**Проект по тестированию морфологических тагеров (TreeTagger)**

*Проект выполнили: Зеленкова Валерия, Иванов Вячеслав, Суслова Дарья*

**1. Анализ системы тегов – Иванов Вячеслав**

**2. Функциональное тестирование программы TreeTagger – Суслова Дарья**

**3. Обработка файла с помощью анализатора – Зеленкова Валерия**

1. Анализ системы тегов

1.1. Сколько частей речи учитывает система; какие части речи в системе отсутствуют, а Вы считаете, что эти части речи необходимо выделять (ответ мотивируйте).

*Система учитывает следующие части речи:*

|  |  |
| --- | --- |
| **Название** | **Код** |
| Существительное | N |
| Глагол | V |
| Прилагательное | A |
| Местоимение | P |
| Наречие | R |
| Предлог | S |
| Союз | C |
| Числительное | M |
| Частица | Q |
| Междометие | I |
| Аббревиатура | Y |
| Другое | X |

*Мы считаем, что в TreeTagger можно было бы отдельно выделять причастия и деепричастия. Их он разбирает как формы глагола (идущий Vmpp-smafen идти; делая Vmgp---a-e делать), из-за чего разбор этих форм выглядит довольно большим.*

1.2. В какие pos-классы попадают местоимения?

*Все местоимения (личные, возвратные, относительные, притяжательные, указательные) попадают в pos-класс P (местоимения).*

1.3. Как лемматизируются причастия?

*Причастия разбираются и лемматизируются как глагол (например, причастие «идущий»). Однако это не всегда так: анализатор допускает погрешность (например, «существующей» - анализатор ошибочно считает леммой словоформу прилагательного «существующий»):*

|  |  |  |
| --- | --- | --- |
| Форма | Разбор | Лемма |
| *существующей* | Afpfsif | *существующий* |
| *идущий* | Vmpp-smafen | *идти* |

1.4. К одной или разным леммам будет отнесены словоформы «*нашедший*» и «*находившего*», «*дал*» и «*давал*»?

*Данным словоформам будут приписаны разные леммы:*

|  |  |  |
| --- | --- | --- |
| Форма | Разбор | Лемма |
| *нашедший* | Vmps-smafpn | *найти* |
| *находившего* | R | *находившего* |
| *дал* | Vmis-sma-p | *дать* |
| *давал* | Vmis-sma-e | *давать* |

*Мы видим, что словоформе «нашедший» присваивается лемма «найти» (что верно); словоформе «находившего» присваивается лемма «находившего», поскольку TreeTagger считает его наречием (на самом же деле должен разбирать его как причастие прошедшего времени в родительном падеже от леммы «находить», следовательно, разбор неверен); словоформе «дал» - лемму «дать» (что верно); словоформе «давал» - лемму «давать» (что так же верно).*

1.5. Напишите правило пересчета тегов системы на теги из ЗС для анафорических местоимений (он, она и т.п.) и наречий.

*Золотой стандарт предусматривает следующий разбор:*

|  |  |  |  |
| --- | --- | --- | --- |
| Wordform\_GS2 | Lemma\_GS2 | POS\_GS2 | Gram\_GS2 |
| *Он* | *он* | SPRO | 3p,m,nom,sg |

*В то же время разбор в Tree Tagger выглядит так:*

|  |  |  |
| --- | --- | --- |
| Форма | Разбор | Лемма |
| *он* | P-3msnn | *он* |

*Правила пересчёта из Tree Tagger в Золотой стандарт для анафорических местоимений:*

|  |  |
| --- | --- |
| Tree Tagger -> Золотой стандарт | Значение |
| P -> PRO | местоимение |
| 1 -> 1р | первого лица |
| 2 -> 2р | второго лица |
| 3 -> 3p | третьего лица |
| m -> m | мужского рода |
| f -> f | женского рода |
| n -> n | среднего рода |
| s -> sg | единственного числа |
| p -> pl | множественного числа |
| n -> nom | именительного падежа |
| g -> gen | родительного падежа |
| d -> dat | дательного падежа |
| a -> acc | винительного падежа |
| l -> loc | местного падежа |
| i -> ins | творительного падежа |
| n -> S (ставится перед тегом PRO) | именное (обладающее свойствами существительного) |

*Правила пересчёта для наречий:*

|  |  |
| --- | --- |
| Tree Tagger -> Золотой стандарт | Значение |
| R -> ADV | наречие |

2. Функциональное тестирование программы TreeTagger.

