Олимпиада НИУ ВШЭ для студентов и выпускников – 2019 г.

Демонстрационный вариант и методические рекомендации по направлению: «Фундаментальная и прикладная лингвистика»

Профиль: «Компьютерная лингвистика и цифровые методы в гуманитарных науках» **КОД - 310**

Время выполнения задания – 180 мин., язык – русский.

I. ОБЩАЯ ЧАСТЬ

Вопрос 1.

Задание

Прочтите пост из блога, посвященного автоматической обработке естественного языка(http://nlpers.blogspot.com/). Составьте краткое резюме этого поста на русском языке (объемом в один абзац, но не более 1000 знаков), отразив в нем основной тезис автора поста.

NLP as a study of representations

Ellen Riloff and I run an NLP reading group pretty much every semester. Last semester we covered "old school NLP." We independently came up with lists of what we consider some of the most important ideas (idea = paper) from pre-1990 (most are much earlier) and let students select which to present. There was a lot of overlap between Ellen's list and mine (not surprisingly). . The whole list of topics is posted as a comment. The topics that were actually selected are here.

I hope the students have found this exercise useful. It gets you thinking about language in a way that papers from the 2000s typically do not. It brings up a bunch of issues that we no longer think about frequently. Like language. (Joking.) (Sort of.)

One thing that's really stuck out for me is how much "old school" NLP comes across essentially as a study of representations. Perhaps this is a result of the fact that AI -- as a field -- was (and, to some degree, still is) enamored with knowledge representation problems. To be more concrete, let's look at a few examples. It's already been a while since I read these last (I had meant to write this post during the spring when things were fresh in my head), so please forgive me if I goof a few things up.

I'll start with one I know well: Mann and Thompson's rhetorical structure theory paper from 1988. This is basically "the" RST paper. I think that when a many people think of RST, they think of it as a list of ways that sentences can be organized into hierarchies. Eg., this sentence provides background for that one, and together they argue in favor of yet a third. But this isn't really where RST begins. It begins by trying to understand the communicative role of text structure. That is, when I write, I am trying to communicate something. Everything that I write (if I'm writing "well") is toward that end. For instance, in this post, I'm trying to communicate that old school NLP views representation as the heart of the issue. This current paragraph is supporting that claim by providing a concrete example, which I am using to try to convince you of my claim.

As a more detailed example, take the "Evidence" relation from RST. M+T have the following characterization of "Evidence." Herein, "N" is the nucleus of the relation, "S" is thesatellite (think of these as sentences), "R" is the reader and "W" is the writer: relation name: Evidence constraints on N: R might not believe N to a degree satisfactory to W constraints on S: R believes S or will find it credible constraints on N+S: R's comprehending S increases R's belief of N the effect: R's belief of N is increased locus of effect: N

This is a totally different way from thinking about things than I think we see nowadays. I kind of liken it to how I tell students not to program. If you're implementing something moderately complex (say, forward/backward algorithm), first write down all the math, then start

implementing. Don't start implementing first. I think nowadays (and sure, I'm guilty!) we see a lot of implementing without the math. Or rather, with plenty of math, but without a representational model of what it is that we're studying. The central claim of the RST paper is that one can think of texts as being organized into elementary discourse units, and these are connected into a tree structure by relations like the one above. (Or at least this is my reading of it.) That is, they have laid out a representation of text and claimed that this is how texts get put together. As a second example (this will be sorter), take Wendy Lehnert's 1982 paper, "Plot units and narrative summarization." Here, the story is about how stories get put together. The most interesting thing about the plot units model to me is that it breaks from how one might naturally think about stories. That is, I would naively think of a story as a series of events. The claim that Lehnert makes is that this is not the right way to think about it. Rather, we should think about 3 stories as sequences of affect states. Effectively, an affect state is how a character is feeling at any time. (This isn't quite right, but it's close enough.) For example, Lehnert presents the following story: When John tried to start his care this morning, it wouldn't turn over. He asked his neighbor Paul for help. Paul did something to the carburetor and got it going. John thanked Paul and drove to work.

The representation put forward for this story is something like: (1) negative-for-John (the car won't start), which leads to (2) motivation-for-John (to get it started, which leads to (3) positivefor-John (it's started), when then links back and resolves (1). You can also analyze the story from Paul's perspective, and then add links that go between the two characters showing how things interact. The rest of the paper describes how these relations work, and how they can be put together into more complex event sequences (such as "promised request bungled"). Again, a high level representation of how stories work from the perspective of the characters.

