

Открытые (?) направления в Kyrgyz NLP

Антон М. Алексеев

2023

Докладчик

Опыт в **индустрии**: высоконагруженные сервисы (Яндекс), чат-боты техподдержки (SoftIT Labs), машинное обучение в рекламе (NativeRoll)

С 2018 года: **наука** в ПОМИ РАН ([лаб. искусственного интеллекта](#))

- Deep Learning, NLP, RecSys, Multi-Modal Data Analysis, Networking, etc.
- проекты с крупными российскими и зарубежными компаниями
- публикации на AAAI, ACL, WSDM, ECIR, LREC, etc.

Преподаватель: ИТМО, НИУ ВШЭ, СПбГУ и Computer Science Center ([youtube](#))

Озвученное сегодня — это мои частные представления, разумеется, можно спорить

План

1. Обработка (кыргызского) языка
2. Задачи, рассматриваемые в рамках диссертации (?)
3. Другие открытые (?) задачи

Где мы сталкиваемся с NLProc?

to the unforeseen circumstances,
Kyrgyz NLP is not ready yet.

are the preliminary results of all
am very much willing to double check

Contextual Spelling Error

Deer Mr. Theodore: **Spelling Error**

I am exceedingly interested in this po
and employment background are app

While working toward my degree, I
small firm. I increased my call volu
success. I will completes my degree
employment in early June.

Grammar Error

АНГЛИЙСКИЙ ↔ ЛАТИНСКИЙ

Ignorance is bliss.

19 / 5 000

Ignorantia sit beatitudo.



g & Duck

Обзор Фото и видео 32 **Отзывы** 122 Особенности

4,4 ★★★★★
302 оценки

Оцените это место

122 отзыва

Напитки • 84%	Атмосфера • 96%	Персонал • 6
54 отзыва	52 отзыва	42 отзыва

Обработка естественного языка (NLProc, NLP)

[Журафски-Мартин, SLP 2 изд.] “Цель этой новой области — заставить компьютеры решать полезные задачи, связанные с человеческим языком, в том числе обеспечение человеко-машинной коммуникации, улучшение коммуникации людей друг с другом или попросту любая полезная обработка текста или речи”

[Wikipedia (en) 2019-08-25] “Обработка естественного языка (NLP) — это подобласть лингвистики, компьютерных наук, обработки информации и искусственного интеллекта, в которой занимаются взаимодействием между компьютером и человеческими (естественными) языками, в частности, тем, как программировать ЭВМ с целью обработки и анализа больших объёмов данных на естественных языках”

Как дела обстоят с кыргызским языком?

- Значительное количество научных работ по теме: КГТУ, КТУ Манас, ОшГУ и др.
- Источники данных: новости, Википедия, открытые документы (?), доступные для скачивания тексты книг (?), несколько корпусов (e.g. Leipzig)
- Готовые инструменты: [alexeyev/awesome-kyrgyz-nlp](https://github.com/alexeyev/awesome-kyrgyz-nlp) (буду рад дополнениям)

Компьютердик лингв

Макалa **Талкуу**

Компьютердик лингвистика (computational linguistics) – тилди компьютерде моделдеп ишт максатын көздөгөн колдонмо лингвистиканын б тармагы. Тилди компьютерде моделдөө жагына информатика, жасалма интеллект, программала сы өндүү кибернетикалык багыттар менен Жалпысынан компьютер аркылуу ишке ган тил илиминдеги бардык багыттар рдик лингвистика

2-май, шейшемби

16:45 Өзбекстан Конституциясынын жаңы редакциясы күчүнө кирди

16:35 COVID-19. Бир аптада 12 учур катталды

Министрлер кабинетинин мажлисинде казино ачууга биринчи расмий лицензия берилди

Как дела обстоят с кыргызским языком?

