WORD SENSE FREQUENCY OF SIMILAR POLYSEMOUS WORDS IN DIFFERENT LANGUAGES

Boris Iomdin, Konstantin Lopukhin,
Anastasiya Lopukhina, Grigoriy Nosyrev

Most frequent sense

Many approaches:

- Macquarie Thesaurus (Mohammad and Hirst 2006)
- WordNet (McCarthy et al. 2007, Bhingardive et al. 2015)
- RuThes-lite (Loukachevitch and Chetviorkin 2015)

Overall sense distribution

Rarely studied:

- Topic modeling based on Hierarchical Dirichlet Processes and on word sense induction (Lau et al. 2014)
- Pattern Dictionary of English Verbs (Hanks and Pustejovsky 2005, Hanks 2008)

Sense frequency and language learning

Possible uses:

- Finding the most relevant senses in dictionaries
- Finding primary meanings for basic vocabulary lists
- Preparing lexicon tests

Bilingual pairs of similar words

- Cognates
- Borrowings
- International words

Not always the same meaning structure

(In fact, hardly ever!)

Sample data

66 Russian-English noun pairs, including:

- Authentic cognates: *брат — brother*, *гусь — goose*, ...
- English borrowings: *бар — bar*, *бизнес — business*, ...
- French borrowings: *анекдот – anecdote*, *батарея – battery*, …
- Latin and Greek: *адвокат – advocate*, *альбом – album, гармония – harmony*, ...

Procedure: English words

Sense frequencies of the English words obtained from SemCor 3.0:

- Total size: more than 200,000 tokens are sense-annotated
- Linked to WordNet 3.0 senses
- Selected words have at least 20 labeled contexts

Procedure: Russian words

Sense frequencies of the Russian estimated automatically by performing WSD on 1000 random contexts sampled from the Russian National Corpus.

Sense inventory and training data from the **Active Dictionary of Russian** (examples, collocations, synonyms).

Challenging task due to low number of training examples.

WSD Pipeline

- Wide **contexts** (10 words on each side), do not take order into account.

WSD Pipeline

- Wide **contexts** (10 words on each side), do not take order into account.
- Assign **weight** for each word ≈ PMI: giving more weight to words that are more likely to appear with the context word than without.

WSD Pipeline

- Wide **contexts** (10 words on each side), do not take order into account.
- Assign **weight** for each word ≈ PMI: giving more weight to words that are more likely to appear with the context word than without.
- Get word2vec vectors for each word, average them using these weights ⇒ **context vector**.

WSD Pipeline

- Wide **contexts** (10 words on each side), do not take order into account.
- Assign **weight** for each word ≈ PMI: giving more weight to words that are more likely to appear with the context word than without.
- Get word2vec vectors for each word, average them using these weights ⇒ **context vector**.
- Average context vectors for all contexts of one sense from the dictionary ⇒ **sense vector**.

WSD Pipeline

- Wide **contexts** (10 words on each side), do not take order into account.
- Assign **weight** for each word ≈ PMI: giving more weight to words that are more likely to appear with the context word than without.
- Get word2vec vectors for each word, average them using these weights ⇒ **context vector**.
- Average context vectors for all contexts of one sense from the dictionary ⇒ **sense vector**.
- When performing disambiguation of an unknown context, calculate the context vector and choose the sense with the **closest sense vector**.

sensefreq.ruslang.ru

ipm	Word	Senses	Est. precision	1st - 2nd	1	2	3	4
3.3	абрикос	2	0.84	0.72	0.86	0.14		
9.1	абсурд	2	0.81	0.43	0.72	0.28		
9.8	авангард	3	0.73	0.19	0.52	0.33	0.15	
7.1	аванс	3	0.61	0.38	0.64	0.26	0.10	
56.0	автомат	3	0.90	0.48	0.73	0.25	0.02	
8.9	автономия	4	0.44	0.07	0.42	0.36	0.14	0.07

