Морфологический тагер TreeTagger

Е. Герасименко, А. Виняр

1. Система тэгов

Table 25. Category table		
Name (en)	Code (en)	Attributes
Noun	N	6
Verb	V	10
Adjective	A	6
Pronoun	P	7
Adverb	R	1
Adposition	S	3
Conjunction	C	4
Numeral	M	6
Particle	Q	1
Interjection	I	1
Abbreviation	Y	4
Residual	X	0

Тестовый файл – tests.txt, с тэгами – tests_result.txt.

1.1.
10 частей речи + аббревиатуры + остаток. По сравнению с разметкой НКРЯ, например – более общие категории, частные можно разнести по этим общим категориям. Например:

предикатив (жаль)
жаль R жаль

вводное слово (кстати)
Кстати R кстати

Но, кстати, парсер не всегда делает это корректно:

числительное-прилагательное (один, второй):
Второй Ncmsnn Второй
Одни P---pna один

1.2. Местоимения:
местоимение-существительное
Я P-1-snn я
она P-3fsnn она
Это P--nsnn это

местоимение-прилагательное
который P--msna который
мой P--msna мой

местоименное наречие
где P-----r где
тут P-----r тут

местоимение-предикатив
некого P---sgr некого

Кажется, все типы местоимений (по НКРЯ) распознаются как местоимения.

1.3. Причастия:

горящим	Vmpp-sna-ei	гореть
сделанное	Vmps-snpfpn	сделать
строящийся	Vmpp-smmfea	строиться
строившийся	Afpmsnf	строившийся

Обычно лемматизируются как глаголы, но причастия несовершенного вида прошедшего времени – как прилагательные, причем с проблемами в лемме.

1.4. Разные формы:

нашедший	Vmps-smafpn	найти
находившего	Afpmpnf	находившего

Это вписывается в общее поведение причастий (1.3). Видно, что второе причастие вообще не лемматизируется.

дал	Vmis-sma-p	дать
давал	Vmis-sma-e	давать

У этих глаголов разные леммы, то есть в этом анализаторе различие по виду – на уровне лемм, что тоже имеет право на существование, на наш взгляд.

1.5.

В качестве правила мы написали программу, берущую на вход от пользователя тэг личного местоимения 3 лица в формате TreeTagger и печатающую его аналог по формату Золотого Стандарта. Потребовалось сделать несколько замен и перестановок. Для наречий – заменить ‘R’ на ‘ADV’. Программа – transformation_rules.py

2. Сложные случаи

Файлы – difficult.txt, difficult_result.txt.

2.1. Токенизация

В целом, отлично, на наш взгляд.

Хорошо:

10,000$	Mc	10,000$
тыс	Ncfpgn	тысяча
рок'н'ролл	Ncmsan	рок'н'ролл
Санкт-Петербурге	Ncmsln	Санкт-Петербург
15го	Afpmsgf	15го
юго-западе	Ncmsln	юго-запад
Superстильно	R	Superстильно
...	SENT	...

Плохо:

Точки везде как SENT, даже после правильно распознанных сокращений.

г. как год, не как город, хотя после него идет имя собственное.

Что-то странное произошло с ‘препинания,представляете’ (нет пробела после запятой):

препинания	Ncnsgn	препинание
,	,	,
п	Afpnsnf	подобный
редставляете	Vmip2p-a-e	редставляете

Объединяет три точки в многоточие, а три вопроса подряд в одну последовательность не объединяет.

2.2. Незнакомые слова
Хорошо определяется тэг, но попыток определить лемму даже не предпринимается.

<u>отличие по одушевленности:</u>
шорьки Ncmpnn шорьки
Noun Type=common Gender=masculine Number=plural Case=nominative Animate=no

зелюки Ncmpny зелюки
Noun Type=common Gender=masculine Number=plural Case=nominative Animate=yes

<u>тут правильно определен падеж:</u>
(по) наве Ncfsdn наве
Noun Type=common Gender=feminine Number=singular Case=dative Animate=no

(в) мове Ncmsln мове
Noun Type=common Gender=masculine Number=singular Case=locative Animate=no

<u>тут тоже все хорошо!</u>
Пырялись Vmis-p-m-e Пырялись
Verb Type=main VForm=indicative Tense=past Number=plural Voice=medial Aspect=perfective

хрюкотали Vmis-p-a-e хрюкотали
Verb Type=main VForm=indicative Tense=past Number=plural Voice=active Aspect=perfective

2.3. Омонимия
Выбирается один вариант. В readme написаны опции, позволяющие выводить тэги с вероятностями больше заданной, но у нас не вышло их применить.

<u>Плохо:</u>
стали Vmis-p-a-p стать
есть Vmip3s-a-e быть
(ну уж либо сталь + быть, либо стать + есть)
видимо, проблема в зависимости разборов друг от друга – два варианта снятия омонимии для каждого, не на что опираться?

Не хочет отделять частицу (‘И тут зашли…’) от союза.

(на) печи Ncfpan печь
Noun Type=common Gender=feminine Number=plural Case=accusative Animate=no
(не было) печи Ncfsnnl печь
Вообще непонятный тэг, наверное, за этим стоит:
Noun Type=common Gender=feminine Number=singular Case=nominative Animate=no + locative
что все равно не очень, увы

<u>Хорошо:</u>
(в) столовой Ncfsln столовая
столовые (приборы) Afpmpaf столовый

<u>Очень хорошо!</u>
блестящие Afpmpaf блестящий
побрякушки

блестящие Vmpp-p-a-en блестеть
на солнце украшения

Структура нам кажется не очень ясной, иногда контекст срабатывает на ура (причастия vs прилагательные), иногда – очевидный – не очень (падежи). Ну и, конечно же, большая проблема с типами, едящими в цехе.

3. Анализ файла

Сопоставление первых 500 словоформ ЗС и выдачи тагера, а также выделенные ошибки тагера – в файле compar.xlsx.

Система приписывает всем токенам леммы и тэги.
3.1. однозначность – 1, анализатор снимает омонимию.
3.2. accuracy по леммам – 470/500 = 94%
3.2. accuracy по тэгам. Тут есть интересная вещь – в глаголах аспект как будто зеркально отражен – во всех глаголах (включая причастия) аспект должен быть противоположным. Выглядит как недоразумение, поэтому приводим две цифры, с учетом глагольного аспекта и без:
accuracy по всему – 420/500 = 84%
accuracy исключая ошибки в аспекте – 469/500 = 93,8%.