По классификации Ж. Мирзахвалова (Унив. Южной Флориды), у кыргызского статус “**Scraping-By**”:
размеченных данных на кыргызском языке доступно меньше, чем для узбекского, казахского и турецкого, и требуется коллективное усилие для изменения ситуации

Table 2: Number of L1 speakers for each Turkic language, its corresponding language code and the associated category from the resource taxonomy¹

Language Name	Codes	Speakers (L1)	Category
Turkish	tr, tur	85.0M	The Underdogs (4)
Uzbek	uz, uzb	27.0M	The Rising Star (3)
Azerbaijani	az, aze	23.0M	The Scraping-Bys (1)
Kazakh	kk, kaz	13.2M	The Rising Star (3)
Uyghur	ug, uig	10.0M	The Scraping-Bys (1)
Turkmen	tk, tuk	6.70M	The Scraping-Bys (1)
Tatar	tt, tat	5.20M	The Scraping-Bys (1)
Kyrgyz	ky, kir	4.30M	The Scraping-Bys (1)
Bashkir	ba, bak	1.40M	The Scraping-Bys (1)
Chuvash	cv, chv	1.04M	The Scraping-Bys (1)

Jamshidbek Mirzakhlov. Turkic interlingua: A case study of machine translation in low-resource languages. Master's thesis, University of South Florida, 2021.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. The state and fate of linguistic diversity and inclusion in the nlp world. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 6282–6293, 2020.

Языки, упоминающиеся в заголовках конференции TurkLang 2013-2022

	2013 (Астана)	2014 (Стамбул)	2015 (Казань)	2016 (Бишкек)	2017 (Казань)	2018 (Ташкент)	2019 (Симферополь)	2020 (Уфа)	2021 (Кызыл)	2022 (Нур-Султан)
“Тюркские”	5	2	7	2	4	4	5	7	12	3
Турецкий	2	8	2	1	1	2	4		2	2
Казахский	26	8	11	8	7	5	3	5	2	10
Татарский	6	5	10	1	8	4	6	6	11	5
Узбекский	1		2		3	13	2	3	8	3
Кыргызский	1		1	2	5	2	1		1	
Башкирский	1	1	1		1	1		2	2	
Якутский (саха)			3		2	2				
Тувинский			2		2	1		1	11	
Чувашский			3		4					
Азербайджанский					1	3	2	1	1	
Хакасские							1	1	1	
Английский	4	1	3	1	5	3		1	1	
Русский	4	1	5	1	4	2	2	8	6	1
Китайский	1			1						

ОЧЕНЬ формальный подход, подсчёт буквальных вхождений; точно есть погрешности + это всего лишь одна конференция, т.е. картина — неполная

К чему это всё сейчас было? [1/2]

Для решения даже самых распространённых на практике задач обработки кыргызского языка “открытых” **ресурсов и инструментов не хватает!**

Почему?

- 1) Не так велика потребность? (она есть)
- 2) Финансирование? (бизнес и государство предпринимают шаги)
- 3) Отсутствие базовых инструментов мешает разработке более сложных?

Изначально в обработке языка инструменты появлялись в том числе как результаты выполнения домашних заданий или подготовки курсовых :)

К чему это всё сейчас было? [2/2]

К сожалению, без набора

- инструментов обработки языка и
- языковых ресурсов (оцифрованных словарей, тезаурусов, примеров парафразы, морфологической разметки, синтаксических treebanks и т. д.)

...двигаться дальше — практически невозможно

Почти любая из самых стандартных задач требует чего-то, чего для кыргызского языка ещё нет!

Зачем вообще нужны датасеты и языковые ресурсы?

Видел работы, где “из головы” придумывается метод и докладывается **без численных оценок качества на общедоступных данных**

Такой подход **не позволяет убедительно** демонстрировать улучшения и, в целом, сравнивать подходы

Чтобы это изменилось, можно:

- подготовить размеченные наборы данных (например, перевести),
- получить на них оценки качества доступных и новых подходов

Так можно сделать практически для любой задачи NLP применительно к кыргыз тили

Нюанс

Критика: “Разметить текст и применить известный метод — это не научная задача” — отчасти справедливо!

