Морфологический разбор и вероятностные модели агглютинативных языков

Эллэй Шамаев eshamaev{at}mail.ru
Алена Кирилловна Прокопьева СВФУ
Оксана Афанасьевна Домотова СВФУ
Олеся Дмитриевна Слепцова СВФУ

Дистрибутивная гипотеза

Лингвистические единицы, встречающиеся в схожих контекстах, имеют близкие значения.

Наибольший вклад был внесен в 1950-х годах
Л. Блумфилдом (Leonard Bloomfield) и З. Харрисом (Zellig Harris).

Дистрибутивная гипотеза

винтокрылый летательный полет
многоцелевой поставки крушение десантный
рухнул упал боевые воздушное билет

вертолет, самолет, ракета

авиакомпании российские

дорога, шоссе.

WordPiece, Token Embeddings и Bert

Текст на английском

кодирование с помощью словаря WordPiece с размером 30 522 слов

кодирование в векторы 768 или 1024 с помощью Token Embeddings
(обучается вместе с BERT)

обработка трансформером BERT
(сложно объяснить, но ...)

Текст на агглютинативном языке

кодирование с помощью словаря (токенизация)

кодирование в векторы 768 или 1024 с помощью Token Embeddings

обработка трансформером

WordPiece: весь английский в словаре 30 тыс. слов

токенизация на токены

playing -> play _ing

mysophobia -> my _so _phobia

Токенизаторы

WordPiece, 2012 - алгоритм, часть токенов бессмысленна

Unigram LM, 2018 - алгоритм, часть токенов бессмысленна

BPE Byte pair encoding, 1994 - алгоритм, выдает бессмысленные токены

Учитывать морфологию

При подсчете статистики (обучении нейросети) слово “сөтүө-лээ-биппит-тэн” в тексте практически не встретится.

Нехватка данных для подсчета статистики.

Нужно “сөтүөлээ #быппыт #тан”.

Токенизатор должен учитывать морфологию

Эксперименты с моделями языка

Natural Language Processing Methods for Language Modeling
Derivational Morphology Improves BERT's Interpretation of Complex Words
KR-BERT: A Small-Scale Korean-Specific Language Model
Byte Pair Encoding is Suboptimal for Language Model Pretraining
How Much Does Tokenization Affect Neural Machine Translation?
и др.

Токенизатор ≠ морфологический анализатор

Нейросети с текущими токенизаторами показывают посредственный результат.

Состояние развития NLP в якутском языке

1. Не хватает размеченных эталонных корпусов.
2. Нет токенизатора, работающего с учетом морфологии слова.

Сделанные шаги:

1. Разрабатываются морфологические анализаторы. СВФУ
2. Русско-якутский машинный перевод. Яндекс
3. Оцифрованы тексты. Национальная библиотека
4. Нейросети для лемматизации ~89%. Вероника Николаева. СВФУ
5. Частеречная классификация ~85%. Оксана Домотова, А.К. Прокопьева, О. Д. Слепцова. СВФУ

Состояние развития NLP в якутском языке

https://github.com/domotova

Сорокин А.А. MORPHOLOGICAL PARSING OF LOW-RESOURCE LANGUAGES

https://github.com/nlp-sakha

Спасибо за внимание!

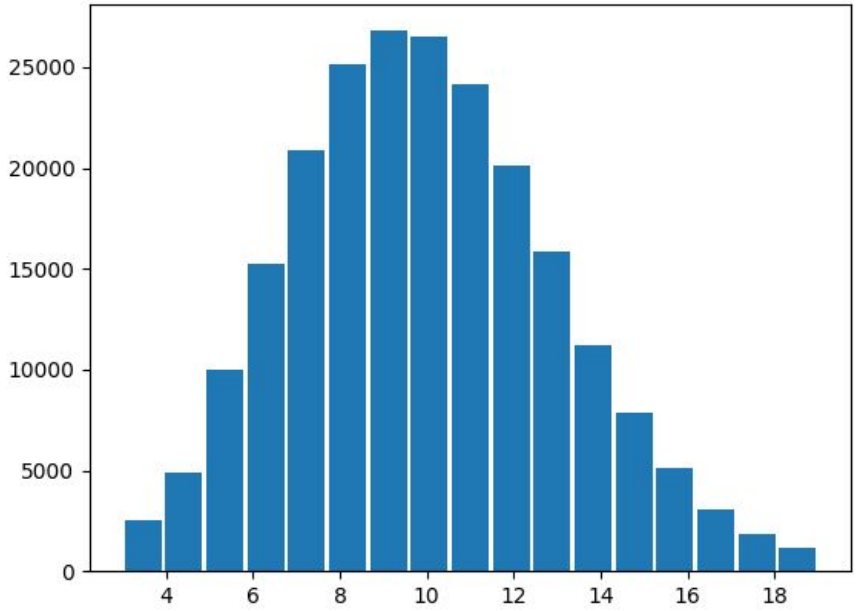
Английский, китайский

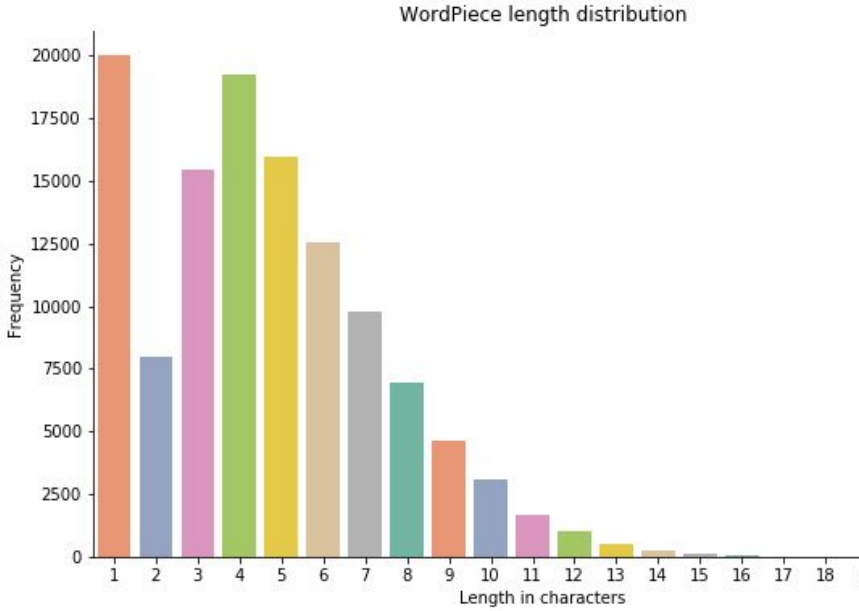
- Наиболее большие и качественные корпусы в млрд слов.
- Определенный порядок слов в предложениях английского языка.
- Большие инвестиции и большие исследовательские команды.

Агглютинативные

- *
- Многообразие словоформ.
- Прямая адаптация статистических (нейросетевых) методов, сконструированных для английского языка.

Многообразие словоформ





Средняя длина словоформы 9-10 букв. Средняя длина основы 5-6 букв.