Pymorphy2 + UDPipe morphological parsing

Репозиторий проекта

Документация pymorphy:
https://pymorphy2.readthedocs.io/en/latest/user/guide.html

Документация UDPipe:
http://ufal.mff.cuni.cz/udpipe
https://github.com/ufal/udpipe

1. Описание парсеров

Парсер	Метод	Требования	Характеристика системы	Система тегов
pymorphy	Словари, Deterministic Acyclic Finite State Automaton	15 Мб оперативной памяти, Python 2/3	Производит морфологическую разметку, лемматизацию, не дизамбигуирует, не токенизирует	словари и граммемы OpenCorpora (с небольшими изменениями) http://opencorpora.org/dict.php?act=gram
UDPipe	на базе Universal Dependencies.	C++ 2015+ Python 2.6+/3+	Производит токенизацию, лемматизацию, POS-tagging. снятие неоднозначности реализуется алгоритмом MEMM (марковский метод максимальной энтропии)	UPOS https://universaldependencies.org/u/pos/

2. Output
   pymorphy2

UDPipe

3. PYMORPHY2:
    a. 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.?

       Числа и десятичные числа правильно определяются при попадании в формат (без тире вместо дефиса, без лишних знаков препинания внутри)

       Сокращения и аббревиатуры в подавляющем большинстве случаев правильно лемматизируются и тегаются частью речи, ошибки чаще всего в случае падежа, т.к. из сокращений без контекста он не определяется.

       Смешанные с цифрами и латиницей слова лемматизируются верно.

       Все остальные символы, полностью некириллические токены — unknown.

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

       В большинстве глаголов и существительных по характерным морфемам лемматизация верная, чаще всего верны и грамматические характеристики. В случае очень коротких слов не из главных частей речи или случайного совпадения посимвольно с морфемами (трям, чопк) лемматизация неверна, приписывает характеристики других частей речи.

    c. 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.

       Предлагаются все возможные разборы, отсортированные по вероятности (score), включая очень маловероятные.

    d. 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.)

       Частеречная омонимия кроме служебных частей речи разбирается очень хорошо, почти всегда первый разбор верен. Для служебных частей (предлог

vs. сравнительная степень “точнее” и т.п.) больше ошибок, т.к. учета контекста нет.

Падежная омонимия, как и омонимия местоименных форм, разрешается крайне слабо: разборы всегда отсортированы в некотором дефолтном порядке (номинатив, потом аккузатив; мужской род, потом средний), поэтому первый из нескольких возможных вариантов не всегда совпадает с реальным контекстом.

UDPIPE:

e. 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.?

   Г. В годе парсится как год, все ок. Когда в числительном больше двух точечных разделителей, отбрасывает элемент с последней точкой.

   Ссылки проходят как иностранные слова, названия парсит как имена собственные.

   Имена собственные иногда парсят как наречия (Фаберже-Фаберж). Скорее всего связано с алгоритмом лемматизации(по 4 последним символам токена)

   Аббревиатуры иногда парсятся как глаголы (АСУ - Ас) или как предлоги (ЭСУ - Эс)

   Буквы+цифры парсит как одну лемму.

   Элементы со спецзнаками делятся этими спецзнаками на несколько лемм.

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

   Зависит от слова. Лучше смотреть таблицу, будет отмечено цветом.

g. 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.

   Предполагается только одна из возможных словоформ которая выбирается MEMMом

h. 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.)

Ему не нравится суффикс -ет: отстраняет - отстранявать, визуализируется - визуализируться. Омонимия в именах собственных решается в пользу существительных.

4. UDPIPE

Тип обработки	Проблема	Пример	Контекст	Правильный ответ	Ответ системы
Лемматизация	глаг. vs. прилаг.	Данный	нам вчера	Давать / дать	0
	Краткие прилагательные	Морозостоек	-	Морозостойкий	1(морозостойка)
	Именованные сущности	Тянь-Шаня	леса	Тянь-Шань	1
	Междометия	Трям		Трям	1(три)
	Совершенный - несовершенный	Визуализируется	не	Визуализироваться	1(визуализируться)
	Род прилагательных	Внутрибрюшной	стенки	Внутрибрюшной	1(внутрибрюшный)
	Разбивание на предложения			По точкам	
	Именованные сущности	Борзых	премьера	Борзых	1
	Превосходная степень	Величайшей	осторожностью	Большой	0
	Приставочная деривация	Напущенного	На локальном хосте	Напустить	1(напущить)
	Управление	Кружок	Стояли в _	Кружок	1(кружка)

	Причастия	Обособленные	Водные объекты	Обособленный	1(обособлеть)
Токенизация	Слова с девисами	Бело-кремовое	пятнышко	Бело-кремовый	1
	Числительные с запятой	478,8	Долларов США	478,8	0
	Числительные с дефисом	Две-три	недели	Две, три	0
	Пунктуация вокруг слов	“Чпок!	“	Чпок	1
POS-tagging	Причастия	Полученное	количество	ADJ	0

PYMORPHY

Тип обработки	Проблема	Пример	Правильный ответ	Ответ системы
Лемматизация	Неоднозначность	парными	парный	1 Парная, парной
	Омонимия	осени осенить VERB	осень	1
	Синкретизм	людей человек	человек	0
	Латиница	Bullie LATN	bullie	0
	Нестандартный формат	web-технологии web-технология	web-технология	0
	Составные лексемы	две-три две-теретьVERB	Два-три или два, три (зависит от токенизации)	1
	Редкие слова/неологизмы	Порскнулопорскнуть VERB	порскнуть	0
	Нестандартный формат	область/	область	1 (но иногда в аналогичных случаях работает)
	Звукоподражание	Трям трь NOUN трям трямый ADJS	трям	1
	Составные существительные	морякам-подводникам моряк-подводник	моряк-подводник	0
	Имена	тартасов	тартасов	1 тартас

	собственные			
POS-tagging	Причастия/прилагательные	дикорастущих	ADJF	0
	Целые числа	3	NUMB intg	0
	Пунктуация	-	PNCT	0
	Аббревиатуры	сша	NOUN plur Geox	0
	Неоднозначность	точнее	CONJ	0, 1 для варианта COMP
	Действительные числа	478,6	NUMB real	0
	Предикативы	можно	PRED pres	0
	Аббревиатуры	гэс	NOUN femn	0
	Сокращения	г	год NOUN	0
	Составные слова	37-процентный	ADJF	0

5. Для первой 1000 словоформ о п р е д е л и т ь:
    a. уровень оставшейся неоднозначности: число элементов в Output(W) для всех слов тестируемого текста, поделенное на число слов в тексте. Если алгоритм работает однозначно, то этот параметр равняется 1.
    b. лексическая точность алгоритма - число слов текста, для которых лемма приписана правильно, поделенное на общее число обработанных системой слов.
    c. Точность по частям речи - число слов текста, для которых хотя бы одна аннотация правильна (в случае морфологического анализа без дизамбигуации)
    d. Точность по полным грамматическим характеристикам - число слов текста, для которых хотя бы одна аннотация правильна (в случае морфологического анализабез дизамбигуации)
    e. Точность по полным грамматическим характеристикам - число слов текста, для которых аннотация правильна, поделенное на общее число обработанных слов.
    f. Полнота – отношение числа словоупотреблений, которым система приписала правильные теги, к числу словоупотреблений с приписанными тегами в золотом стандарте.

Пункт задания	**Pymorphy2**	**UDPipe**
a	1,8133	1
b	94,4%	93,5%
c	95,5%	98,3%
d	96,5%	92,9%
e	37,96%	92,9%
f*	37,96%	92,9%