Но как-то двигаться дальше всё равно надо :)

+ с новыми задачами **неизбежно выяснится что-то новое**

Что касается **ценности**, в целом, в NLProc-сообществе очень приветствуется подготовка новых наборов данных

Есть специальные конференции, посвящённые языковым ресурсам: **LREC** (CORE C; Scopus, WoS) и другие

Наиболее престижные конференции по NLP

Акроним	Rnk	Комментарий	Крайний срок подачи
Наиболее престижные конференции по NLP			
ACL	A*	Самая престижная NLP-шная конференция	Срок подачи прошёл: https://2023.aclweb.org/
EMNLP	A	Самая релевантная для прикладных работ и эмпирики; тоже очень престижная	abstract: June 16, 2023, paper: June 23, 2023 https://2023.emnlp.org/calls/main_conference_papers/
COLING	A	Тожe довольно престижная конференция, работа из любой части NLP должна вкатить	В этом году не будет, раз в два года проводится: https://aclanthology.org/venues/coling/
NAACL	A	Тожe довольно престижная конференция, работа из любой части NLP должна вкатить	Пока тишина, никаких анонсов https://naacl.org/ https://twitter.com/naacl
EACL	A	Менее престижная, но читаемая конференция, работа из любой части NLP должна вкатить	Только что прошла: https://2023.eacl.org/
ECIR	A	Флёр информационного поиска, но NLP тожe ОК	Прошла в апреле https://ecir2023.org/
CICLing	B	Конференция поменьше, но тожe уважаемая	Кажется, в этом году не будет
LREC	C	Максимально релевантная конференция, хоть и не очень крутая: посвящена языковым ресурсам	Проводится раз в два года, в этом не будет http://www.lrec-conf.org/

Релевантные конференции

По тюркским языкам			
ICTL	–	???	дедлайн 31 марта https://ictl.uni-mainz.de/
TurkLang	–	Компьютерная обработка тюркских языков (РИНЦ в 2023)	дедлайн 1 октября http://www.turklang.net/ru/
Tu+		Workshop on Turkic and Languages in Contact with Turkic , публ. Linguistic Society of America (похоже, не CompLing)	дедлайн был в январе https://sites.google.com/view/tuplus8workshop
ConCALL		Американская конференция по языкам ЦА, лингвистика	что-то уже 2 года не наблюдаю
Просто Scopus из ближайших			
AIST	Ru	Хорошая изначально российская конференция, Scopus; публикации будут в Springer, но могут появиться не сразу	30 июня <u>эбстракты</u> , 15 июля тексты https://aistconf.org/

Можно также подать на ARR: <https://aclrollingreview.org/cfp> Рецензирование 2 месяца, в случае успеха возможность подать на некоторые из мероприятий из прошлого слайда (никогда не пробовал).

План

- ~~1. Обработка (кыргызского) языка~~
2. Задачи, рассматриваемые в рамках диссертации
3. Другие открытые (?) задачи

Задачи, рассматриваемые в рамках диссертации

1. Тематическая классификация текстов (multi-label)
 - а. В том числе оценка качества неконтекстных векторных представлений
2. Выделение именованных сущностей
(с использованием переводной и ручной разметки)
3. Тематическое моделирование

Общая цель — разработка соотв. наборов данных и выработка эффективных стратегий обучения методов агрегации информации

Обработка естественного языка

В рамках диссертационного исследования будут разработаны интеллектуальные методы, позволяющие эффективно обрабатывать большие объёмы текстовой информации на кыргызском языке

- текстовая классификация (тематическая)
- выделение (“распознавание”) именованных сущностей
- тематическое моделирование

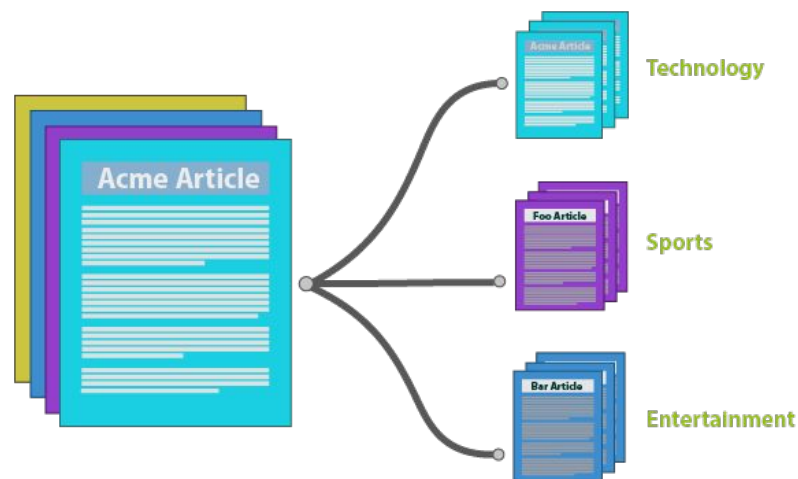
В первую очередь потребуются подготовить наборы данных для настройки и оценки качества соответствующих методов машинного обучения