So now I, W, hope that you, R, have an increased belief in the title of the post. Why do I think this is interesting? Because at this point, we know a lot about how to deal with structure in language. From a machine learning perspective, if you give me a structure and some data (and some features!), I will learn something. It can even be unsupervised if it makes you feel better. So in a sense, I think we're getting to a point where we can go back, look at some really hard problems, use the deep linguistic insights from two decades (or more) ago, and start taking a crack at things that are really deep. Of course, features are a big problem; as a very wise man once said to me: "Language is hard. The fact that statistical association mining at the word level made it appear easy for the past decade doesn't alter the basic truth. :-)." We've got many of the ingredients to start making progress, but it's not going to be easy!

II. СПЕЦИАЛЬНАЯ ЧАСТЬ

Выберите и выполните <u>только один</u> из блоков заданий специальной части в соответствии с выбранной вами программой магистерской подготовки.

Вопрос 2. (Вариант для направления “Компьютерная лингвистика”)
Решите задачу.

В алфавите языка племени УЫУ всего две буквы: У и Ы, причем этот язык обладает такими свойствами: если из слова выкинуть стоящие рядом буквы УЫ, то смысл слова не изменится. Точно так же смысл слова не изменится при добавлении в любое место слова буквосочетания ЫУ или УУЫЫ. Можно ли утверждать, что слова УЫЫ и ЫУУ имеют одинаковый смысл?

Вопрос 2. (Вариант для направления “Цифровые методы в гуманитарных науках”)
Задание.

Вы с командой единомышленников решили сделать глубоко размеченный корпус русских романов для количественных исследований этого жанра (в синхронии и диахронии).

Напишите подробный план такого исследования. В план нужно включить информацию о том:
—как будет собираться корпус, какие данные и откуда вы планируете получить,
—какие компьютерные инструменты придется найти или разработать для сбора этого корпуса
—какую разметку вы сделаете в этом корпусе? проявите максимум фантазии и предложите как можно больше уровней разметки
—какие компьютерные инструменты, технологии придется найти или разработать для осуществления разметки. С чем возникнут сложности?
—какие количественные исследования (синхронные и диахронические) могут быть произведены на основе вашего корпуса. Что нового они расскажут о русском романе?
—можете ли вы предположить, в каких направлениях возможен переход от количественного анализа к интерпретации.
Возможны ли кросс-культурные исследования с применением вашего корпуса? Если да — какие?

Вопрос 3 (Вариант для направления “Компьютерная лингвистика”)
Задание
Перед вами он-лайн система “поздравлятор”. Она сочиняет поэтические поздравления по запросу пользователя. Для того чтобы система выдала оригинальный стихотворный текст, пользователь должен ввести определенную информацию: имя, пол, возраст, способ обращения (на ты или на вы) к имениннику, метрические характеристики. В результате работы системы пользователь получает осмысленный, грамматически правильный, ритмически организованный и рифмованный текст, содержащий в себе поздравление с днем рождения. Несмотря на то, что каждый раз система выдает новые стихи, все предыдущие накапливаются в банк данных, и их можно посмотреть.Ваша задача как компьютерного лингвиста – разработать методологию тестирования качества работы лингвистических модулей системы. В описании вашей методологии должны быть отражены ответы на следующие вопросы:
1) Какие именно функции, связанные с обработкой и генерацией текстов на
естественном языке, важны для предлагаемого сервиса и почему? Какие из них абсолютно необходимы, а без каких можно обойтись?
2) Каким образом качество работы этих функций может быть протестировано? Что должно быть предусмотрено в системе, для того чтобы была обеспечена возможность такого тестирования?
3) Какой могла бы быть система рейтингов (штрафов, баллов и т.п.) для разных лингвистических функций? Как получить и интерпретировать результирующую
оценку качества лингвистической системы в целом?

Вопрос 3. (Вариант для направления “Цифровые методы в гуманитарных науках”)
Задание.
Современные цифровые гуманитарные исследователи постоянно заимствуют новые технологии и применяют их для своих целей. Так в Digital Humanities пришли корпусные/квантитативные методы анализа текстов, сетевой анализ, геоинформационные системы, машинное обучение, компьютерное зрение. Выберите любую современную технологию (из перечисленных — или иную) и опишите как можно более подробно несколько примеров исследований, которые можно осуществить с ее помощью. Какие новые научные результаты может дать применение этой технологии?

