Creating a list of word alignments from parallel Russian simplification data

Anna Dmitrieva

Supervisors: Jörg Tiedemann, Ulla Vanhatalo, Roman Yangarber

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Recap: my thesis work

Topic: Automatic text simplification of Russian texts using means of machine translation

Previous stages:

- A pilot version of a multi-task simplification model is developed and trained on English data.
- A Russian-Simple Russian parallel dataset for text simplification is made, baseline models have been trained.

Current stage: experiments on Russian simplification data

RuAdapt

- RuAdapt is a parallel dataset in which target texts are simplified versions of the source texts, i.e. a parallel Russian-Simple Russian dataset;
- Consists of texts adapted for learners of Russian as a Foreign Language and their original versions;
- Subcorpora:
  - Adapted literature (largest);
  - Encyclopedic entries;
  - Fairytales.
- Has paragraph and sentence level alignments that were created automatically.

This study

- Goal: develop a list of monolingual word alignments taken from parallel Russian simplification data.
- What can be done with this data?
  - Create lexical simplification systems;
  - Create a neural monolingual word alignment model suitable for simplification data;
  - Study the vocabulary of adapted texts.

source	target
поглядывает	смотрит
припомнилось	вспомнилось
брякнет	скажет
коль	если
Экой	Какой
обабился	баба
трогательное	волнительное
сюда	здесь
взоры	взгляды
кличешь	зовёшь

Lexical simplification

- Changing “complicated” words in the text for “simpler” synonyms;
- Often requires a special word list where each source word has a simpler target word;
- Can be rule-based or ML-based: for example, use lexically constrained decoding.

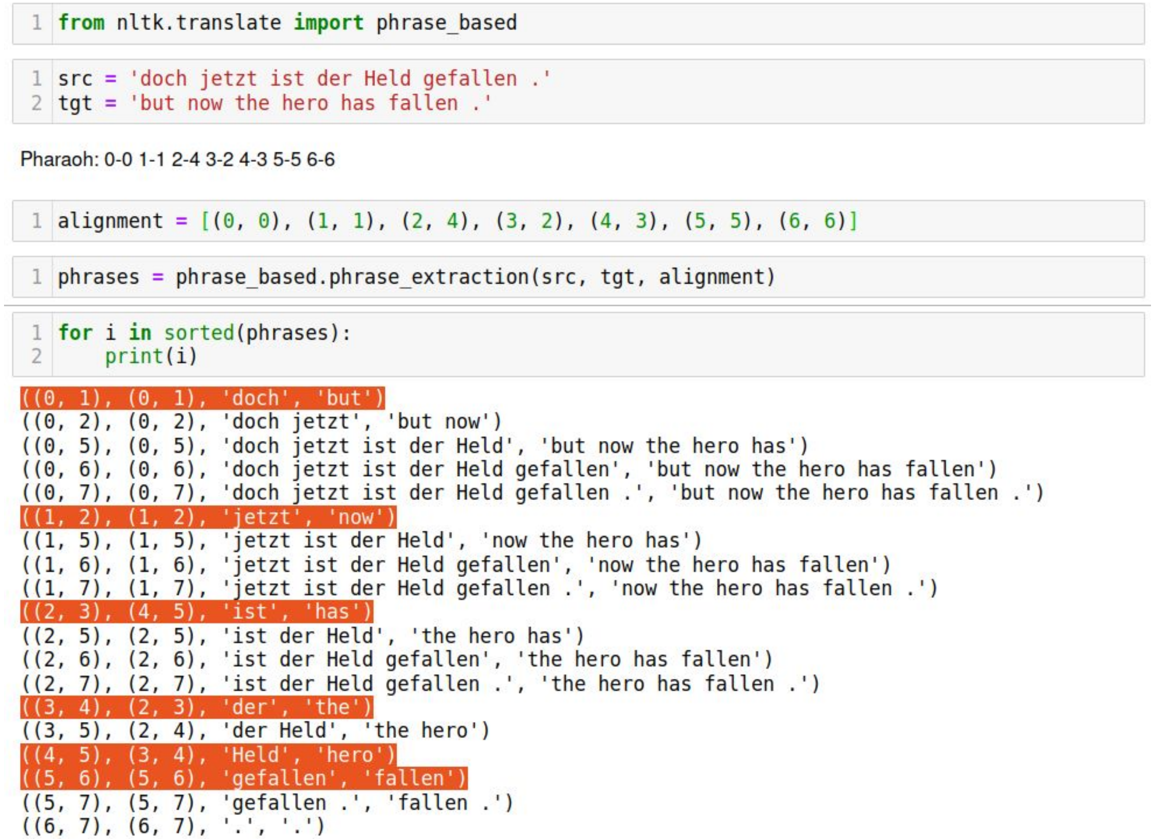
Creating the list(s)

- For now, only adapted literature was used, since it it the larger subcorpus;
- 15156 sentence pairs to extract word pairs from
  - Cosine similarity between sentences: 0.99 > cos_sim > 0.31
- Word aligners: 1 unsupervised statistical aligner, 1 neural aligner;
- Only one-word alignments are included;
- Alignments will be post-edited by experts.