1. Текстовая классификация

Отнесение текста к одной из заранее заданных категорий; примеры:

- речевой акт в диалоге: вопрос, уточнение, согласие, отказ, приветствие, прощание и т. д.
- тип документа,
- **темы новостей/запросов в техподдержку/поисковых запросов,**
- положительный или отрицательный тон у комментария/реплики,
- содержит ли предложение совет,
- оскорбителен ли текст комментария,
- и другое

Существует также постановка с одновременным отнесением документа к нескольким категориям: **многозначная классификация** (multi-label classification)



1. Текстовая классификация

Конкретнее — для начала подготовка **бенчмарка**

- Тексты 24.kg в разделе “Кыргызча”, более 23к
- Вручную размечены **тематики** 1’500 текстов
- Методы:
 - Multi-Label k Nearest Neighbours,
 - Multi-Label Support Vector Machine и другая классика,
 - нейросетевые методы на основе разных рекуррентных сетей,
 - нейросетевые методы на основе предобученных “Трансформеров” (например, **multi-lingual-bert-cased**)

Вопросы: какая токенизация помогает лучше всего? поможет ли морф. анализатор? многоязычный BERT — способен помочь? и т. д.

1а. Неконтекстные векторные представления

Дистрибутивная семантика: ‘oculist and eye-doctor... **occur in almost the same environments**’, ‘If A and B have almost identical environments. . . we say that they are synonyms’ (З. С. Харрис)

‘You shall know a word by the company it keeps!’ (Дж. Р. Фирс)

Я купил *ИКС* в ближайшем хозмаге.

Пришёл домой, растянул *ИКС* на балконе, повесил брюки.

Заключённые спустились по *ИКС* из окна камеры.

Что можно подставить вместо ***ИКС***? Список вариантов можно **ВЫЧИСЛИТЬ**

1а. Неконтекстные векторные представления

Зачем? Например, чтобы использовать в классификации коротких текстов как признаки для нейросети

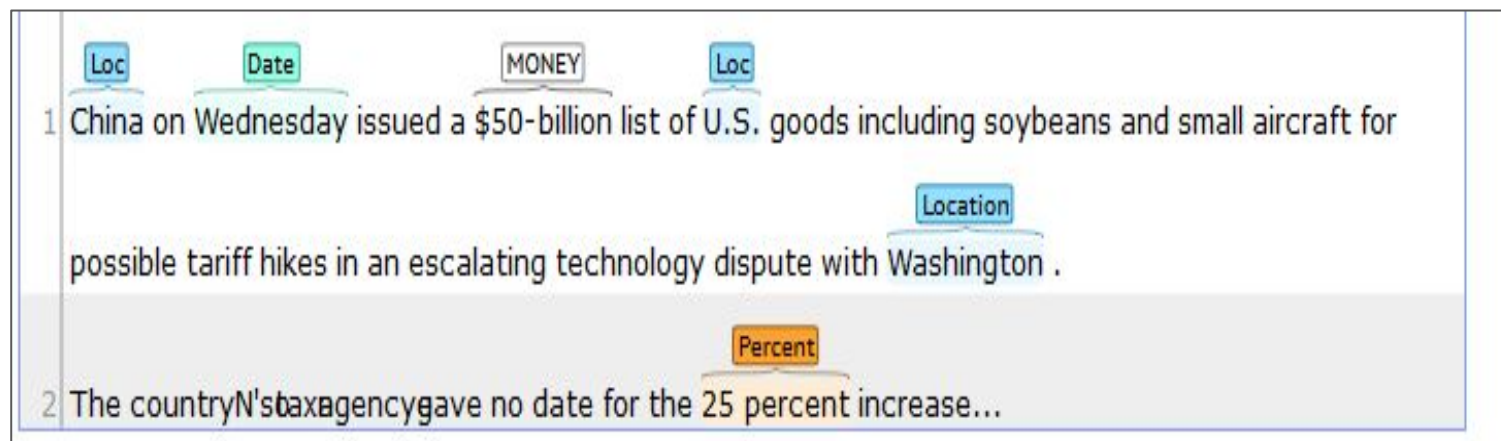
Конкретнее — подготовка **бенчмарка**

- Переведён на кыргызский язык наиболее полный из русских **набор для оценки качества эмбеддингов НЈ**
- Обучены векторные представления

НО: в дальнейшем стоит подготовить такой же набор непереведённых пар

(так как в НЈ неизбежно много заимствований из русского + некоторые пары потеряли смысл)

2. Выделение именованных сущностей



1 China on Wednesday issued a \$50-billion list of U.S. goods including soybeans and small aircraft for possible tariff hikes in an escalating technology dispute with Washington .

