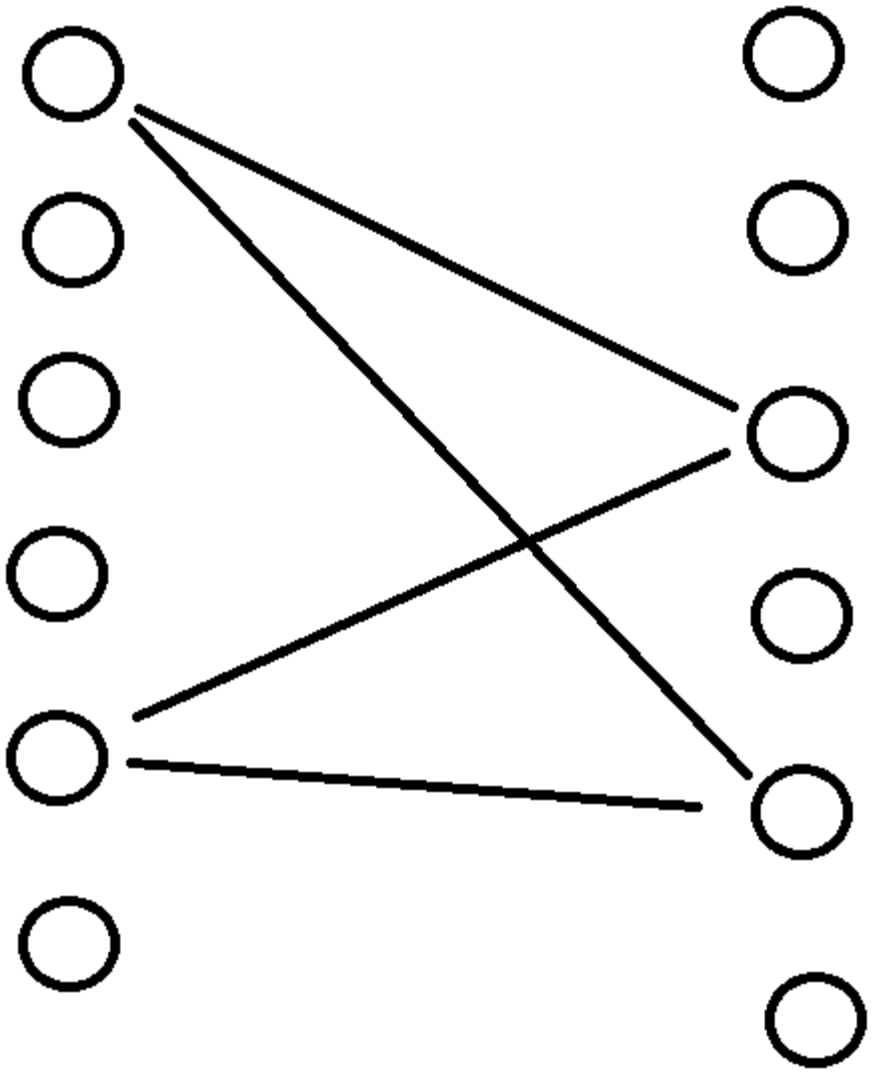
the german - russian dictionary of food industry

немецко - русский словарь по пищевой промышленности

SMT Phrase Table

English	Russian	Src Freq	Dst Freq	Joint Freq
literature	литературный	120847	39363	5674
calculate	рассчитывать	65010	74748	5734
dozens	множество	4799	213344	2632
modernise	модернизация	2199	101669	1230
renovated	обновленный	22106	55299	2875
bureau	контора	86191	16353	3087
everyday	будни	57935	7809	1749
taste	привкус	154219	4859	2251
attachment	крепление	54653	36396	3667
protect	охранять	296098	19074	6178
symptom	признак	113082	145808	10555
task	целевой	488541	176207	24111
precious	ценный	61070	99043	6391
legacy	устаревший	26935	21189	1963
generated	произведенный	35493	42473	3190
adjoining	смежный	9349	17468	1050
leaflet	буклет	16218	17295	1376
quote	смета	60958	19246	2814
renowned	прославленный	28969	5758	1061
staff	посох	508534	6345	4665
controversial	Неоднозначный	19921	8035	1039