Word alignment: eflomal

Eflomal (Efficient Low-Memory Aligner) is a system for efficient and accurate word alignment using a Bayesian model with Markov Chain Monte Carlo (MCMC) inference.

- An additional dataset of around 2.5 mil. paraphrases (from Opusparcus and paraphraser.ru) was used to train the aligner on;
- Outputs Pharaoh-format alignments. A dedicated instrument from NLTK called phrase_based was used to obtain word-to-word alignments.
- It is possible to also match phrases.



```
from nltk.translate import phrase_based
```

```
src = 'doch jetzt ist der Held gefallen .'
tgt = 'but now the hero has fallen .'
```

Pharaoh: 0-0 1-1 2-4 3-2 4-3 5-5 6-6

```
alignment = [(0, 0), (1, 1), (2, 4), (3, 2), (4, 3), (5, 5), (6, 6)]
```

```
phrases = phrase_based.phrase_extraction(src, tgt, alignment)
```

```
for i in sorted(phrases):
    print(i)
```

```
((0, 1), (0, 1), 'doch', 'but')
((0, 2), (0, 2), 'doch jetzt', 'but now')
((0, 5), (0, 5), 'doch jetzt ist der Held', 'but now the hero has')
((0, 6), (0, 6), 'doch jetzt ist der Held gefallen', 'but now the hero has fallen')
((0, 7), (0, 7), 'doch jetzt ist der Held gefallen .', 'but now the hero has fallen .')
((1, 2), (1, 2), 'jetzt', 'now')
((1, 5), (1, 5), 'jetzt ist der Held', 'now the hero has')
((1, 6), (1, 6), 'jetzt ist der Held gefallen', 'now the hero has fallen')
((1, 7), (1, 7), 'jetzt ist der Held gefallen .', 'now the hero has fallen .')
((2, 3), (4, 5), 'ist', 'has')
((2, 5), (2, 5), 'ist der Held', 'the hero has')
((2, 6), (2, 6), 'ist der Held gefallen', 'the hero has fallen')
((2, 7), (2, 7), 'ist der Held gefallen .', 'the hero has fallen .')
((3, 4), (2, 3), 'der', 'the')
((3, 5), (2, 4), 'der Held', 'the hero')
((4, 5), (3, 4), 'Held', 'hero')
((5, 6), (5, 6), 'gefallen', 'fallen')
((5, 7), (5, 7), 'gefallen .', 'fallen .')
((6, 7), (6, 7), '.', '.')
```

Obtaining word-to-word alignments from Pharaoh format

Example from https://github.com/clab/fast_align, code from https://www.nltk.org/api/nltk.translate.phrase_based.html

Word alignment: AWESOME

Awesome-align (Aligning Word Embedding Spaces Of Multilingual Encoders) is a tool that can extract word alignments from multilingual BERT and allows users to fine-tune mBERT on parallel corpora for better alignment quality.

- No large training corpus needed;
- Can match single words or phrases.

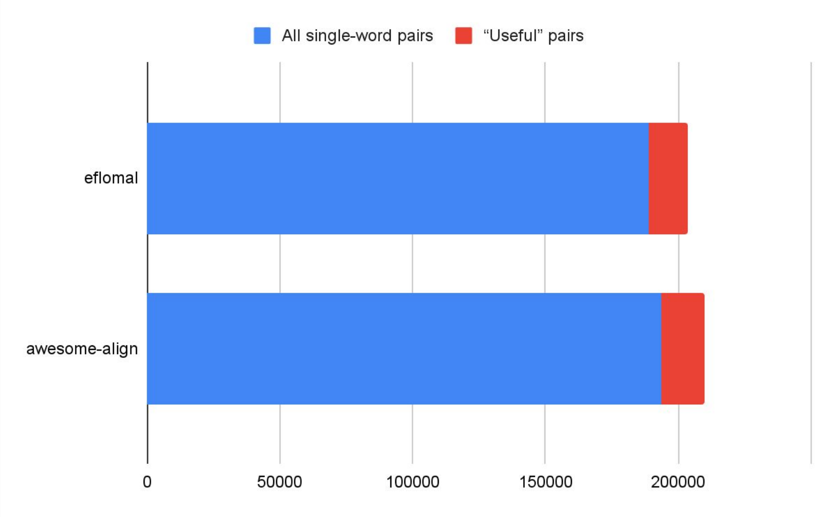
Word alignment: stats

Statistic	eflomal	awesome-align
All single word pairs	188706	193778
Pairs consisting of different words, cleaned from noise	19687	22767
Unique pairs	14807	15989
Unique pairs in common	8403	