Примеры правильных ответов

Задание 1.

Автор блога излагает некоторый новый взгляд на обработку естественного языка (naturallanguage processing). Он заключается в обращении к идеям, которые царили в этой области несколько десятилетий назад. И основная из них предлагает рассматривать NLP как изучение языка сквозь призму его «представлений», схем. В качестве примеров таких схем автор приводит теорию риторических структур (TPC) У.Манна и С.Томпсон и работу о структуре рассказа В.Ленхарт. Теория риторических структур, стремясь раскрыть коммуникативную роль организации дискурса, представляет текст как иерархически организованный список 8 элементарных дискурсивных единиц. А в работе Венди Ленхарт предлагается новый взгляд на устройство рассказа: в ее интерпретации оно представляет собой не череду событий, а некоторую последовательность эмоциональных состояний героев. Ее теория подкрепляется разработкой схематического представления стандартной структуры рассказа. Автор статьи предлагает применить такой подход и в новой сфере – в области автоматической обработки естественного языка (NLP). По его мнению, использование этого опыта предыдущих исследований позволит достичь определенного прогресса в данной области.

Задание 2.

При любой разрешенной нам операции добавления или выкидывания куска слова количества букв У и Ы в этом куске равны. Это означает, что разность между числом букв У и букв Ы в слове не изменяется. Это можно проследить на примере

Ы -> ЫЫУ -> ЫУУЫЫЫУ -> ЫУЫЫУ

Во всех этих словах букв Ы на одну больше, чем букв У. Вернемся к решению. В слове УЫЫ разность равна (-1), а в слове ЫУУ равна 1. Значит, из слова УЫЫ нельзя разрешенными операциями получить слово ЫУУ, и следовательно, нельзя утверждать, что эти слова обязательно имеют одинаковый смысл.

Задание 3.

1. Подобная система могла бы быть устроена следующим образом.

Её основой могли бы стать шаблоны – готовые короткие стихи с пропусками, в которые будут вставляться данные, полученные от пользователя (например, обращение к имениннику). Нужны заготовки шаблонов на разные метрические характеристики (чем больше шаблонов будет подготовлено для каждого типа метрики, тем разнообразнее будут получаться поздравления, но для минимального варианта системы должно быть достаточно и одного шаблона на метрику). Чтобы формы для стихов могли «приспосабливаться» под данные, вводимые пользователем, некоторые позиции в шаблонах должны быть «подвижными», т.е. должны подразумевать возможность замены одних элементов на другие. Например, в маленьком фрагменте заготовки «Дорогая_____, Нету тебя краше !» должна быть учтена возможность замены первого прилагательного на другое, в зависимости от имени, которое вводит пользователь, ср.:

(1) Дорогая Маша, Нету тебя краше!

(2) Милая Наташа, Нету тебя краше! (

3) Дорогой наш Саша, Нету тебя краше!

Для того, чтобы подобные замены были возможны, необходимо создать словарь с особой структурой. В нём должны храниться слова-замены двух типов: разного рода эпитеты, т.е. прилагательные, и короткие слова, которые могут увеличивать количество слогов в строке, существенно не меняя смысла, вроде усилительных частиц типа уж, местоимений типа ты, наш, ср. пример выше или вариант «Дорогой ты наш Владимир». Возможно, осмысленным будет и включение некоторых глаголов, взаимозаменяемых в стихотворных поздравлениях (ср. обожаем/уважаем). В словаре же должны быть указаны релевантные