SMT Phrase Table



SMT Phrase Table

$$TranslationSimilarity(ru_a, ru_b) = Pr(ru_a|ru_b)Pr(ru_b|ru_a)$$

$$Pr(ru_a|ru_b) = \sum_i Pr(ru_a|en_i)Pr(en_i|ru_b)$$

$$Pr(ru_b|ru_a) = \sum_i Pr(ru_b|en_i)Pr(en_i|ru_a)$$

$$Pr(en_i|ru_a) = \frac{Count(en_i, ru_a)}{\sum_{en_i} Count(en_i, ru_a)}$$

Data

Phrase-table

English-Russian

sum of all joint counts: 2.91 billion

Word2vec, Glove

200 mln sentences

Candidate Synonyms

Training set: 26’281 positive, 2’657’507 negative examples

Test set: 26’461 positive, 2’614’819 negative examples

Only 58.9% of initial reference pairs were found among candidates

Candidate Synonyms

Rank = 0

ганс	ханс
лишь	только
обеспечивать	обеспечить
семьдесят	шестьдесят
затраты	расходы
заросший	поросший
двуокись	диоксид
анакин	энакин
флеш	флэш
дек	окт
негативно	отрицательно
двадцать	тридцать
нарезанный	порезанный
болезнь	заболевание
значительный	существенный
рамсфелд	рамсфельд
содействовать	способствовать
невролог	невропатолог
двигатель	мотор
мухаммад	мухаммед
авг	окт
двести	триста
насладиться	наслаждаться
безусловно	несомненно
дзен	дзэн

Candidate Synonyms

Rank = 100’000

закусочная	кафе
жертвоприношение	подношение
вечнозеленый	хвойный
вдохновлять	окрылять
чудак	чудик
выходить	ехать
горестный	душераздирающий
живописный	уютный
актер	фильм
посредством	при
всего	примерно
костюм	смокинг
нефтехимия	химия
вытянуть	спустить
застрелить	расстреливать
докапитализация	рекапитализация
кладовка	чулан
сателлит	спутник
индонезия	сингапур
рафсанджани	рафсанжани
петь	сочинять
истолочь	толочь
потоптать	топтать

Candidate Synonyms

Rank = 545’220

дебаты	диалог
знатный	могучий
разработчик	создатель
злясь	разгневавшись
конструкционный	профилированный
оглушенный	подавленный
вызвать	навлечь
путь	часть
существо	сущий
выпечки	духовка
сгрести	смести
распутница	шлюха
дискета	накопитель
переправить	попасть
влияние	свойство
видеосъемка	фотосессия
осмотрев	присмотревшись
вполне	где
опава	опавский
приспособление	сооружение
обесцениваться	снижаться

Reference Synonyms

Wiktionary.org (Russian)

99394 single-word synonym pairs

42509 distinct queries

Abramov's dictionary of Russian synonyms and similar words (1915).

34930 pairs

12527 distinct queries

The intersection: 6616 pairs

Missing Synonym Pairs

%	Human judgement	Example
36	Good candidates	ряженка – варенец (~milk product) пацанва – детвора (~children)
25	Both words are rare or unknown	тонемика – тонология (~tonology?)
21	One word is rare or unknown	гуртоправ – гуртовщик (~drover) скворец – кокако (~starling)
6	One word is slang or obscene	записка – малявка (~note)
5	Words are not synonyms	важный (~important: adjective) – царственно (~kingly: adverb)
4	One word is not Russian	галоген – галоїд (~halogen)
3	Obsolete meaning	сплетник (~gossip) – трубач (~trumpeter) управлять (~to control) – рядить (~to dress?)