2 The country's tax agency gave no date for the 25 percent increase...

Выделение имён, топонимов (или адресов), дат, названий организаций – из текстовых сообщений, статей, новостей, постов и т. д.

- Словари, газеттиры, рег. выражения и иные **эвристики справляются не всегда** (конечны + как правило, слишком много вариантов написания)
- Результаты работы используются для последующей **агрегации** и иных задач information extraction, например, для **выделения из текста отношений**

2. Выделение именованных сущностей (NER)

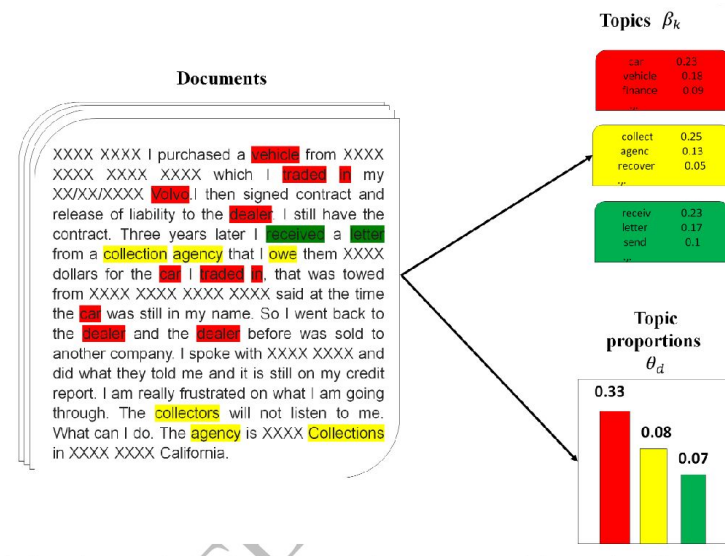
Также подготовка бенчмарка

- обучающая выборка (“серебряный стандарт”)
 - 14к предложений CoNLL2003 переведены на кыргызский “Я.Переводом”
 - предстоит придумать и отладить метод, как поточнее **перенести метки с английского** на кыргызский перевод (с помощью word alignment)
 - на таком наборе **можно будет обучать NER-методы** (**пока** набора такого размера для кыргызского просто нет) — для выделения ORG, LOC, PER
- также классические стат. методы и более современные подходы

А прямо сейчас (спасибо!) — ведётся ручная разметка (студентами-практикантами КГТУ и волонтерами) набора данных для оценки качества NER, это ещё ценнее

3. Тематическое моделирование

- Задача “обучения без учителя” – по **неразмеченной** коллекции текстов выделить набор тематик; по сути, вероятностная бикластеризация, т.е. одновременная кластеризация документов и тем
- **Тема – распределение над словами, документ – распределение над темами**
- Обучение – максимизация вероятности порождения коллекции текстов моделью (варьируются параметры распределений)
- Широко применяется в гуманитарных науках и для предварительного эксплораторного анализа любых крупных наборов текстов



3. Тематическое моделирование

Разработка “ноу-хау” (в целом, просто пайплайна) для эффективного тематического моделирования на кыргызских текстах

- Переиспользование новостных корпусов, их автоматическая предобработка (токенизация, лемматизация, фильтрация лемм по частоте)
- Апробирование классических байесовских методов (pLSA, LDA, ARTM и др.) + нейросетевых (ETM, ProdLDA, top2vec, BERTopic)
- Адаптация соответствующих скриптов для оценки качества (оценки когерентности, diversity, сопоставление с темат. метками)

Насколько мне известно, этого никто не делал

План

- ~~1. Обработка (кыргызского) языка~~
- ~~2. Задачи, рассматриваемые в рамках диссертации~~
3. Другие открытые (?) задачи

