Автоматическая морфоразметка корпусов русскоязычных социальных медиа: обучение и оценка качества

Селегей Д. [2], Шаврина Т. [1], Селегей В.[2,3], Шаров С. [2,4]

[1]Московский Государственный Университет, Россия

[2]Российский государственный гуманитарный университет, Россия

[3]ABBYY, Russia

04.06.2016

Актуальность

Для лингвистических исследований нужны большие размеченные интернет-корпуса актуального русского языка.

Корпуса в десятки миллиардов словоупотреблений непригодны для использования при отсутствии разметки.

Такая разметка может делаться только автоматически, но как получить высокое качество не только частеречной, но и всей разметки? Автоматическое снятие омонимии?

Проблемы морфоразметки русских текстов social media

Имеются и используются для разметки корпусов open-source т.н. POS-taggers. Например, tnt
Но: они обучены на вручную размеченном подкорпусе с существенно иной жанровой структурой -- т.н «снятике» НКРЯ. Прямой перенос их языковой модели на сегменты социальных медиа не позволяет получить морфоразметку нужного качества.
Отсутствует эталонная разметка (золотые стандарты) для оценки качества морфоанализа social media.
Есть разнообразные проблемы с используемым в разаметке тагсетом – это сейчас смесь противоречивых требований к обучению парсера и использования в корпусных запросах.

Задачи исследования

1. Создание нового стандарта морфоразметки (тагсета), сочетающего полноту категорий, их потенциальную определимость и полезность для исследователей.

2. Создание достаточно большого тестового корпуса (золотого стандарта), размеченного в соответствии с этим тагсетом и доступного общественности

3. Перенос этой разметки на 50-миллиардный ГИКРЯ

Предлагаемое решение

- Использование для разметки тестовогого и обучающего корпусов «тяжелого» синтаксического парсера (ABBYY Compreno).
- Разработка нового стандарта разметки на основании MSD с учетом известных проблем
- Мэппинг и оптимизация разметки Compreno на GICR-MSD
- Публикация золотого стандарта на сайте GICR
- Обучение TNT-парсера
- Преренос разметкаи на всеь GICR
- Эксперименты с другими схемами парсинга

```
69
70  #ПРИЛАГАТЕЛЬНЫЕ
71  A = PartOfSpeech::Adjective & GrammaticalType::(GTAdjectiveAttributive|GTAdjective)
72  A1::-                                                                          ## 4
73  A2::p/c/s      = DegreeOfComparison::DegreePositive/DegreeComparative/DegreeSuperlative
74  A3::m/f/n      = Gender::Masculine/Feminine/Neuter
75  A3::-          = Number::Plural                                                ## Ge
76  A4::s/p        = Number::Singular/Plural
77  A5::n          = (AdjectiveShortness::AdjectiveFullForm & Case::Nominative) | (AdjectiveShortnes
78  A5::g/d/a/l/i  = AdjectiveShortness::AdjectiveFullForm & Case::(Genitive|Partitive)/(Dative|Dati
79  A6::s/f        = AdjectiveShortness::AdjectiveShortForm/AdjectiveFullForm
```

Результаты

Мы представляем новый стандарт разметки, сочетающий разнообразие грамматических категорий и высокое качество автоматического снятия омонимии Abbyy Compreno и удобство популярного формата MSD. По запросу доступен корпус в 2 млн словоформ, а также на webcorpora.ru – демо-вариант в 50 000 словоформ.

Category of homonymy	**Old precision (on TnT)**	**New precision**
Inanimate nominative	0,792	0,947
Inanimate accusative	0,858	0,884
animate accusative	0,661	0,980
Animate genitive	0,890	0,890
Substantives	0,680	0,916
Adjectives similar with substantives	0,900	0,918

Качество частеречной разметки

На Живом Журнале:

Part of speach	Precision	Accuracy	F-measure
Noun	0,989	0,960	0,945
Verb	0,995	0,979	0,987
Adjective	0,978	0,978	0,978
Pronoun	1,000	0,896	0,945
Adverb	0,940	0,935	0,937
Preposition	1,000	0,972	0,986
Conjunction	0,927	0,962	0,944
Numeral	0,957	0,978	0,967
Particle	0,964	0,891	0,926
Interjection	1,000	0,585	0,738
Predicative	1,000	0,900	0,947
Parenthesis	0,954	0,807	0,874
Макроусреднение	**0,970**	**0,904**	**0,931**

Стандарт MSD-GICR

Новый тагсет и качество разметки на реальных текстах можно оценить на сайте: webcorpora.ru/news/282
(размещен демонстрационный вариант).

Всем желающим предоставляется золотой стандарт, содержащий 2 миллиона слов из Живого Журнала (публикации и комментарии к ним).

Мы призываем всех заинтересованных исследователей обратиться по адресу: geekrya@gmail.com
и использовать данный золотой стандарт для обучения парсеров и улучшения морфоразметки на текстах social media.

Спасибо за внимание!
Ждем вас у стенда

GEEKPЯ