Top synonym candidates for “проворный”(“agile”) by descending translation similarity

Rank 0-11	Rank 12-22	Rank 23-34
верткий (~nimble)	вертлявый (~fidgety)	динамичный (~dynamic)
ловкий (~agile) **+W +A**	подвижный (~mobile)	оживленный (~brisk)
поворотливый (~agile)	оперативный (~operational)	своевременный (~timely)
шустрый (~nimble) **+W**	**стремительный** (~rapid) **+W**	находчивый (~resourceful)
юркий (~nimble) **+W**	скорейший (~early)	изворотливый (~quirky)
прыткий (~nimble)	подсказанный (~prompted)	безотлагательный (~urgency)
маневренный (~maneuvering)	**бойкий** (~spirited) **+W +A**	пробужденный (~awakened)
быстрый (~fast) **+W +A**	незамедлительный (~immediate)	активный (~agile)
быстро (~quickly)	сноровистый (~nimble)	скорый (~fast)
расторопный (~agile) **+W**	подвижной (~mobile)	сообразительный (~witted)
гибкий (~flexible)	подсказывающий (~prompting)	**резвый** (~spirited) **+W**
		бодрый (~brisk)

+W – intersects with Wiktionary
+A – intersects with Abramov dictionary

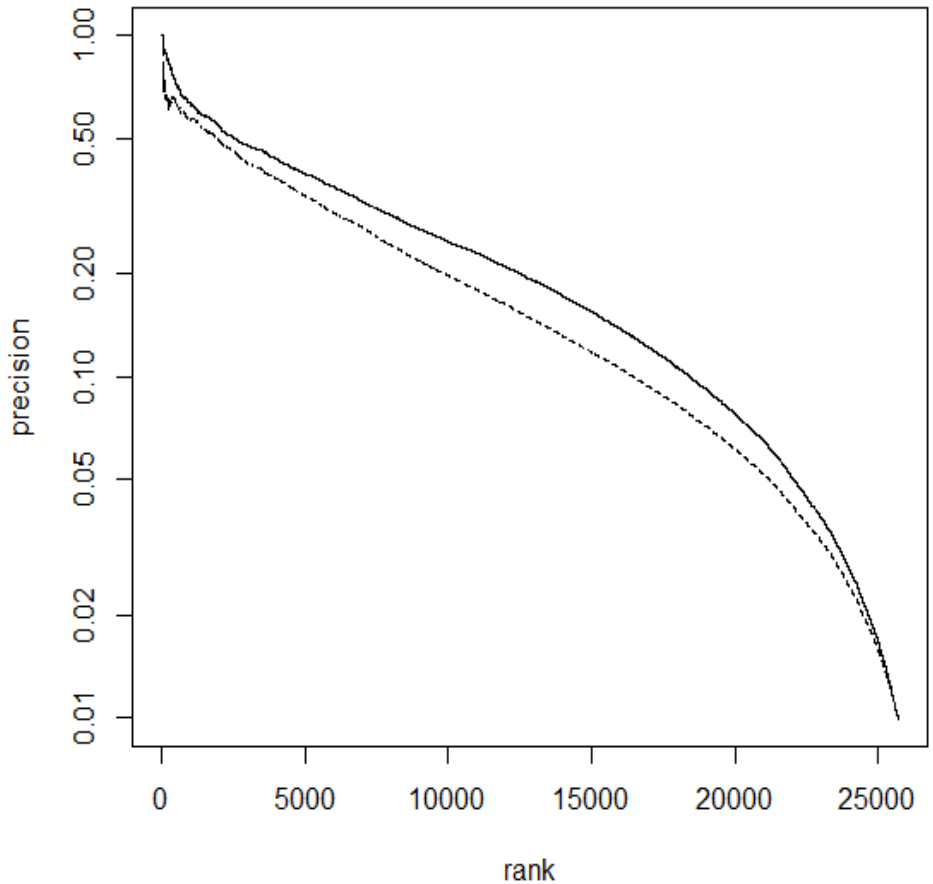
Metrics

average precision

average rank (AveRank)

median of ranks (Median)

