Semantic Recommendation System for Bilingual Corpus of Academic Papers

Anna Safaryan Petr Filchenkov Weijia Yan Andrey Kutuzov Irina Nikishina

October 16, 2020

Do We Need Bilinguality? The Answer is "Yes".

Current RusNLP

- Search engine for academic papers
- Dialogue, AIST and AINL
- English papers

→

To Do

- Bilingual recommendations
- Cross-lingual word embeddings
- Off-the-shelf vs. self-made?

Table 1: RusNLP corpus statistics

Conference	**Since**	**Texts**	**Russian**	**English**
Dialogue	2000	1,785	1,424	361
AIST	2012	91	21	70
AINL	2015	96	0	96
Total texts		**1,983**	**1,445**	**527**

https://nlp.rusvectores.org/

How does it look like?

Target Paper:

MULTI-PRONUNCIATION LEXICON FOR RUSSIAN AUTOMATIC SPEECH RECOGNITION (PILOT STUDY)

Shirokova A. (anna_a@stel.ru),
Telesnin B. (telesnin_ba@stel.ru),
Rogozhina V. (mind_your_own_business@rambler.ru)

Stel CS, MSLU, Moscow, Russia

Our pilot study is aimed at building a lexicon of effective pronunciation variants on the basis of canonical pronunciations, for implementing it into the automatic speech recognition system for Russian. We focus on phonetic changes in word pronunciation caused by different factors operating in spontaneous speech. Our speech data includes three different corpora of the conversational type. Manual expert processing and analysis of the audio data are used. The lexicon construction procedure is given. Some statistics for pronunciation variation in Russian, obtained from the speech data, is presented. A description of frequent types of this phenomenon is given. Parallel and sequential pronunciation variants are discussed. Ways of formulating general phonetic variation rules and predicting potential contexts, in which pronunciation variation is likely to appear, are considered. Test data, phoneset used, and automatic speech recognition (ASR) parameters are described. Preliminary results for ASR and key word spotting (KWS) are shown. The appropriateness of using multi-pronunciation lexicon is discussed.

Keywords: Russian spontaneous speech, pronunciation variants, Russian pronunciation, spontaneous speech, pronunciation lexicon, reduction, Russian ASR

How does it look like?

Транскрибирование, структурирование и временной анализ речевого корпуса эстонского языка при выборе единиц в системе синтеза (текст-речь)

Transcrbing, structuring and temporal analysis of fluent speech corpus for a unit selection tts system for Estonian

Meelis Mihkla (meelis@eki.ee), **Indrek Kiissel** (indrek@eki.ee),
Tõnis Nurk (tonis@eki.ee), **Liisi Piits** (liisi@eki.ee)
Institute of the Estonian Language, Tallinn, Estonia

В статье рассматриваются проблемы создания системы синтеза, основанной на выборе единиц из корпуса эстонского языка (текст-речь). Авторы предлагают правила транскрибирования и принципы фонологического структурирования, облегчающие выбор языковых единиц. Исследуется также интенсивность коллокации (сочетаемости) в зависимости от темпа речи и разрабатываются соответствующие модели длительности.

Evaluation the quality of Estonian text-to-speech synthesis and diphone corrector for the TTS system*

Meelis Mihkla, Einar Meister, Indrek Kiissel, Jürgen Lasn

Abstract

The main tasks of the Estonian text-to-speech synthesis project have in principle now been fulfilled: an Estonian diphone database has been created and the linguistic processing of the text and prosody modelling has been realised. The planning of further developments required an interim evaluation of the present state of the synthesis as far as the intelligibility, smoothness and naturalness of the synthesised speech was concerned. Speech intelligibility depends to a great extent on the selection of speech units and their segmental quality. We use the Esprit/SAM test. Part of the test material was generated as VCV, VC and CV words, using 17 Estonian consonants in the environment of the extreme vowels *a*, *i* and *u*. The other set of stimuli was made up of the most frequent VCV, VC and CV combinations occurring in the Estonian language. To improve the smoothness of synthetic speech it seems reasonable if we combined some words in the sentence into prosodic compounds. These unusual compounds will inevitably produce some unknown diphones. The same problem occurs in the pronounciation of foreign words and names. Therefore we need a diphone corrector. We also discuss about future developments of Estonian TTS synthesizer.

Keywords: Estonian TTS, quality evaluation, SAM test, diphone corrector

Lemmatization for Ancient Languages: Rules or Neural Networks?

