Сравнение статистических характеристик корпусов деревьев зависимостей

Шамаева Е.Д.

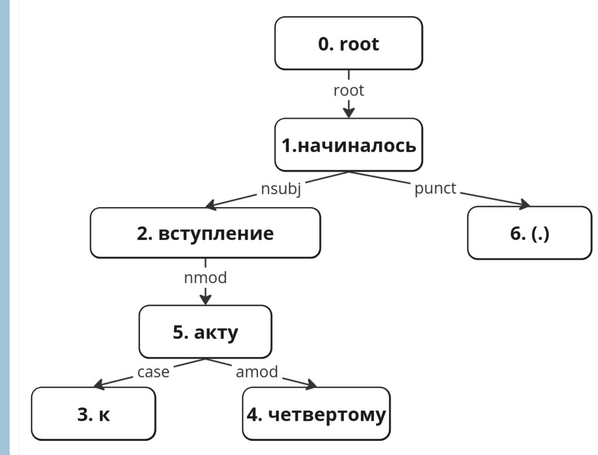
МГУ имени М.В. Ломоносова

Оценка результатов синтаксического анализа

Дерево зависимостей: вершины — токены предложения (слова, знаки препинания), ребра — отношения между токенами

Оценка результатов синтаксического анализа [1, 2]:

1. группировка токенов тестовых предложений: по предложениям или по определенному признаку токенов
2. вычисление доли корректно определенных главных токенов (и опционально типов связи) внутри группы
3. для группировки по предложениям — усреднение по предложениям долей «корректных» токенов



Цель исследования

Для каждого вида разбиения на группы

- оценить количество элементов каждой группы
- сравнить соотношения групп между разными языками

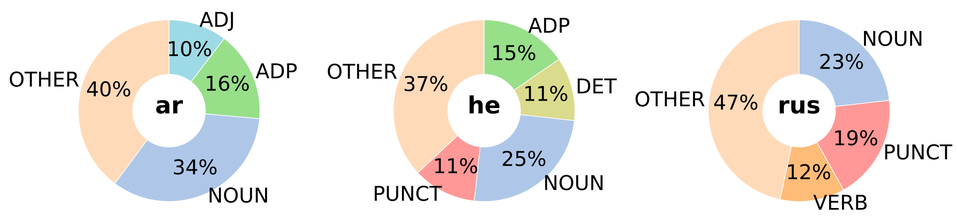
Корпуса синтаксически размеченных текстов

В данном исследовании использованы тестовые выборки корпусов текстов из проекта Universal Dependencies [3], набор языков — из [2]

Датасет	AR_PADT	BDT_BDT	CHI_GSD	ENG_EWT	FI_TDT	HE_HTB	HI_HDTB	IT_ISDT	JA_GSD	KO_GSD	RU_SynTagRus	SV_Talbanken	TR_IMST
Объем	680	1799	500	2077	1555	491	1684	482	543	989	8800	1219	1100

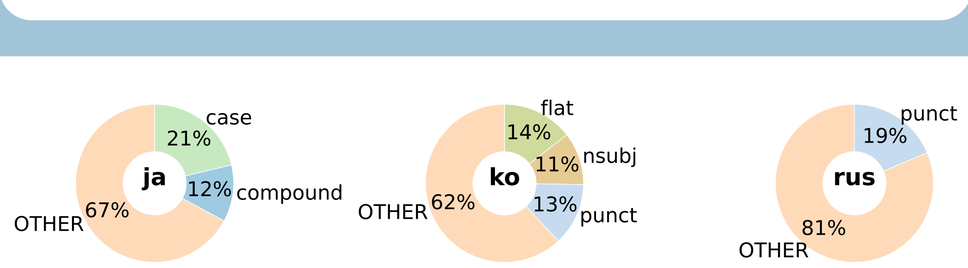
Группировка токенов по частям речи

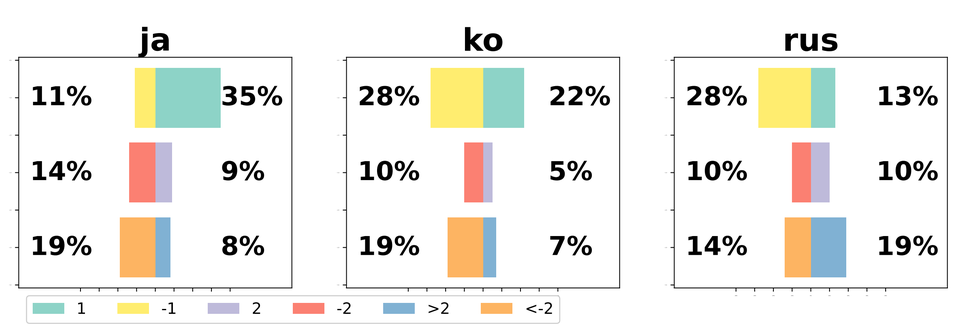
- В разных языках количество частей речи колеблется от 14 до 17
- Во всех языках не более 3-4 частей речи встречаются часто (10% и более)
- Практически для всех языков более 80% групп содержит более 100 элементов
- Во всех языках существительные (NOUN) — наиболее частотная группа



