Проектные задачи
компьютерной лингвистики

Автоматическое глоссирование малоресурсного языка

Савина Дарья

4 курс

2025

Проблема

***

Language documentation is a critical aspect of language preservation, often including the creation of Interlinear Glossed Text (IGT). Creating IGT is *time-consuming and tedious*, and automating the process can save valuable annotator effort.

[Ginn 2023]

Проблема автоматизации

Плюсы:

① Более последовательная разметка

② Быстрее и эффективнее с большими данными

Минусы:

① Недостаток данных для малоресурсных языков

② Для изолятов нет возможности дообучить модель с родственным языком

Этапы разметки

Оригинал: к’ыскӷу тиӷр к’сот мыринысь

Перевод: Кошки стали подниматься по дереву.

Морфологическая сегментация

к’ыск-ӷу	тиӷр	к’со-т	мыр-ины-дь

Глоссирование как задача классификации (Labeled Segmentation + IG)

к’ыск-ӷу	тиӷр	к’со-т	мыр-ины-дь
STEM.NOUN-PL	STEM.NOUN	STEM.VERB-CONV.3.PL	STEM.VERB-INCH-IND
к’ыск-ӷу	тиӷр	к’со-т	мыр-ины-дь
КОШКА-PL	ДЕРЕВО	ВЛЕЗАТЬ-CONV.3.PL	ПОДНИМАТЬСЯ-INCH-IND

Данные

Возможных классов: 88

Количество морфем: ~300

	Gloss	Category	Morph
0	AVERT	aspect	ирут’ез
1	COMPL	aspect	гар, гит, гыт, ғар, ғыр, ғыр, ғыт, ҕыт, кар, хыт, ғр, ғырт, кир
2	CONC	aspect	гин
3	DIM	aspect	ё
…	…	…	…
84	IND	mood	ҕана, ҕан, д, дь, нд, т, ть, к, кан, қан
85	PL	number	го, гу, гун, гуну, ғу, ғун, ҕун, ҕуну, ку, куну, у, ху
86	FUT	tense	н, на, ны, ы, н, нин, ңы
87	CAUS	voice	ң, ңг, к, г, гу, ғу, ку, қу

Предшествующие исследования

Метод

X-ICL (crosslingual incontext learning)

Подходы

Рандомное сэмплирование

Aggregate WordRecall

Сэмплирование на основе WordRecall / WordPrecision

Метрика chrF

Morpheme Recall (используется Morfessor)

We study strategies for selecting in-context examples, finding significant impacts to performance. Our best-performing systems outperform transformer model baselines, despite involving no training whatsoever.

[Ginn et al. 2024]

Предшествующие исследования

[Coates 2023]

- An Ensembled Encoder-Decoder System for Interlinear Glossed Text

[Elsner & Liu 2025]

- Prompt and circumstance: A word-by-word LLM prompting approach to interlinear glossing for low-resource languages.

[Ginn et al. 2024]

- Can we teach language models to gloss endangered languages?

[Elsner & Liu 2025]

- Boosting the Capabilities of Compact Models in Low-Data Contexts with Large Language Models and Retrieval-Augmented Generation

Исследовательская цель

Посмотреть на способности русскоязычных моделей справляться с задачей автоматического глоссинга методом few-shot.

Модели

YandexGPT/GigaChat

План работы

Подходы

- few-shot: случайные примеры + список возможных глосс
- RAG-подобная модель, возвращающая примеры с похожими морфемами

*В идеале: RAG для ретрива релевантной информации из грамматики для правильного глоссирования

Две модели

- Пайплайн: сегментация + классификация глосс
- «Совмещённая» модель

GlossLM

①

the largest corpus of multilingual IGT data

②

pretrained multilingual neural model for automatic generation of IGT

③

achieved a new SOTA on automatic IGT generation (5/7 языков)

Сравнение дообученного GlossLM
с лучшими результатами few-shot.

Метрики

① Morpheme Accuracy

② Word Accuracy

③ precision, recall, F1-мера

SIGMORPHON 2023 Shared Task on Interlinear Glossing

* В идеале: подсчитать метрику chrF
(character n-gram F-score for automatic MT evaluation)
[Popović 2015]

Литература

Okabe & Yvon. 2023. *Towards Multilingual Interlinear Morphological Glossing.* Findings of the Association for Computational Linguistics: EMNLP 2023, pp. 5958–5971.

Ginn. 2023. *SIGMORPHON 2023 shared task of interlinear glossing: Baseline model.*

Ginn et al. 2024. *GlossLM: A Massively Multilingual Corpus and Pretrained Model for Interlinear Glossed Text.* Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pp. 12267–12286.

Shandilya & Palmer. 2024. *Boosting the Capabilities of Compact Models in Low-Data Contexts with Large Language Models and Retrieval-Augmented Generation.*

Elsner & Liu. 2025. *Prompt and circumstance: A word-by-word LLM prompting approach to interlinear glossing for low-resource languages.*