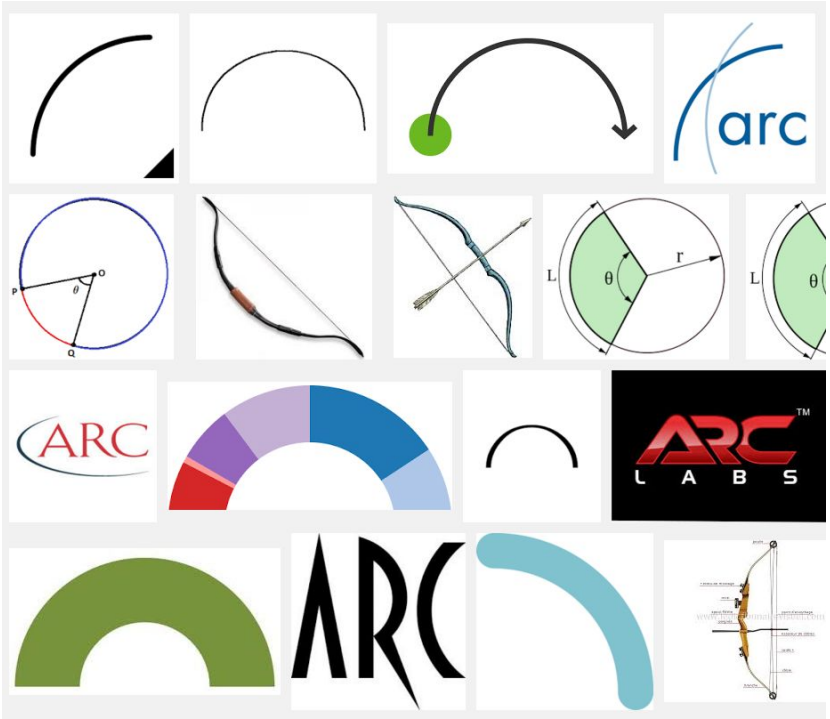
Some results

We observed 3 cases:

- Dominant senses match, but others do not
- Some senses match, but different dominant senses
- Senses do not match at all

No match

арка — arc



No match

вагон — wagon

русский ↔ английский

вагон

Английско-русский словарь *На данный момент в разработке*

вагон *существительное, муж. род*

car *сущ.*

wagon *сущ.*

менее частотные:

carriage *сущ.* · carload *сущ.* · waggon *сущ.* · van *сущ.*

⚠ В результате пожара никто из пассажиров поезда не пострадал, однако один вагон поезда полностью выгорел изнутри, в другом вагоне повреждены внутренняя обшивка, сидения, крыша.

telegraf.by

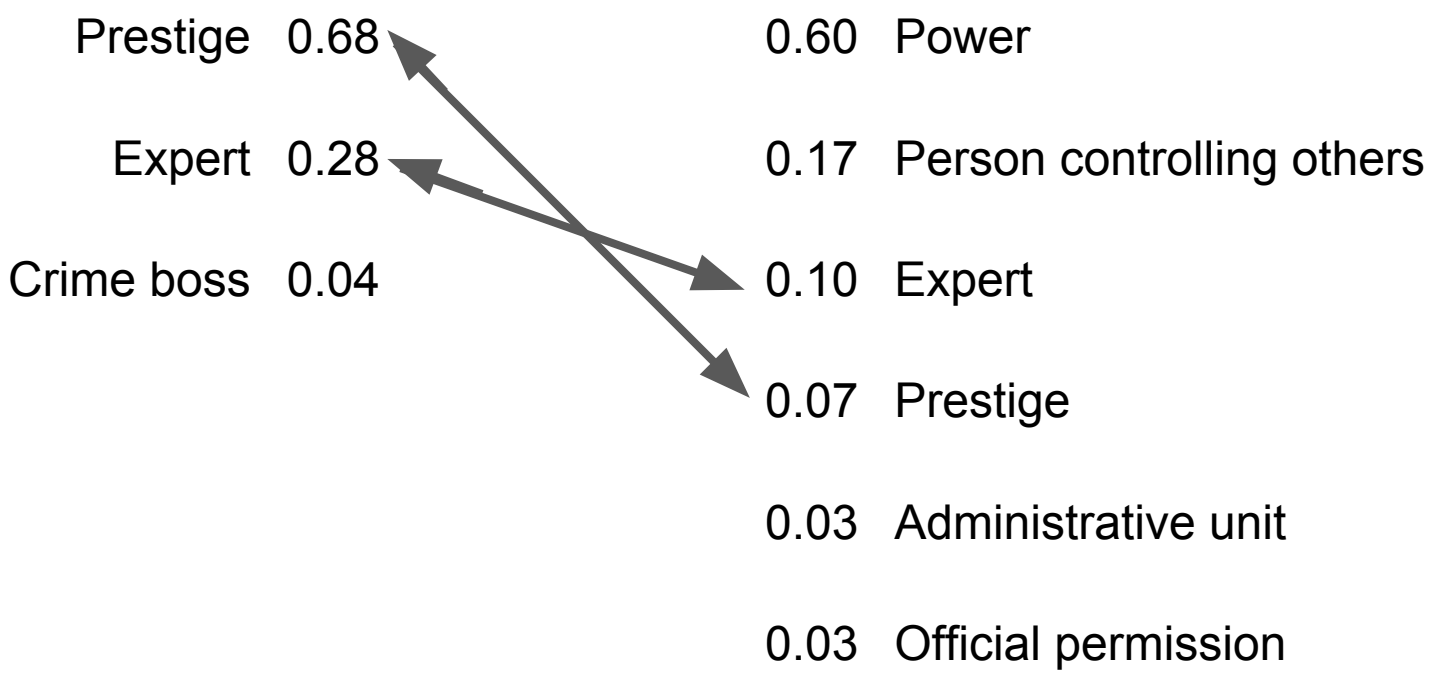
⚠ None of the train passengers were injured in the fire, but a wagon train completely burnt out from the inside. There were also damaged interior panels, seats and roof in another wagon.