Authors   Authors and affiliations

Oksana Dereza

Conference paper
First Online: 27 September 2018

Part of the Communications in Computer and Information Science book series (CCIS, volume 930)

Abstract

Lemmatisation, which is one of the most important stages of text preprocessing, consists in grouping the inflected forms of a word together so they can be analysed as a single item. This task is often considered solved for most modern languages irregardless of their morphological type, but the situation is dramatically different for ancient languages. Rich inflectional system and high level of orthographic variation common to these languages together with lack of resources make lemmatising historical data a challenging task. It becomes more and more important as manuscripts are being extensively digitized now, but still remains poorly covered in literature. In this work, I compare a rule-based and a neural network based approach to lemmatisation in case of Early Irish (Old and Middle Irish are often described together as "Early Irish") data.

Keywords

Early Irish   Natural language processing   Under-resourced languages   Lemmatisation   Neural networks   Sequence-to-sequence learning

Speech analysis and synthesis systems for the tatar language

Publisher: IEEE   Cite This   PDF

Aidar Khusainov ; Alfira Khusainova   All Authors

Abstract

Document Sections

I. Introduction
II. Continious Speech Recognition System for the Tatar Language
III. Automatic Speech Synthesis System for the Tatar Language
IV. Language Identification System
V. Conclusion

Abstract:
In this paper we describe our recent work of creation speech human-machine interface for the Tatar language. Our work consists of three main elements: speech recognition system, speech synthesizer and language identification system. These systems will be used in mobile and desktop applications, for instance, machine translation system, smart assistant.

Published in: 2016 IEEE Artificial Intelligence and Natural Language Conference (AINL)

Date of Conference: 10-12 Nov. 2016   **INSPEC Accession Number:** 16792402
Date Added to IEEE *Xplore*: 06 April 2017   **Publisher:** IEEE
ISBN Information: **Conference Location:** St. Petersburg, Russia

I. Introduction
Using speech as a tool for manipulating electronic devices is becoming more and more common. This fact can be proved by lots of desktop and web-based services that provide

ЛАРИНГАЛИЗАЦИЯ В ОЦЕНОЧНЫХ РЕПЛИКАХ РУССКОГО ДИАЛОГА

А.М. Андреева

«СТЭЛ – Компьютерные Системы»

anna_a@stel.ru

О.Ф. Кривнова

МГУ им. М. В. Ломоносова

Ларингализация функционирует как на сегментном, так и на супрасегментном уровнях. Для последнего относительно хорошо исследовано ее использование в синтаксической функции, частично в экспрессивно-модальной. В нашей работе уточняются и расширяются данные об употреблении глоттальной смычки и ее заменителей в экспрессивно-оценочных репликах русского диалога. Предполагается большое количество иллюстраций (осциллограмм и интонограмм), полученных на собственном экспериментальном материале.

So, what should we do?

Preprocessing

Generation of a text representation

Search for the closest papers

Pretrained UDPipe models:

Russian	English
(SynTagRus)	(ParTUT)

Bilingual dictionary from Facebook

Pretrained monolingual embeddings (Wikipedia) or Pre-aligned cross-lingual embeddings (Wikipedia)

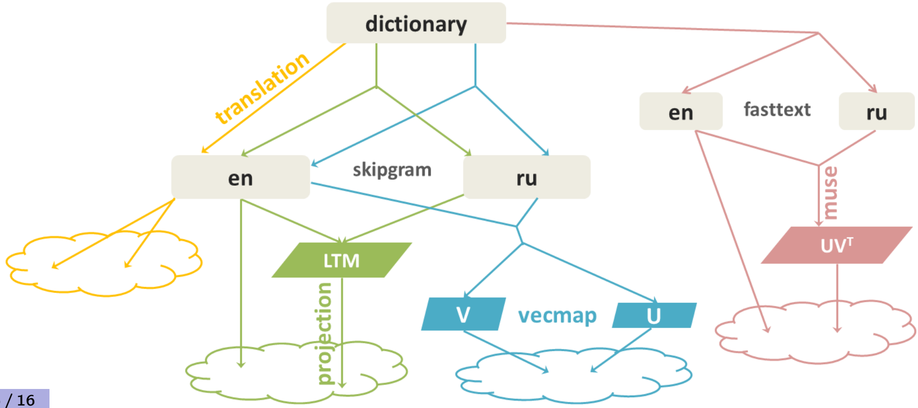
What are the ways to get cross-lingual representations?

Self-Made

- Translation
- Linear Projection
- VecMap (Bilingual Word Embedding Mappings)

Off-the-Shelf

- MUSE (Multilingual Unsupervised and Supervised Embeddings)



What are the ways to get cross-lingual representations?

Self-Made

- Translation
- Linear Transformation
  - [Mikolov et al., 2013a]
- VecMap
  - [Artetxe et al., 2018]

- Skip-gram
  - [Mikolov et al., 2013b]
- Lemmatised + POS tags

Off-the-Shelf

- MUSE
  - [Lample et al., 2018]

- Fasttext
  - [Bojanowski et al., 2017]
- Not preprocessed

How to evaluate recommendations?

