Лицей «Физико-техническая школа»
Санкт-Петербургского Академического университета

Курсовая работа (отчет о практике)

Опыт автоматизированной классификации контекстов при морфологической разметке корпуса

Работу выполнила
Ученица: Голуб Анна (11 А класс)
Научный руководитель:
Грановский Дмитрий Владимирович
Место прохождения практики:
проект «Открытый корпус» (OpenCorpora)

Санкт-Петербург
2018

Аннотация

В рамках проекта OpenCorpora производится лингвистическая разметка корпуса русского языка с помощью краудсорсинга: пользователи выполняют задания на сайте OpenCorpora. Единичное задание имеет вид: для данного слова в данном контексте выбрать тот грамматический признак, которым оно обладает. Единообразие разметки обеспечивается, в частности, при помощи проверки ответов модераторами. Одной из важных задач является фильтрация поступающих ответов для того, чтобы на модерацию не попадали те из них, которые с очень низкой вероятностью окажутся неправильными. Одним из способов фильтрации является проверка, стоит ли данное слово в некотором заданном ошибкоопасном контексте. Соответственно, задачей является выявление критериев такого контекста.

Оглавление

Введение

1. **Что такое корпус?**

Корпус — это база текстов. Примерами корпуса являются:

- собрание текстов на данном языке,
- собрание текстов на одном языке и их переводов на другой язык,
- собрание статей из определенной научной области.

Примерами использования корпуса являются:

- электронный словарь,
- машинный перевод,
- автоматическая проверка орфографии и пунктуации,
- тестирование программ разметки.

Корпус текстов *OpenCorpora* [1] используется преимущественно для тестирования программ разметки.

2. **Что такое разметка?**

Разметка — это информация о лингвистических характеристиках различных составляющих текста.

Разметка может быть:

- фонетической (например, место образования для согласного звука)
- морфологической (например, род для существительного)
- синтаксической (например, простое/сложное для предложения)

В *OpenCorpora* осуществляется морфологическая разметка, размечаемой единицей является **токен**. (Здесь и далее имеется в виду именно морфологическая разметка.) Разметка токена состоит из одной или нескольких (в случае омонимии) интерпретаций [2]. Каждая интерпретация обязательно содержит указание на класс токена (словарный, несловарный). **Словарь** корпуса [3] – файл, где собраны все размечаемые слова корпуса и их грамматические характеристики. Словарными называются токены, содержащиеся в словаре; несловарными — те, которые помещать туда не имеет смысла ввиду, например, их количества (интернет-адреса, химические формулы и т. п.) (Токены, которые могут быть размечены, но пока не внесены в словарь, не относятся ни к одному из классов.) Для словарных токенов интерпретация также включает:

- идентификатор леммы из словаря,
- часть речи,
- набор значений обязательных для данной части речи грамматических

категорий (например, число для имен существительных),

- набор меток, обозначающих особенности конкретного употребления словоформы в тексте (например, «опечатка», «безличное употребление глагола»).