Word alignment: stats

A “useful” pair of words is a pair that can be included in the list of word alignments. In this pair of words:

- Source word and target word are different;
- There is no noise (punctuation instead of words, etc.)



Instructions for editors: eflomal alignments

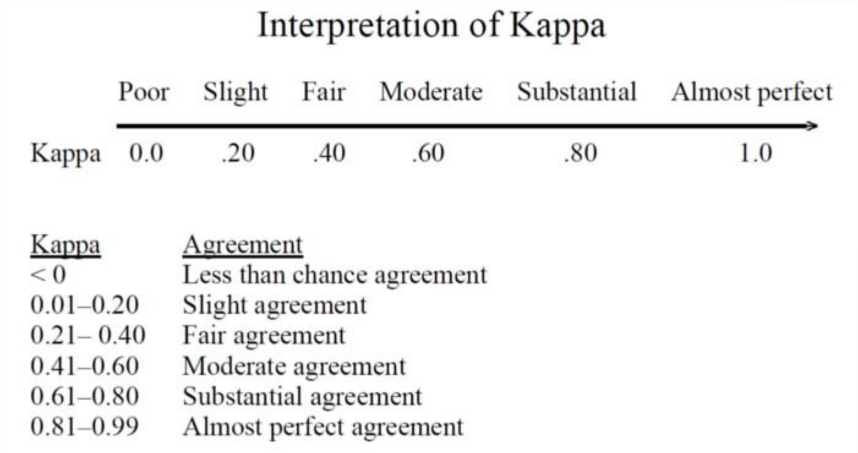
- 10 files, each one ~1480 pairs in length, every file has a copy => 20 documents altogether;
- An editor picks two documents to edit, but cannot assign themselves to a document and its copy. No document is edited by the same person twice;
- The editor gives a score to each word pair.

Instructions for editors

- Score 0 is given to:
  - noisy pairs (non-synonyms);
  - pairs consisting of the same word or different forms of the same word.
- Score 1 is given to:
  - pairs that can only be considered synonyms in a certain context;
  - different words with the same root (звать/позвать);
  - synonyms that are presented by different parts of speech.
- Score 2 is given to:
  - pairs that are considered synonyms in most contexts;
  - pairs where the source word is an older form of the target word (кофий/кофе).

Current post-editing situation

- Files checked (total): 12/20
- Word pairs checked (without copies): 13326/14807
- Pairs that received score 2 at least from one editor: 1517
- Pairs that received score 1 at least from one editor: 2884
- Cases where both the file and its copy has been checked: 3/10
- Mean inter-annotator agreement score (Kohen’s cappa): 0.35



Interpretation of Kappa

	Poor	Slight	Fair	Moderate	Substantial	Almost perfect
Kappa	0.0	.20	.40	.60	.80	1.0

Kappa	Agreement
< 0	Less than chance agreement
0.01–0.20	Slight agreement
0.21– 0.40	Fair agreement
0.41–0.60	Moderate agreement
0.61–0.80	Substantial agreement
0.81–0.99	Almost perfect agreement

Editing the rest of the alignments

- 7586 pairs from awesome-align left to post-edit;
- First step: eliminating words with same roots
  - Exclude all pairs where two words have at least one same root;
  - Use NeuralMorphemeSegmentation (https://github.com/AlexeySorokin/NeuralMorphemeSegmentation) to split words into morphemes;
  - 796 pairs detected, after manual check 775 pairs were excluded.
- Next steps: possibly training a classifier on existing alignments.

