Домашнее задание № 1.
Графематический и морфологический анализ текста
Отчет

Задание F:

Провести исследование качества разрешения морфологической омонимии одного из морфоанализаторов для русского языка (см. вариант А), подключив его к своей программе.

Исследование можно провести вручную, взяв нескольких текстов небольшого размера (20-25 предложений), либо автоматически, используя как эталон размеченные тексты. В последнем случае следует вычислить точность разрешения омонимии по леммам, части речи, а также по всем морфологическим характеристикам/тегам (см. материалы MorphoRuEval-2017). Для исследования можно взять одну из следующих пар анализатор – размеченный текст: mystem – тексты НКРЯ (RNC); *pymorphy* – *OpenCorpora*; CrossMorphy – *OpenCorpora*. Поскольку в общем случае морфологические теги уже размеченных текстов могут не совпадать в точности с тегами исследуемого анализатора, необходимо выполнить конвертацию тегов (см., например, тут).

Отчет: Подсчитанные показатели качества разрешения омонимии (в удобной, обозримой форме), описание способа их вычисления, указание использованных размеченных текстов, программа с комментариями (при автоматическом тестировании), выводы по исследованию.

1 Введение

Отчет содержит описание исследования качества разрешения омонимии морфологического анализатора русского языка *pymorphy2*. Исследование проведено в автоматическом режиме, в качестве эталона использованы размеченные тексты корпуса *OpenCorpora* со снятой омонимией. Так как *pymorphy*2 согласован с *OpenCorpora*, перевод тегов не требуется.

2 Процесс исследования

OpenCorpora предлагает корпус из большого количества текстов со снятой вручную омонимией. Кроме того, есть существенно меньший подкорпус, который содержит только те предложения, в которых распознаны все слова (без "UNKN"). Так как *pymorphy* не использует соседние токены при определении тегов, то мы можем использовать бóльший корпус, удалив из него не распознанные слова. Таким образом мы получим максимальное число размеченных токенов для анализа.

Всего корпус со снятой вручную омонимией содержит 93 466 размеченных токена, из них для анализа использовались только распознанные слова. После удаления знаков препинания и нераспознанных слов осталось 67 314 токенов.

Далее все полученные словоформы были проанализированы с помощью pymorphy. Морфологический анализатор принимает словоформу и возвращает все возможные (известные словарю) варианты разбора. Кроме того, на основе статистики анализатор располагает варианты в порядке убывания вероятности (того, что встретится слово с таким вариантом разбора). Вероятность рассчитана на основе наблюдений с *OpenCorpora* (250 000 наблюдений на момент подсчета авторами), либо на основе эмпирических правил, либо считается равномерной. Авторы анализатора утверждают[1] , что выбранный таким образом вариант разбора верен в 79% случаев. Анализ, описанный в этом отчете должен уточнить эти данные.

[1] https://pymorphy2.readthedocs.io/en/latest/user/guide.html

Итак, в рассматриваемом наборе 67314 токенов. В таблице ниже срез из набора, он содержит названия токенов, их тег из *OpenCorpora* (проставлен вручную), тег, присвоенный *pymorphy*, колонка ambig принимает значение True, если *pymorphy* известно несколько вариантов разбора токена (то есть может существовать омонимия). Токенов, для которых существует несколько вариантов разбора, в оставшемся наборе данных оказалось 29 623. Для них мы и будем считать точность разрешения омонимии.

Примечание: в дальнейшем исследовании участвуют только те слова, для которых возможна омонимия. В принципе, возможно, что анализатор допускает другие ошибки при проставлении тегов, и эти ошибки мы не увидим при анализе. Но нас интересует именно разрешение омонимии, и при рассмотрении только 29 623 токенов мы экономим ресурсы.

	tokens	morph_POS	OC_POS	morph	OC	ambig
55	в	PREP	PREP	PREP	[PREP]	True
57	тут	ADVB	ADVB	ADVB,Dmns	[ADVB, Dmns]	True
58	же	PRCL	PRCL	PRCL	[PRCL]	True
59	появившихся	PRTF	PRTF	PRTF,perf,intr,past,actv plur,gent	[PRTF, perf, intr, past, actv, plur, gent]	True
60	рекламных	ADJF	ADJF	ADJF,Qual plur,gent	[ADJF, Qual, plur, gent]	True
63	отсутствовавших	PRTF	PRTF	PRTF,impf,intr,past,actv plur,gent	[PRTF, impf, intr, past, actv, plur, gent]	True
64	на	PREP	PREP	PREP	[PREP]	True
66	Культуре	NOUN	NOUN	NOUN,inan,femn sing,datv	[NOUN, inan, femn, sing, loct]	True
71	с	PREP	PREP	PREP	[PREP]	True
72	одной	ADJF	ADJF	ADJF,Apro femn,sing,gent	[ADJF, Apro, Anum, femn, sing, gent]	True

Таблица 1: Срез из набора данных, в котором оставлены только те слова, для которых возможна омонимия (колонка ambig == True)

Далее, в этом наборе наблюдений будем находить неверно расставленные морфологическим анализатором *pymorphy* теги. Существует несколько видов ошибок при распозновании омонимии.

Первый вид ошибки: неверное определение части речи. Это может быть частеричная омонимия (если леммы совпадают для двух разных частей речи) или лексико-морфологическая (если леммы не совпадают для двух частей речи, но слова совпадают в какой-либо другой форме кроме начальной).