telegraf.by

Different dominant senses

авторитет ‘prestige’	*authority* ‘power’
акция ‘share’	*action* ‘something done’
артист ‘performer’	*artist* ‘painter’
банда ‘a group of criminals’	*band* ‘a group of musicians’

авторитет — authority



Non-native usage (translations)

*The head of the Shelkov criminal gang, a **criminal authority**, has been arrested in Thailand.*

< *криминальный авторитет* ‘a mafia boss’

*The **touristic base** built in Erzhei for the 'Oktai' ensemble is now open not only to the young singers but to everybody*

< *туристическая база* ‘a camping site’

File:Abandoned touristic base in Biysk 03.JPG

Google

Переводчик

иврит | английский | русский | Определить язык

английский | иврит | французский | Перевести

Криминальный авторитет
Система блоков
В вагоне

Criminal authority
blocks system
In the wagon

Immigrant usage

*Когда сегодня человеку проще сесть в машину и проехать два **блока** в магазин, то ему надо помнить о своем здоровье*
(Chajka, a Russian magazine published in the USA)

*Так нет же, надо было ему именно такое место для обеда выбрать, что бы перекрыть чуть ли ни самый главный выезд из даунтауна. Двадцать минут в очереди, что бы выехать из гаража и еще сорок, что бы проехать два **блока***
(Livejournal.com)

Finding mistakes: an online experiment

- Respondents had to find mistakes in 15 English sentences taken from real texts
- Beside 10 filler sentences, there were 5 sentences containing one of the words under study: *arc, authority, base, block, wagon*
- Each of these words was used in 2 sentences: one in a correct sense absent in Russian, the other in an incorrect sense characteristic for the Russian cognate.

Families were walking beside wagons pulled by teams of oxen
(standard usage of *wagon*, but absent in Russian *вагон*)

When they were travelling, a wagon accidentally disconnected from their train
(standard usage of *вагон*, but absent in English *wagon*)

Results

English speakers reported significantly less mistakes in sentences where the words under discussion were used in their dictionary meanings.

Russian speakers reported less mistakes in sentences where these words were used with meanings absent in English dictionaries but natural to the Russian cognates of these words.

Errors reported, %	Native usage	Non-native usage
English speakers	1%	21%
Russian speakers	8%	10%

Conclusions

- A pilot study of sense frequency distributions was performed: we compared meaning structures of English and Russian nouns with respect to sense frequencies and found out that they are dissimilar
- This discrepancy causes various mistakes
- We can obtain data useful for learners of Russian or English (or, potentially, other languages) and for lexicographers and computational linguists dealing with machine translation or deep semantic analysis

Future plans

- Apply the method of estimating word sense frequencies used for Russian by using the data from the MacMillan dictionary and the DANTE database
- Study parallel corpora to investigate sense structures of cognates used by native and non-native speakers, in different translations
- Create and update a database of Russian-English cognates comparing their senses according to their frequency
- **Expand the method to closely related languages to provide material for comparative semantic and lexicological studies**
- **Expand the method to non-cognate translation equivalents (*вещь* vs. *thing*, *встреча* vs. *meeting*)**