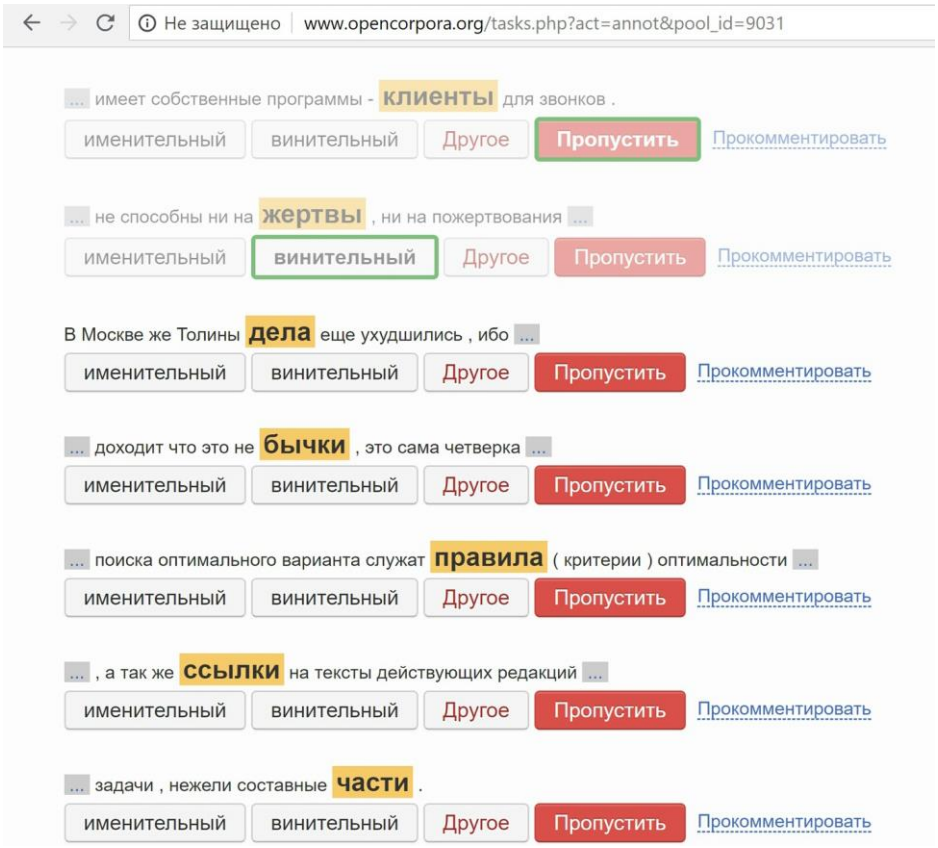
3. **Создание разметки.**

В создании разметки выделяется два этапа: автоматическая разметка и разметка с помощью краудсорсинга.

Автоматическая разметка — разметка с помощью специально для этого созданной компьютерной программы, сопоставляющей каждому слову в корпусе множество его интерпретаций из словаря. У большого числа слов существует несколько интерпретаций, поэтому после того, как корпус разметила программа, для них нужно **снять омонимию**, т. е. для каждого слова в контексте выбрать из множества интерпретаций единственно верную. Снятие омонимии осуществляется с помощью краудсорсинга.

Краудсорсинг (от англ. c*rowd* - *«*толпа» и *sourcing* - «использование ресурсов»*)* - привлечение к решению тех или иных проблем инновационной производственной деятельности широкого круга лиц для использования их творческих способностей, знаний и опыта по типу субподрядной работы на добровольных началах с применением инфокоммуникационных технологий [4]. В рамках проекта *OpenCorpora* пользователи выполняют задания на сайте *OpenCorpora*. Единичное задание имеет вид: для данного слова в данном контексте выбрать тот грамматический признак, которым оно обладает (например, "именительный падеж" vs "винительный падеж").

Так выглядит интерфейс с заданиями на сайте *OpenCorpora* (в данном случае предлагается определить, в именительном или в винительном падеже стоит выделенное существительное).



Для единообразия разметки каждую языковую единицу размечают 3-4 пользователя. В случае несовпадения ответов хотя бы двоих из них задание отправляется на рассмотрение к модератору (специалисту с лингвистическим образованием), который окончательно решает вопрос о разметке данной сущности. (Ответ модератора считается верным.) Безусловно, роль модератора в процессе создания разметки очень важна, но использование краудсорсинга позволяет отсеивать случаи однозначной разметки и размечать только сущности с наиболее трудно определимыми грамматическими характеристиками.