Однако этими задачами NLP не ограничивается :)

Рассмотрим наиболее “популярные”, “мейнстримные” задачи, в которых ещё можно очень много сделать

Условно поделим на несколько групп

- 1) уровень “ниже слов” — символы, морфемы и пр.
- 2) “уровень слов” — семантика и пр.
- 3) “уровень предложений” — синтаксис и извлечение информации
- 4) “уровень документа”

(1) Морфология

- **Частеречная разметка (PoS-tagging)**
- **Морфологический анализ:** т. е. более подробная разметка (см. постер)
- **Лемматизация:** приведение к нормальной форме слова (к лемме, неопределённой форме, инфинитив)
- **Морфологическая сегментация:** деление слов на морфемы (разбиение на корень, суффиксы и т. п.)

Инструменты: сейчас из **открытых** инструментов есть только не разрешающий морфологическую неоднозначность **apertium-kir**

Данные:

- КТМУ: [780 предложений](#) с морфосинтаксической разметкой (открытый доступ)
- КТМУ: корпус из 2.5М словоформ — осенью собираются открыть
- [Verbal paradigms for Kyrgyz \(100 Kyrgyz verbs fully conjugated in all tenses\)](#) by Aytmatova Alima, annotation for Unimorph by E. Chodroff

Минимальный пример работы с apertium-kir (Python)

```
import apertium
import streamparser
from typing import Iterable

apertium.installer.install_module("kir")
analyzer = apertium.Analyzer("kir")

text = "Апам рамканы жууду."
apertium_parsed = analyzer.analyze(text)
apertium_parsed = "^" +
    "$^".join([str(an)
                for an in apertium_parsed]) + "$"
lexical_units = streamparser.parse(apertium_parsed)

for lexical_unit in lexical_units:
    print(lexical_unit)
```

Output:

Апам / *Апам

рамканы / *рамканы

жууду / жуу<v><tv><ifi><p3><pl>

/ жуу<v><tv><ifi><p3><sg>

./.<sent>

<v> = глагол,

<iv> = транзитивный,

<ifi> = прошлое определённое,

<p3> = третье лицо,

<pl> = мн. число,

<sg> = ед. число

(1) Иные задачи

- **Постановка слов в заданную форму**
(apertium-kir, кажется, умеет)
- **Исправление орфографических ошибок**
(если не ошибаюсь, занимается А. Б. Турдубаева)
- **Транслитерация в обе стороны**
 - в латиницу — несложно, есть работы кыргызоведов, в которых предлагаются конкретные схемы + наверняка есть государственный стандарт
 - из латиницы обратно — сложнее
 - другие варианты! например, арабица для кыргызского из СУАР (Китай)

(2) Семантика!

- **Машиночитаемый тезаурус**
(какой-то словарь синонимов есть у tamga software, но там строгая лицензия; полезная вещь во многих задачах)
- Словари **слов с “тональной окраской”**
(вещь простая, сложно лишь обосновать выбор этих слов)
- Наборы данных: **оценка “сходства слов” и “отношений между словами”**

Помимо моего переводного, хорошо бы сделать ещё один с нуля + набор, на котором можно проверять “отношения”:

$$v(\text{король}) - v(\text{мужчина}) + v(\text{женщина}) = v(\text{королева})$$