Группировка токенов по типу связи

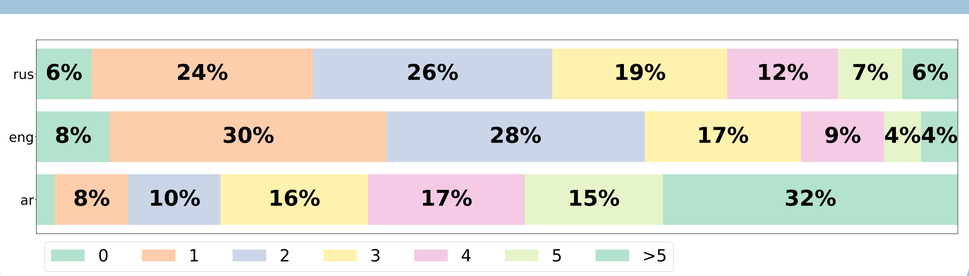
- В разных языках количество эталонных типов связей колеблется от 24 до 48
- Во всех языках не более 1-3 эталонных типов связей встречаются часто (10% и более)
- Только в хинди, арабском, русском и идише доля достаточно частотных групп (100 и более элементов) превышает 70%





Группировка токенов по расстоянию до главного токена

- Во всех языках для более 60% токенов расстояние до соответствующего главного токена не превышает 2
- Во всех языках небольшое количество токенов находится на расстоянии более 10 от соответствующего главного токена
- Только в арабском, идиш и японском количество токенов, стоящих левее от главного токена, меньше количества токенов, стоящих правее от главного

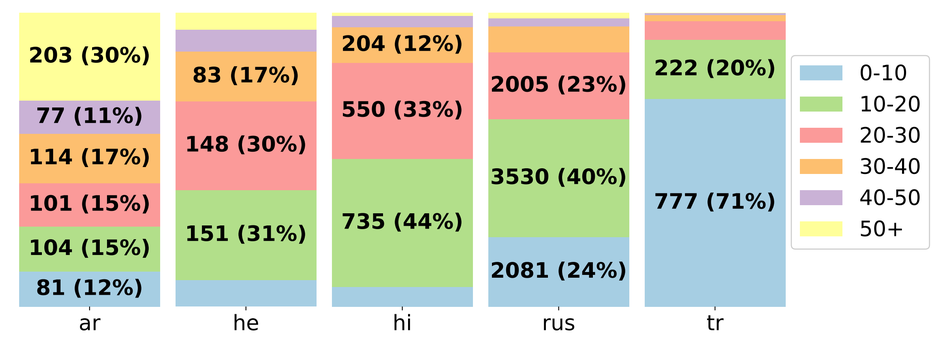


Группировка токенов по расстоянию до корневого токена

- Только в арабском языке более 30% токенов находится на глубине более 4 от корневого токена
- В остальных языках более 60% токенов находится на глубине 1-3 от корневого токена
- Для большинства языков группы токенов глубиной более 7 не являются достаточно представленными (менее 100 токенов)

Статистика по группам предложений с заданной длиной

- В большинстве языков более 50% предложений короче 30 токенов
- В большинстве языков группа длиной 10-20 токенов содержит больше всего предложений
- Группы предложений длиной 30-40 токенов имеет смысл анализировать в хинди, арабском и русском, 40-50, 50+ — только в арабском
- Почти в половине языков размер любой группы меньше 500 предложений (арабский, идиш, итальянский, китайский, корейский и японский)



Результаты

- При оценке качества синтаксического анализа на группе предложений или токенов необходима проверка объема группы
- Пропорциональное соотношение объемов групп существенно зависит от языка

Список литературы

1. Zuhra F. T., Saleem K. Hybrid embeddings for transition-based dependency parsing of free word order languages. Information Processing & Management. 2023. T. 60.
2. Kulmizev A., Lhoneux M. et al. Deep Contextualized Word Embeddings in Transition-Based and Graph-Based Dependency Parsing — A Tale of Two Parsers Revisited. EMNLP-IJCNLP. Hong Kong, China: Association for Computational Linguistics, 2019. 2755–2768.
3. Marneffe M., Manning C.D., Nivre J., Zeman D. Universal Dependencies. Computational Linguistics. 2021. 47 (2). 255–308.

Реализация исследования: https://github.com/Derinhelm/treebank_statistics