Постановка задачи

Важной задачей является облегчение работы модератора путем фильтрации части попавших к нему на рассмотрение заданий. Одним из способов решения данной задачи является выявление критериев контекста, в котором стоит размечаемое слово, при которых ответ пользователя с приемлемой вероятностью окажется неправильным.
Предлагается решить эту задачу для заданий вида *"именительный падеж"* vs *"винительный падеж"* для существительных. В качестве статистических данных возможно использование информации о соответствующих заданиях из корпуса текстов *OpenCorpora.*

Методика

1. Проанализировать массив данных о проверенных ответах данного типа заданий. Проверить, существуют ли случаи, когда все ответы, данные на одно и то же задание разными людьми, неверные (ответ модератора дан и считается верным), оценить частоту этих случаев. (Нужна автоматизация, размер данных порядка десятков тысяч рядов.)
2. Если такие случаи достаточно частотны (пороговая величина будет дана), разработать набор формальных критериев, по которым можно с приемлемой точностью автоматически выделять такие случаи. Предполагаемый вид критерия - "перед/после целевого слова в пределах K слов находится слово с такими-то грамматическими характеристиками или токен такого-то вида (например, число)".

Для написания кода программ использовать язык программирования Python [5].

Результат

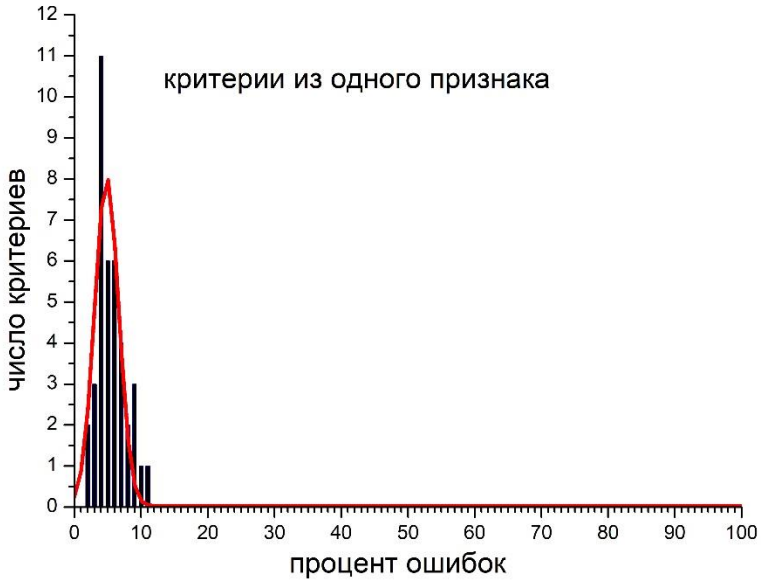
1. Написан код для подсчета числа таких случаев, когда все ответы, данные разными людьми на одно и то же задание, неверные. (Здесь и далее имеются в виду задания вида *"именительный падеж"* vs *"винительный падеж"* для существительных.) Таких случаев оказалось около 0.2% на выборке размером 35199 примеров, что превышает данную пороговую величину (0.1%). Соответственно, имеет смысл оптимизация разметки заданий данного типа, т. к. случаи, когда все отвечавшие дали неверный ответ, достаточно частотны.
2. Разработан набор признаков контекста, в котором стоит целевое слово. Признак может представлять собой:
   - характеристику целевого слова (например, является ли оно аббревиатурой),
   - характеристику контекста (например, было ли при разметке показано всё предложение),
   - характеристику предыдущего слова (например, является ли оно глаголом).