Второй вид ошибки: верно определена часть речи, но неверно определен падеж, число, ...Это может быть ошибка в разрешении морфологической омонимии (совпадают словоформы для разных форм одной лексемы). Либо, если для данной словоформы существует лексико-морфологическая омонимия, это может быть верным определением части речи, но неверным определением других параметров тэга.

Статистика считалась отдельно по случаям и отдельно для частей речи. Например, в случае, если часть речи разбираемой словоформы — существительное, далее проверяются теги падежа, числа, рода, одушевленности. Если какой-то из этих тегов не совпадает, то этой словоформе присваивается соответствующий маркер. В случае, если совпадают эти характеристики, то считается, что омонимия разрешена верно. На самом деле, в ручном разборе *OpenCorpora* встречаются некоторые другие теги: разговорное слово, местоименное прилагательное и т.д. Отсутствие этих тегов в разборе *pymorphy* не считалось ошибкой. В то же время, иногда *pymorphy* проставляет теги, которых нет в разборе OC. Например, для слова 'пятерых' *pymorphy* определил одушевленность (действительно, мы используем подобные собирательные числительные только для одушевленных), а в разборе OC этот тег отсутствует. Это усложняло процесс подсчета статистик.

Подобным образом рассмотрены прилагательные, глаголы, причастия (+вид сов/несов, переходность, время), числительные (только падеж) и остальные встречавшиеся теги. Для

Morphol_homonimy_Noun_Case	4086	55.58
Morphol_homonimy_ADJF_Case	1270	17.28
POS_homonymy	1024	13.93
Morphol_homonimy_ADJF_Gender	294	4.00
Morphol_homonimy_Noun_Gender	226	3.07
Morphol_homonimy_Noun_Animacy	180	2.45
Morphol_homonimy_NPRO_Case	108	1.47
Morphol_homonimy_PRTF_Gender	71	0.97
Morphol_homonimy_PRTF_Case	44	0.60
ROMN_homonimy	15	0.20
Morphol_homonimy_Noun_Number	14	0.19
Morphol_homonimy_NUMR_Case	13	0.18
LATN_homonimy	3	0.04

наречий и числительных дополнительные параметры не рассматривались.

Далее, по всем тегам подсчитаем, как часто встречается те или иные ошибки при разборе омонимии в рассматриваемом наборе данных. Напомним, что в рассматриваемом наборе 29 623 словоформы с возможной омонимией.

В 'UNKNOWN' попали слова, которые почему-то не получилось сравнить автоматически. Их всего 6, и то, что там нет ошибок было проверено вручную.

Слов, в которых теги были расставлены верно, 22 045, это 74.4% слов, в которых омонимия была возможна. Таким образом, верно распознаны 89% всех слов. Это выше, чем подсчитано разработчиками в 2013 году, но они, возможно, учитывали другие ошибки.

Среди всех ошибок при распознавании слов с омонимией 56% приходится на ошибки при определении падежа существительных. Самая частая ошибка среди них — в различии винительного и именительного падежей. 17.3% ошибок совершено при распознавании падежа прилагательных. Это значительно меньше, но, скорее всего, связано с тем, что прилагательных было существенно меньше в нашем наборе данных.

При распознавании части речи совершено 13.9% ошибок. Это в основном частеричная омонимия в которой одна лемма может быть и существительным и прилагательным (например, 'согласные', 'прохожие'). Но встречается и лексико-морфологическая омонимия.

Интересно, что встречается сравнительно много ошибок в определении рода существительных. Например, *pymorphy* определяет слово 'пол' женским родом (от 'половина'), а слово 'арене' — мужским (от слова 'арен').

Интересно еще проследить разбор отдельно стоящих заглавных латинских букв. В некоторых случаях символ определяется как латинская буква, а в некоторых — как римская цифра. В общей сложности выявлено 18 ошибок в разборе таких примеров. Они отражены в таблице с маркерами 'ROMN_homonimy' и 'LATN_homonimy'.

Далее посчитана статистика по словам. В таблице ниже представлены слова, в которых чаще всего была неверно разобрана неоднозначность, и число таких ошибок. Это одни из самых часто встречающихся в тексте слов. В основном неверно определялся поадеж существительных, часть речи для местоимений и частиц.

России	52
тоже	45
этом	44
США	42
человек	32
мира	29
Ссылки	26
этой	25
века	25

3 Выводы по исследованию

Морфологический анализатор *pymorphy2* хорошо расставляет теги для русских слов. Не смотря на то, что он использует словарь и статистический метод для расставления меток и не учитывает контекст, верный тег расставляется в 89% случаев на исследуемом корпусе.

Существует довольно мало размеченных вручную текстов на русском языке, на которых можно было бы обучать и проверять морфологические анализаторы. При расширении базы возможных употреблений слов можно улучшить работы анализатора, даже в том случае, если он не учитывает конекст. Кроме того, так как OpenCorpora самый надежный и популярный ресурс, предоставляющий вручную размеченный корпус, то анализатор построен и обучен (статистически) на этом корпусе. Проверять работу анализатора стоило бы на другом корпусе. Наверняка тогда процент правильных тегов был бы ниже, но более верно отражал бы количество ошибок на новых текстах.

Комментарии

Исследование выполнено на языке Python 3.6

Используемые библиотеки:

- стандартные
    1. pandas
    2. numpy
- специализированные (для анализа текстов)
    1. pymorphy2 (источник)
    2. opencorpora (источник)