Precision vs. Rank

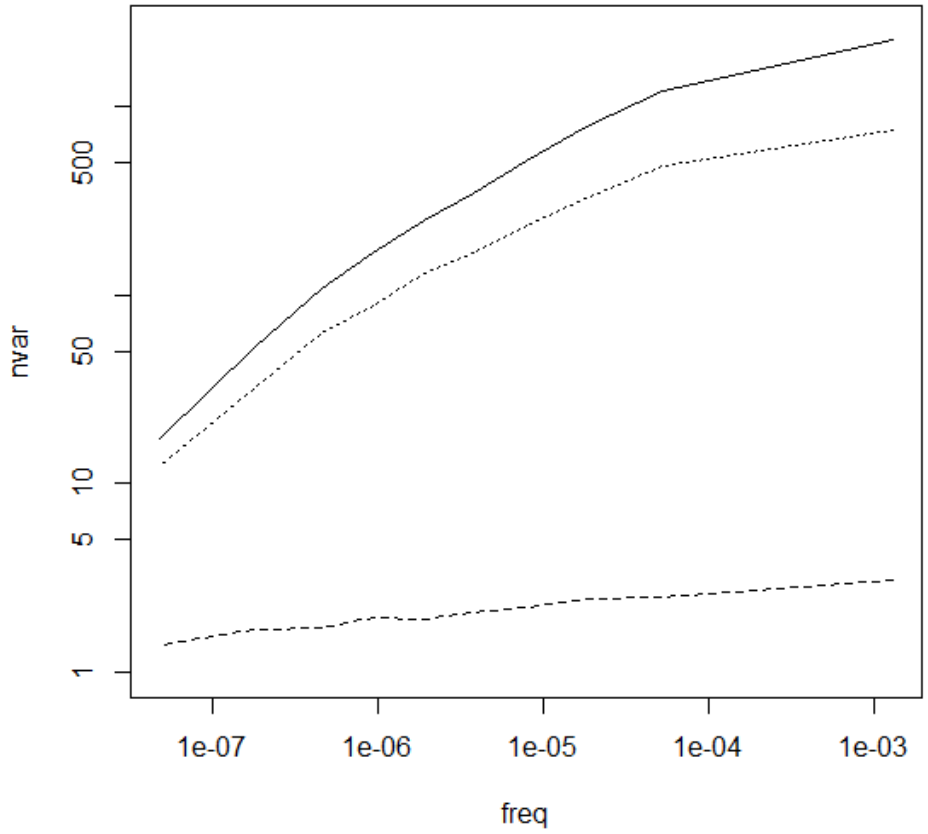


Precision vs rank.
Solid line – translation similarity, dashed line – word2vec

Ranking Results

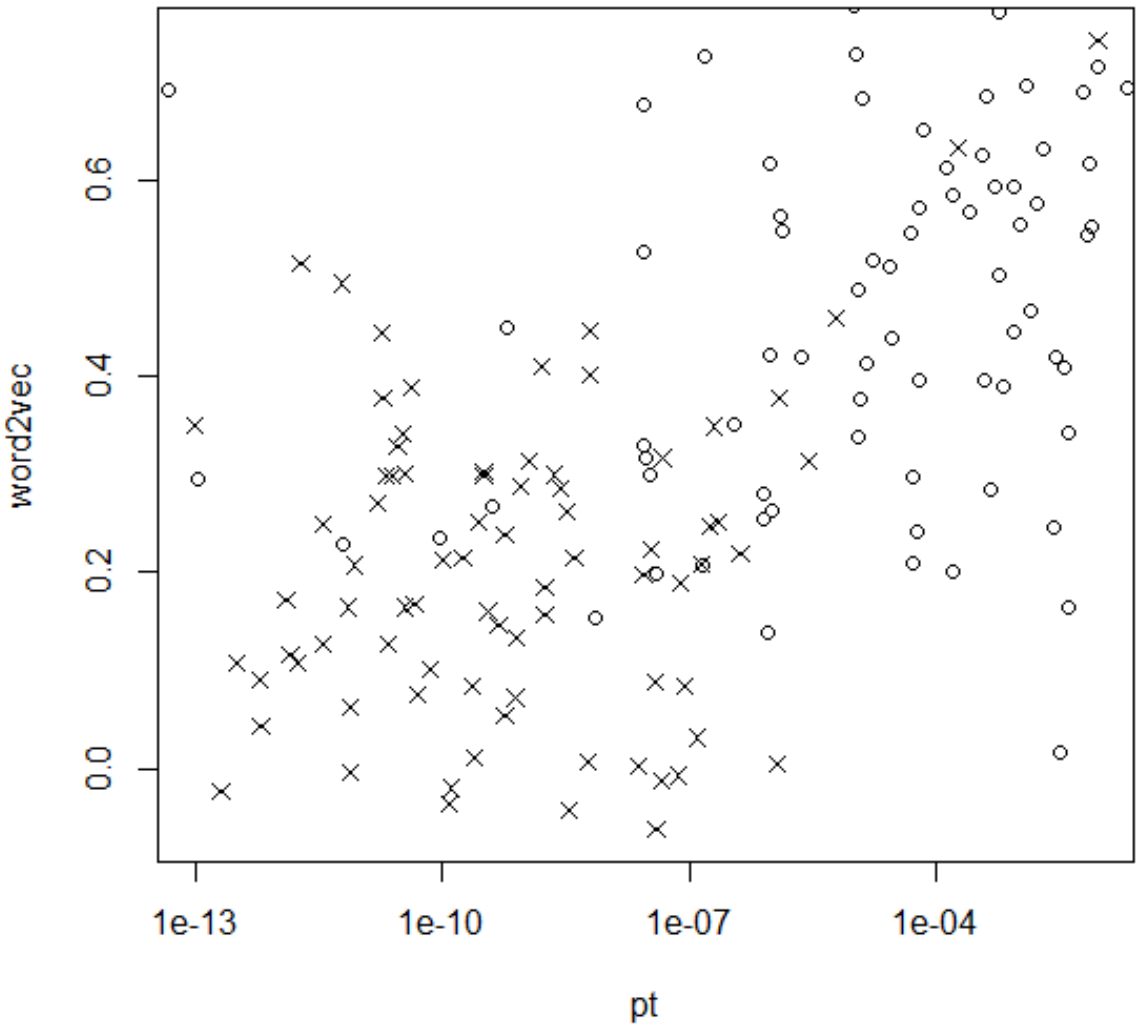
Feature combination	AveP	AveRank	Median
Wiktionary dataset			
Word2Vec	0.165	407228	132683
TranslationSimilarity	0.237	222513	66802
TranslationSimilarity + Frequencies	0.247	212819	65304
Word2Vec + TranslationSimilarity + Frequencies	**0.303**	**181381**	**50232**
Glove	0.117	560219	219061
Glove + TranslationSimilarity + Frequencies	0.299	182327	50338
GloveSynt	0.058	803457	467868
GloveSynt + TranslationSimilarity	0.274	204993	57393
GloveSynt + TranslationSimilarity + Frequencies	0.291	191225	54492
Abramov dataset			
Word2Vec	0.025	516313	244381
Word2Vec + TranslationSimilarity + Frequencies	0.068	250096	79268
Glove	0.031	506889	220102
Glove + TranslationSimilarity + Frequencies	**0.0751**	**238683**	**73224**
TranslationSimilarity	0.0491	272115	99969