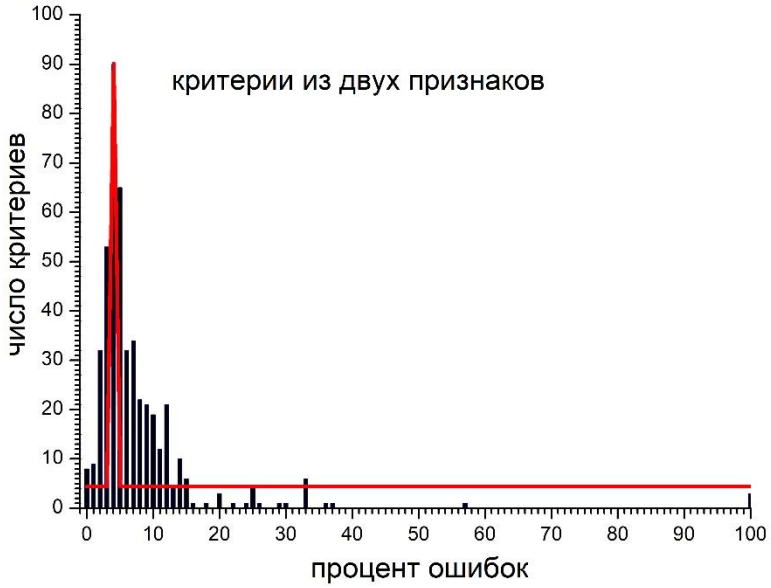
   Все признаки представлены в таблице 1.

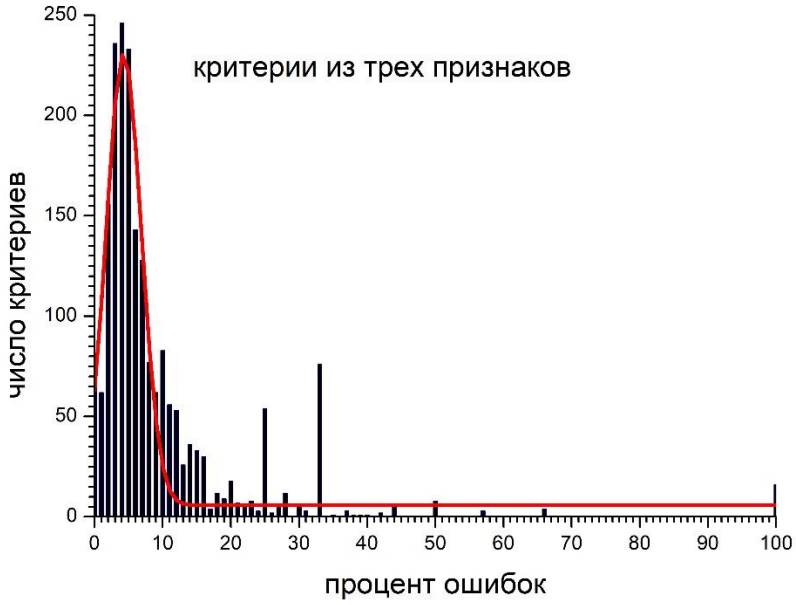
Таблица 1

Признак	Признак: предыдущий токен - …
при разметке было показано всё предложение	число
слово первое в предложении	название географического объекта
слово последнее в предложении	отчество
слово написано с большой буквы	аббревиатура
в слове есть дефис	имя
слово последнее в клаузе	несклоняемое существительное
слово - название географического объекта	существительное
слово - аббревиатура	глагол
слово - имя	глагольный инфинитив
слово - несклоняемое существительное	количественное/собирательное числительное
слово - в словосочетании, стоящем в кавычках	полное прилагательное
следующее слово - глагол	краткое прилагательное
одно из следующих 3 слов - глагол	предлог
одно из следующих 5 слов - глагол	наречие
одно из следующих 10 слов - глагол	частица
	местоимение
	категория состояния
	союз
	прилагательное в сравнительной степени
	полное причастие
	краткое причастие
	междометие
	деепричастие
	собирательное числительное