2.1. Как решаются проблемы токенизации: что происходит с числами, десятичными числами, сокращениями типа г., словами с дефисами, апострофом, знаками препинания? спецзнаками типа $ или &, смешанными элементами (буквы+цифры, вкраплениями другого алфавита) etc.?

*При токенизации знаки препинания отделяются в отдельные токены, слова с дефисами (например, по-разному, Нью-Йорк, экс-министр) остаются неразделенными по дефису, сокращения отделяются в отдельные токены от цифр (если это единицы измерения, например, 10см. -> 10 отдельный токен. см отдельный токен и точка отдельный токен) и последующих знаков препинания, слова с апострофом (для русского это имена собственные) не делятся и остаются цельным токеном, числа со спецзнаками (например, 150°С) остаются неделимыми.*

2.2. Что происходит с незнакомыми словами? Насколько хорошо предсказываются их грамматические характеристики, их леммы?

*У незнакомых слов в целом хорошо угадываются грамматические характеристики, хотя иногда у существительных неправильно угадывается одушевленность, род и то, имя ли собственное это. Также у незнакомых прилагательных иногда неправильно угадывается падеж. В основном леммы угадываются неправильно и в качестве леммы записывается сама словоформа. Примеры: реновации (Ncfsgn реновации), эссенциям (Ncfpdn эссенциям), прокрастинацией (Ncfsin прокрастинацией), котиками (Ncmpin -> Ncmpiy котиками), Людовике (Ncmsln ->Npmsly Людовике), излагалась (Vmis-sfm-p излагалась).*

2.3. Что происходит с омонимичными словоформами: предлагается только один максимально вероятный вариант, предлагаются все возможные варианты, предлагаются все варианты, за исключением очень маловероятных случаев или случаев, снимаемых "надежными" правилами и т.п.

*Для омонимичных словоформ TreeTagger предлагает только они максимально вероятный вариант.*

2.4. Какие проблемные случаи омонимичных разборов разбираются хорошо, в каких часто возникают ошибки и т.п. (например, (а) частеречная омонимия: прилагательное vs. существительное, глагол vs. прилагательное, наречие vs. частица; (б) падежная омонимия; (в) омонимия различных местоименных форм и т.д.)

*В основном хорошо разбираются случаи падежной омонимии: в текстуру* ***волос (Ncmpgn волос)****, из московской* ***мэрии (Ncfsgn мэрия)****, хотя есть случаи и с ошибками: в сеанс* ***видео (Ncnsnn->Ncnsgn видео)****, добавить взбитые* ***белки (Ncfpny->Ncmpan белка)****. Также есть ошибки частеречной омонимии, например, неправильный разбор предлога в течение:* ***в (Sp-a в) течение (Ncnsan течение)*** *15 лет, или разбор наречия как прилагательного: Для приготовления савоярди очень* ***важно (Afpnsns ->R важный)****, но дороговато* ***(Afpnsns ->R дороговатый)*** *для среднестатистического родителя оплачивать.*

3. Обработка файла с помощью морфологического анализатора и сравнение с Золотым стандартом

Код с необходимыми преобразованиями – в файле comparison.py

Уровень оставшейся неоднозначности – 1

Лексическая точность - 0.696

Точность – нет значения (так как всем леммам что-то приписано)

Accuracy - 0.775

(Accuracy только по POS-тегам - 0.915)