Studying adapted texts’ vocabulary

Strategies:

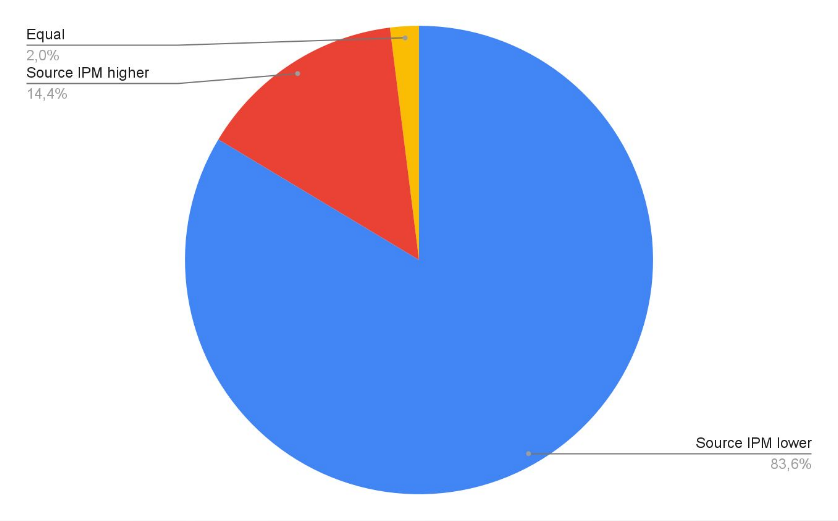
- Compare word pairs to CEFR grade lists: lists of words that a learner should know on a certain level of language acquisition
  - In a given pair, **source** word is supposed to have a **higher grade level** than target word (for example, an A2 level word should be replaced with an A1 synonym);
- Compare word pairs to a frequency dictionary
  - **IPM** (instances per million words) of the **source** word should be **lower** than IPM of the target word
  - Dictionary used: The frequency dictionary of modern Russian language (O. Lyashevskaya, S. Sharov - Azbukovnik, Moscow, 2009. http://dict.ruslang.ru/freq.php)

Current statistics

All 1517 word pairs that received grade 2 from the editor(s) were checked.

Frequencies:

- 1360 pairs where both words have IPM values in dictionary;
- In 1137 cases the source word's IPM is lower than target's.
- Sometimes IPM is equal, mostly in cases where source word has non-modern spelling (несчастие/несчастье), because in such cases source and target are lemmatized the same way.

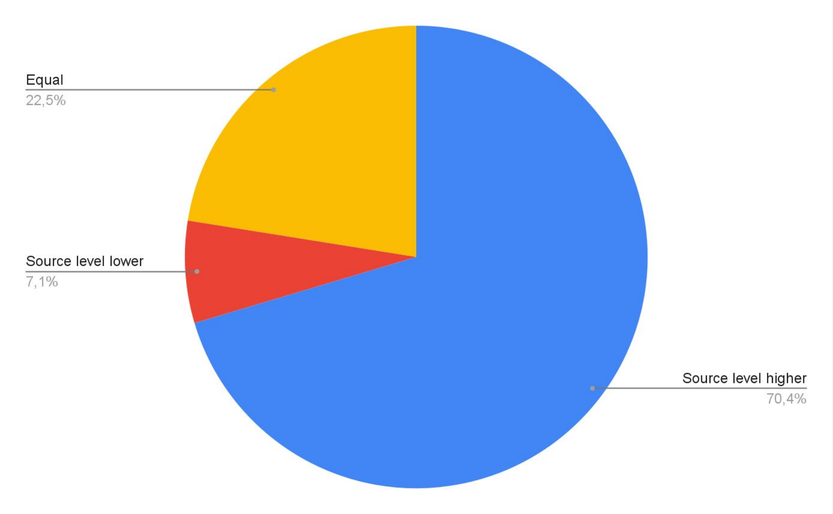


Current statistics

All 1517 word pairs that received grade 2 from the editor(s) were checked.

Words’ grade levels:

- 645 pairs where both words are present in the CEFR vocabulary lists;
- In 454 cases the source word’s level is higher than target’s;
- In 145 cases the levels are equal.



Problems

- In simplification data, source can be a lot longer than target.
  - Solution: for now, use human post-editing. Later, more efficient monolingual word aligners can be trained.
- Should words be lemmatized before the list is given to editors?
  - Solution: for now, do not lemmatize the words before human editing.
- How to handle different words with the same root (звать/позвать)? How to handle synonyms that are presented by different parts of speech?
  - Solution: give them a 1 score: not exactly a pair of synonyms but also not noise.
- Entries in the vocabulary lists (CEFR vocabulary lists, the frequency dictionary), as well as in our list, are ambiguous in meaning.
  - Solution: none?

Conclusions

- Not many cases of actual lexical simplification (i.e. synonym replacement) are happening in adapted literature. OR: such phenomena cannot be fully captured without special word aligners for parallel simplification data.
- When lexical simplification does happen, in most cases it is “good”, reasonable replacements where the replacement (target) word is simpler and/or more frequently used than the replaced (source) word.