Number of Candidate Synonyms



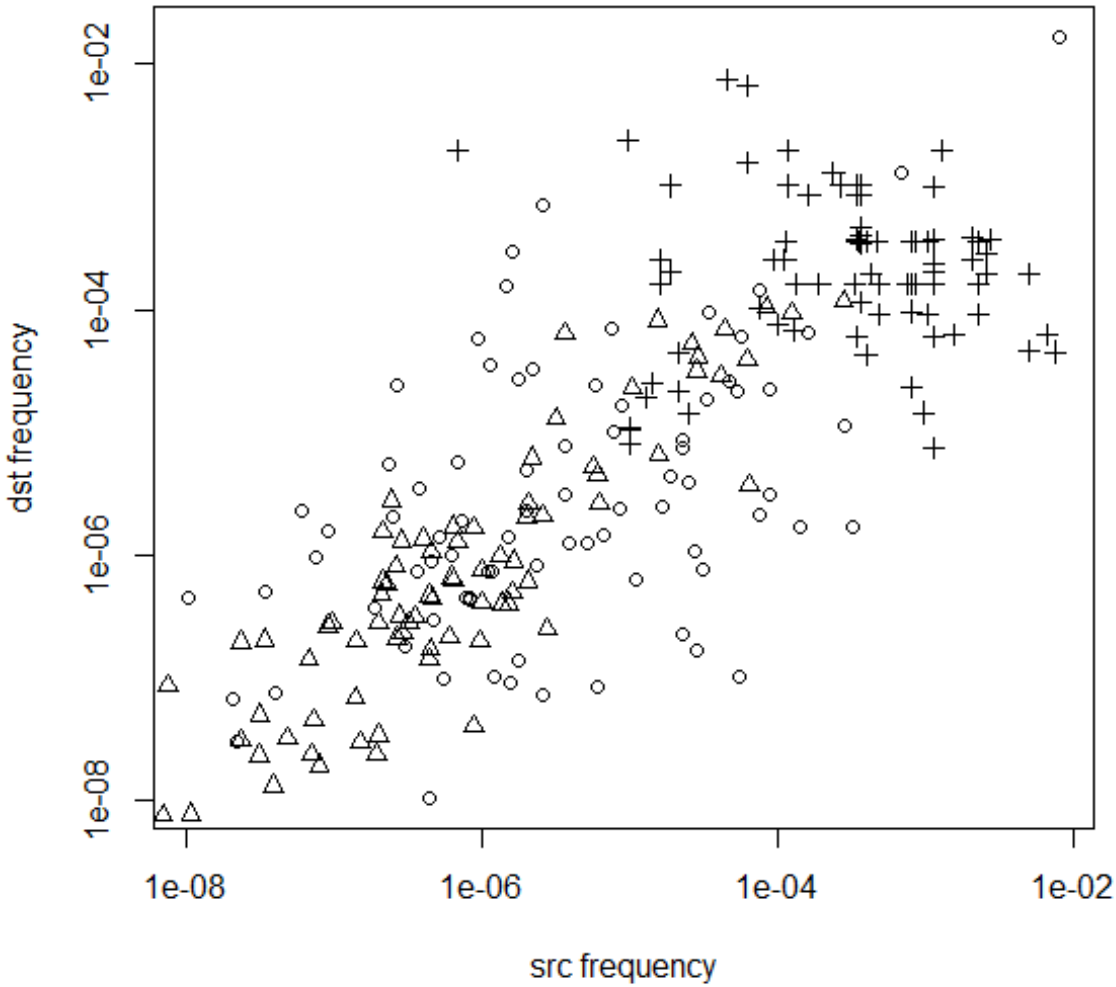
Dependence between the number of candidate synonyms and the query frequency. Bottom line – reference synonyms, upper line – all candidates by translation model, the middle line – number of candidates per one reference pair

Translation Similarity / word2vec distance



positive (o) and negative (x) examples

Distribution of candidate pairs w.r.t. word frequencies



o - random reference pairs

Δ - translation similarity

+ - top by vector models

Conclusion

Candidate pairs from phrase-table

Long tail of the distribution

Word2vec – for the most frequent words

Phrase-table – rare and polysemous words

Reproducible metric

Unusual results: Glove + phrase-table, Abramov dictionary

More examples

http://translate.yandex.ru/

http://translate.yandex.com/ - English version

Yandex

The impact of different data sources on finding and ranking synonyms for a large-scale vocabulary

(2015) RUSSE: The First Workshop on Russian Semantic Similarity

Human-oriented translation dictionaries

SMT Phrase Table

Quality Evaluation