(3) Синтаксис

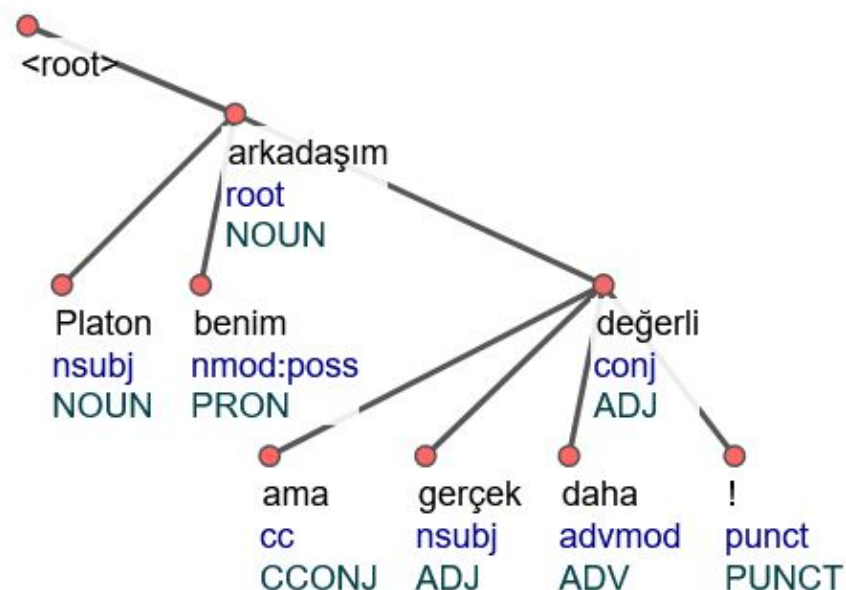
Автоматический **синтаксический разбор** для тюркских языков сложен, но наработок для других языков — много

Данные: выложены деревья от КТМУ ([780 предложений](#)), на этих данных обучаться нельзя, но **можно проверять качество**

Есть как минимум парсеры для турецкого и уйгурского языков

В целом, сложная задача, непонятно, как подступиться, когда нет большого TreeBank-а

Platon benim arkadaşım ama gerçek daha değerli !



(3) Около синтаксиса: везде нужны наборы данных, хотя бы небольшие

- **Shallow parsing/Chunking**: выделение отдельных групп (phrases) (как в грамматике непосредственно составляющих)
- **Выделение сочинительных связей** (нужен набор данных для оценки!)

Susan works slowly and carefully

- **Разрешение анафор**: к чему в тексте относится местоимение? (можно попробовать разметить на тех же текстах, что и NER-24kg)
- Исправление **синтаксических ошибок**
- Распознавание и “исправление” **эллипсиса** (вставка “пропущенных слов”)

(3) Извлечение информации: что есть, кроме NER

- **Entity Linking**: после NER делают сопоставление элементам баз знаний (например, Википедии или списка персоналий новостного сайта)
- **Relation Extraction**: какая компания какую купила? кто что сделал по отношению к кому? Часто использует NER как предварительный этап
- **Open Information Extraction** ([мой большой обзор](#)): большая надежда — суметь построить этим методом базы знаний; **данные**: WiRe57 можно легко перевести, чуть сложнее перенести разметку – ТОЛЬКО руками

“John managed to open the door”

надо бы извлечь:

(John; managed to open; the door)

извлекать необязательно:

(John; opened; the door)

(4) Уровень документа

- **Анализ тональности текста**
(непонятно, зачем нужно, люди отзывы пишут, в основном, по-русски)
- **Другие задачи классификации:** intent classification
(полезно для поиска и чатботов), speech act classification
(полезно для чатботов), suggestion classification, и другое
- **Информационный поиск**
- **Кластеризация текстов**
- **Тематическое моделирование**
(например, применительно к эпосу)

(4) Задачи, часто требующие генеративного и/или ультрасовременного подхода к решению

- **Языковое моделирование (нейронное)**
как ни крути, придётся исследовать наиболее эффективные стратегии дообучения мультязычных моделей на кыргызских текстах: **нужен большой корпус**
- **Упрощение текста** (например, для не являющихся носителями языка)
- **Перефразирование** (и распознавание, что суть фраз одна и та же)
- **Реферирование текста** (text summarization)
- **Машинный перевод** (есть методы, есть данные, но это бездонная задача)
- **Вопросно-ответные системы** (e.g. дан абзац, надо ответить на вопрос)
- **Диалоговые системы**
- **Мультимодальные задачи**: не только текст, но и картинки/видео/etc.

Заключение

Во всех перечисленных задачах

- **желательно** собрать **набор данных для обучения** (train set; не всегда вручную!)
- **требуется** собрать **набор данных для оценки** (test set)
- для корпусных исследований и обучения self-supervised-моделей требуется также **большой** набор “чистых” текстов, т. е. корпус
- есть место для применения как самых модных методов, так и подходов на правилах (когда сложно собрать train set)
- малый задел в части воспроизводимых численных методов

То есть работы, на мой взгляд, хватит надолго и для всех

Спасибо за внимание!

Антон М. Алексеев

2023