для нас характеристики слов-замен: в первую очередь, количество слогов и место ударения. Могут также храниться отдельно разные формы одного и того же прилагательного, чтобы формы единственного или множественного числа, мужского или женского рода были заданы изначально в словаре – в таком случае можно будет не использовать в нашей системе морфологический модуль. Если же всё-таки включать в систему фрагменты морфологического парсинга и грамматических правил построения словоформ, то придётся добавить ещё ряд модулей: так, нужно будет автоматически определять количество слогов в образованных словоформах (что, в принципе, довольно просто) и место ударения (что сложнее, учитывая подвижность ударения в русском языке). Вероятно, в данном случае рациональнее хранить информацию в готовом виде в словаре, тем более, что для шаблонов будет нужно всего две-три формы одной лексемы, а не вся парадигма. Помимо словаря слов-замен, нужно создать словарь имён. В простейшем случае он будет выглядеть как тройка из имени, количества слогов и места ударения (количество слогов можно не включать в словарь, а считать автоматически). В принципе, этого будет достаточно для выбора правильных слов-замен: указание на пол позволит выбрать правильную форму прилагательного, а введённого пользователем имени будет достаточно для обращения (изменять форму имени собственного необходимости не будет). Информацию о поле человека, который может иметь данное имя, допустимо прикреплять к слову сразу в словаре. Хранящиеся в системе шаблоны должны быть расклассифицированы определенным образом: а) по возрасту, на который они рассчитаны (в зависимости от этого в константных, неизменяемых частях стихотворения будут содержаться разные пожелания – конфет и мороженого ребенку, долгих лет жизни взрослому человеку) б) по полу (опять-таки, разные пожелания для женщин и мужчин) в) по способу обращения. г) по типу метрической системы (как уже было указано выше) Таким образом, данные, введенные пользователем, учитываются на двух этапах порождения текста: сначала при отборе шаблона, а потом при его «подгонке» под имя путем манипулирования словами-заменами. Такая организация системы (генерация только верхней строки) предполагает наличие достаточно большого количества шаблонов и весьма ограниченный простор для автоматического порождения текста. Зато она просто устроена, что хорошо, как минимум, в двух аспектах: а) система быстрее работает б) меньше вероятность ошибки Можно организовать работу системы несколько иным способом, сделав её более автоматизированной. В шаблоне можно оставлять больше свободных позиций (который, например, будут заполняться прилагательными, случайно выбираемыми из некоторого закрытого множества лексем). Можно оставить пробел для выбора номинации, например, девочка/девушка/женщина (ср. Ты лучшая девочка/девушка/женщина в мире, причём выбор будет осуществляться согласно информации, указанной в графе возраст). Можно автоматизировать выбор правильной формы глагола (ср.: Чтоб Вы были счастливы vs. Чтоб ты была счастлива vs. Чтобы ты был счастлив; радуй vs. радуйте). При этом, опять же, в морфологическом анализаторе необходимости нет: в данном случае можно задать условия выбора правильной формы с помощью простых правил, например: оставить в строке свободные места для подстановок (чтоб_ __ был_ счастлив_); в зависимости от того, какие данные ввёл пользователь, заполнять пробелы (вариантов в каждом случае будет не более трёх). Можно даже подключить к работе системы семантический компонент: он будет связан, например, с выбором прилагательного, соответствующего полу и возрасту поздравляемого. Так, например, к женщине применим эпитет нежный, а к мужчине – скорее, нет. Эту информацию можно также указать в словаре, не надстраивая дополнительного семантического модуля. В синтаксическом модуле система нуждаться не будет: работу синтаксического анализатора выполняют заранее сформированные шаблоны (ср. принцип работы Грамматики Конструкций, где для каждой переменной конструкции есть конкретные требования по её заполнению).

2. Чтобы протестировать систему, достаточно задать некоторое количество запросов (задавая различные комбинации параметров) и оценить генерируемые ею стихотворения. Для этого необходимо, чтобы система позволяла делать большое количество запросов за короткий промежуток времени. В качестве корпуса порождённых стихотворений для анализа качества работы системы можно использовать банк данных, в котором накапливаются все созданные системой стихи. 3. Качество работы системы можно оценить, например, следующим образом. Можно задать ряд параметров (в данном случае такими параметрами-критериями будут грамматическая (морфологическая) правильность, соответствие выбранной метрике и, возможно, семантическая адекватность) и оценивать каждый параметр по пятибалльной шкале. Каждому параметру можно присвоить свой вес (в зависимости от того, какой компонент мы считаем более существенным). И в результате мы получим число, посчитанное по следующей формуле: $\Sigma Ok*nk$, где Ok – оценка по данному критерию, а nk – вес данного критерия. Это число можно считать оценкой работы системы.

Методические рекомендации

Олимпиада по направлениям «Компьютерная лингвистика» и «Цифровые методы в гуманитарных науках» проводится с целью конкурсного отбора на одноименные магистерские программы. Основная цель программы – подготовка лингвистов, владеющих теоретическими знаниями о структуре и функционировании естественного языка, владеющих методами и решения задач с помощью технологий автоматической обработки языка, способных участвовать в самых современных проектах, связанных с языковыми технологиями, а также самим их создавать, формулировать новые задачи и предлагать алгоритмы их решения.