Design

- 20 papers in Russian + 20 papers in English (randomly)
- 4 methods $\rightarrow$ 5 closest papers for each target one
- How many recommended papers are relevant to the target one?

Annotators

- Expertise in the field + knowledge of both languages
- Crowdsourcing
- 3 annotators per recommendation

Which recommendations were more relevant?

Table 2: RusNLP experimental results for target papers in both languages: precision

Method	Precision
Translation	54.5
Projection	54.5
VecMap	54.2
MUSE	**58.5**

Are the results consistent?

Table 3: RusNLP experimental results for target papers in both languages: inter-rater agreement

Method	**Krippendorff's α**
Translation	**0.347**
Projection	0.262
VecMap	0.163
MUSE	0.170

Are the results consistent?

Table 3: RusNLP experimental results for target papers in both languages: inter-rater agreement

Method	Krippendorff's α
Translation	**0.347**
Projection	0.262
VecMap	0.163
MUSE	0.170

Any problems?

- Ambiguity of the guidelines
- Not paper-specific evaluation
- Size of the annotation forms

Are the results really cross-lingual?

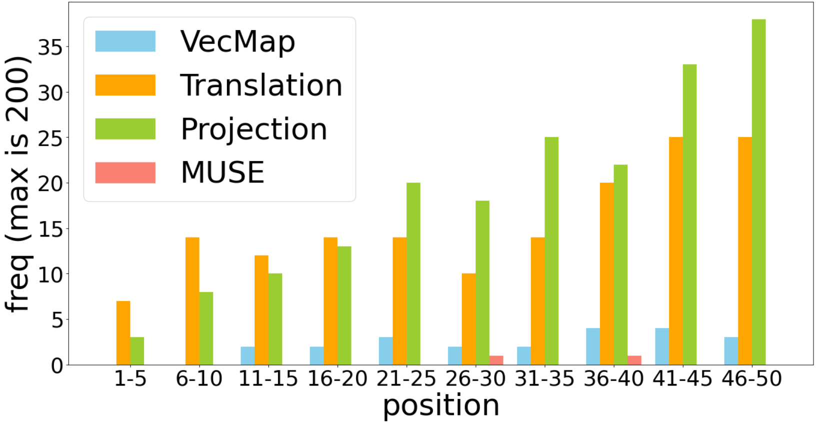


Figure 1: Distribution of cross-lingual recommendations by position

Position — a place of a paper in the list of recommendations sorted by cosine similarity.
Freq — an absolute number of recommended papers written not in the language of the target paper (out of 200 recommendations: 40 target papers $\times$ 5 positions in a bin).

What about the coverage?

Method	English texts			Russian texts		
	Tokens	Vocab	Dict size	Tokens	Vocab	Dict size
Translation	71.53	63.15	296,630	53.91	47.99	19,118
Projection	71.53	63.15	296,630	**89.30**	**85.57**	248,978
VecMap	71.53	63.15	296,630	**89.30**	**85.57**	248,978
MUSE	**89.30**	**83.21**	200,000	86.58	82.84	200,000

Table 4: Coverage (%)

Token coverage — the percentage of tokens from the text length.
Vocabulary coverage — the percentage of unique words from the text vocabulary taken into account when vectorising by each method.

Did Muse Outperform Other Methods?

Outcomes

- MUSE has the best precision (58.5%)
- Most of recommended papers were in the same language
- Low inter-rater agreement for all methods

In the Future

- Changes in the evaluation setup (binary/ranking)
- Dependence on coverage
- Text-level vectorisation
- Specialised embeddings

Source code: `https://github.com/rusnlp/hse_nis`

References I

Artetxe, M., Labaka, G., and Agirre, E. (2018).
Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations.
In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 5012–5019.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017).
Enriching word vectors with subword information.
Transactions of the Association for Computational Linguistics, 5:135–146.

Lample, G., Conneau, A., Ranzato, M., Denoyer, L., and Jégou, H. (2018).
Word translation without parallel data.
In *International Conference on Learning Representations*.

Mikolov, T., Le, Q. V., and Sutskever, I. (2013a).
Exploiting similarities among languages for machine translation.

References II

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b).
Distributed representations of words and phrases and their compositionality.
In *Advances in neural information processing systems*, pages 3111–3119.

Experiment on the Wikipedia

- 54 pairs of articles from the Russian and English Wikipedia with parallel titles.
- For each article it was automatically evaluated whether the article with a parallel title was included into the top-1, top-5, and top-10 recommendations.

Table 5: Wikipedia experimental results for target papers in both languages

Method	Recall@1	Recall@5	Recall@10
Translation	51.85	87.96	95.37
Projection	**56.48**	**91.67**	97.22
VecMap	38.89	85.19	99.07
MUSE	34.26	90.74	**100.00**