Лингвоспецифическая разметка китайских текстов в Русско-китайском параллельном корпусе НКРЯ

Вольф Елена Александровна, ФиКЛ, выпуск 2021 года
Короткова Юлия Олеговна, ФиКЛ, 3 курс

Поставленные цели

Разработать алгоритм автоматической разметки китайских предложений:

- деление на слова
- присвоение грамматической информации
- присвоение транскрипции

А также:

- разработать открытые наборы данных (датасеты)
- написать и опубликовать статьи

Результаты

Поставленные цели — выполнены!

- Сделано 3 датасета по 850 предложений для алгоритмов словоделения, и 1 датасет из 1350 для алгоритмов аннотации пиньиня
- Проведены эксперименты и исследования инструментов для словоделения, PoS-тэггинга, присвоения аннотации пиньиня
- Улучшен алгоритм разметки китайских текстов
  - Репозиторий: https://github.com/ruzhcorp/ruzhcorp_chinese_annotation
- Документы корпуса полностью переразмечены новым алгоритмом

Новая разметка

Ссылка на корпус: https://linghub.ru/rnc_parallel_chinese/search

Примеры запросов:

1) Заимствования из русского (*Агафья*)
2) Традиционные иероглифы (*為 VS 为, 臺灣 VS 台湾*)
3) Пиньинь (*yī, yǐ*)
4) Фонетически омонимичные слова (*了*)

Выступления на конференциях

- Linguistics Colloquium (онлайн). Automatic Chinese Word Segmentation in the Translated Texts: Case Study of the Russian-Chinese Parallel Corpus of RNC
- Buckeye East Asian Linguistics Forum (Огайо/онлайн) Comparative analysis of grapheme-to-phoneme models for the Russian-Chinese parallel corpus;
- Международная конференция «Диалог» (Москва/онлайн). Автоматическая лингвистическая разметка китайских текстов, содержащих заимствования: словоделение, транскрипция, PoS-тэггинг;
- Международная научная конференция «Корпусная лингвистика 2021» (Санкт-Петербург/онлайн). Автоматическая разметка заимствований из русского языка в китайских текстах: проблемы словоделения и морфопарсинга;
- Corpus Linguistics International Conference “CL2021” (Лимерик/онлайн). Enhancing Loanword Detection in the Chinese Texts via Code-Switching and Fine-Tuning: Case Study on the Russian-Chinese Parallel Corpus of RNC.

Запланированные выступления

- XXIV Международная научная конференция «Китай, китайская цивилизация и мир. История, современность, перспективы» (Москва/онлайн). Лингвистическая разметка китайских текстов в Русско-китайском параллельном корпусе НКРЯ;
- 11th International Conference “Slovko 2021” (Братислава/онлайн). Linguistic annotation of translated Chinese texts: Coordinating theory, algorithms and data

Выпускные квалификационные работы

- Александра Коновалова, «Автоматический частеречный анализ для китайского языка с привлечением данных параллельного корпуса»
- Армине Титизян, «Определение переключения кодов в текстах на китайском»

Перспективы

- улучшить функционал поиска по нашей разметке
  - вести поиск по китайским пос-тэгам (к зиме)
  - вести поиск по “обобщенной” орфографии (к зиме)
- выложить на сайте большого НКРЯ (к зиме)
- возможно, добавлять новую разметку
  - семантическую
  - синтаксическую
- мы поможем другим группам - например, той, которая делает пословное выравнивание (как context reverso)

Более подробная информация — на сайте ФГН НИУ ВШЭ:
https://ling.hse.ru/ruzhcorp_annotation

Спасибо за внимание и за предоставленную помощь!

Благодарны советам и открыты для сотрудничества!

ruzhcorp@yandex.ru