3. Создан файл со строками, соответствующими каждому из заданий, вида: предложение с выделенным целевым словом (оно стоит в квадратных скобках); далее число правильных и неправильных ответов на данное задание, далее бинарные коды каждого из признаков (1 – выполняется, 0 – не выполняется).
   Пример:
   Надо признать , обращение к современности выбило из рук писателя самый главный его козырь – [[знание]] скрупулезно описываемых реалий прошлого . 2 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
   (На данное задание было дано два верных и один неверный ответ. При разметке было показано все предложение; слово, стоящее перед целевым, является существительным.)
   Код, проверяющий выполнение признаков, использует формальные признаки токенов в данном предложении (например, при определении, является ли целевое слово первым в предложении), а также словарь корпуса. Ввиду большого размера словаря (порядка 10^6 строк) написан код, перезаписывающий каждое слово из словаря в файл, соответствующий его части речи.
4. Разработаны критерии контекста, каждый из которых представляет собой сочетание одного, двух или трех признаков. Написан код для подсчета процентных соотношений ошибочных ответов пользователей к общему числу ответов для каждого из критериев. Построены графики зависимости частоты выпадения каждого процента от его значения, которые с приемлемой точностью представляют собой графики нормального распределения. Отклонения вызваны случаями корреляции признаков, составляющих критерий (например, трех признаков, определяющих расстояние до ближайшего глагола), а также случаями недостаточно большой выборки примеров с данным критерием контекста.
   (Графики представлены ниже.)


критерии из одного признака
число критериев
процент ошибок
0
1
2
3
4
5
6
7
8
9
10
11
12
0
10
20
30
40
50
60
70
80
90
100


критерии из двух признаков
число критериев
процент ошибок
0
10
20
30
40
50
60
70
80
90
100
0
10
20
30
40
50
60
70
80
90
100


критерии из трех признаков
число критериев
процент ошибок
0
50
100
150
200
250
0
10
20
30
40
50
60
70
80
90
100

5. При анализе полученных статистических данных выявлено, что наиболее ошиобкоопасными контекстами являются контексты, где **целевое слово – аббревиатура** или **предыдущий токен – аббревиатура** или **число**. В таблице 2 представлены критерии, состоящие из трех признаков, для которых соответствующий процент ошибки превышает 25%. (В таблице представлены только критерии с достаточной (более 10) выборкой примеров, не являющиеся сочетаниями коррелирующих признаков.)

Таблица 2

Критерий	Процент ошибки
Целевое слово - аббревиатура, следующее слово – аббревиатура, следующий глагол в числе следующих 10 слов	25.5
Целевое слово написано с большой буквы, предыдущий токен – число, ближайший глагол на расстоянии 5 слов	27.2
Целевое слово написано с большой буквы, предыдущий токен – число, ближайший глагол на расстоянии 10 слов	33.3
Целевое слово написано с большой буквы и является аббревиатурой, предыдущий токен - число	37.3
Целевое слово - аббревиатура, предыдущий токен – число, ближайший глагол на расстоянии 10 слов	38.9

Выводы

Установлено, что существуют такие критерии контекста для заданий вида *"именительный падеж"* vs *"винительный падеж"* для существительных, что процент ошибок при выполнении соответствующего задания достаточно высок. Выявлены эти критерии. В дальнейшем эта информация может быть использована для автоматизации морфологической разметки корпуса текстов *OpenCorpora*.

Благодарности

Я благодарна моему научному руководителю Д. В. Грановскому за постановку задачи и обсуждение проблем, возникавших по ходу ее решения, а также за помощь в поиске необходимых источников информации.

Список литературы

1. Проект «Открытый корпус» (OpenCorpora). http://www.opencorpora.org
2. Грановский Д.В., Бочаров В.В., Бичинева С.В. Открытый корпус: принципы работы и перспективы // Компьютерная лингвистика и развитие семантического поиска в Интернете: Труды научного семинара XIII Всероссийской объединенной конференции «Интернет и современное общество». Санкт-Петербург, 19–22 октября 2010 г. / Под ред. В.Ш. Рубашкина. — СПб., 2010. — 94 с.
3. Словарь корпуса *OpenCorpora* (версия для скачивания plain text) http://www.opencorpora.org/files/export/dict/dict.opcorpora.txt.zip
4. Википедия — свободная энциклопедия http://ru.wikipedia.org
5. Python 3.7.2rc1 documentation. https://docs.python.org/3