Олимпиада предполагает три задания: логическую задачу, творческое задание в формате,краткое отзыв и резюме текста на английском языке. Время написания работы – три часа. Основным предметом оценки для первого задания (логической задачи) будет являться умение логически мыслить и выдвигать аргументированные доказательства. Для второй задачи будет важно знакомство с современными подходами к решению задач в области компьютерной лингвистики, с принципами тестирования и оценки, внимание к языковым свойствам рассматриваемых единиц и конструкций. Также будет важно умение формулировать ход рассуждения и находить оригинальные решения, умение ясно и аргументированно изложить свои мысли, построить текст, не содержащий внутренних противоречий. Объем первого задания не ограничен. В третьем задании будет оцениваться точность и корректность резюме английского текста, имеющего специальную проблематику. В резюме должна быть четко изложена позиция автора по отношению к обсуждаемому вопросу, а также сформулирована собственная позиция автора резюме.

оцениваемые навыки	критерии оценки	баллы
Задание 1		
Содержательная точность	Точность изложения содержания текста, анализа предоставленных данных, адекватность представления авторской позиции	0-15
Качество резюме/ответа	Логичность и краткость изложения, сформулированность собственной позиции автора,	0-10

	отсутствие грубых стилистических, грамматических и орфографических ошибок	
задание 2		
Способность к логическому рассуждению	Наличие правильно решенного задания	0-20
задание 3		
Аргументированность и обоснованность изложения	Полнота, аргументированность, непротиворечивость, структурированность изложения, умение критически мыслить, убедительность приведенных примеров, обоснованность выводов	0-15
Понимание проблематики, творческий подход	Понимание круга проблем кейса, выделение лингвистических компонентов поставленной задачи, структурированное понимание круга проблем, оригинальность решений,	0-20
Содержательная полнота текста	Полнота ответов на ключевые вопросы, наличие внутренней связи между ответами на вопросы, убедительность внутренней структуры работы	0-15
Стилистика и грамотность	Отсутствие грубых стилистических, грамматических и орфографических ошибок	0-5
Итого: максимально 100 баллов		

Перечень и содержание тем олимпиадных состязаний:

1. Уровни обработки лингвистической информации:- токенизация,
- лемматизация (стемминг),
- морфологическая,
- синтаксическая,
- семантическая разметка,
- дизамбигуация.

2. Основные направления компьютерной лингвистики:- машинный перевод,
- реферирование текста,
- автоматическая генерация текста,
- извлечение данных (opinion mining, data mining),
- кластеризация текстов.

3. Основные ресурсы и компьютерные инструменты:

- поисковые системы,
- корпуса,
- базы данных,
- онтологии.

4. Логические задачи.

Список рекомендуемой литературы (компьютерная лингвистика):

1. Леонтьева Н.Н. Автоматическое понимание текстов: Системы, модели, ресурсы. Академия, 2006
2. Jurafsky, Daniel, and James H. Martin. 2009.Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics. 2nd edition. Prentice-Hall. Творческие задания (кейсы) предлагаются по одной из предложенных ниже тем Автоматическая генерация текстов
4. Болдасов М.В., Соколова Е.Г. Генерация текстов на естественном языке – состояние вопроса и прикладные системы // НТИ, Серия 2, No10, 2005, с.12-22.
5. Болдасов М.В., Соколова Е.Г. Генерация текстов на естественном языке – теории, методы, технологии// НТИ. Сер. 2. Информационные процессы и системы. 2006. 11
6. Cécile L. Paris, William R. Swartout, William C. Mann Natural language generation inartificial intelligence and computational linguistics. Springer, 1991
7. Ресурсы, связанные с автоматической генерацией текстов в сети: ACL Special Interest Group on Natural Language Generation (SIGGEN) http://www.siggen.org/ Машинный перевод8. Хроменков П.Н. Современные системы машинного перевода. М., 2005.
9. Иомдин Л.Л. Правиловая система машинного перевода ЭТАП-3: опыт разработки и некоторые уроки. Презентация на семинаре Яндекса.http://download.yandex.ru/company/experience/seminars/etapoverviewrusyandex.pdf 9
10. Hutchins J., Machine translation: history of research and use. In: Encyclopedia of Languages and Linguistics. 2nd edition, edited by Keith Brown (Oxford: Elsevier 2006), vol.7, pp.375-383.http://www.hutchinsweb.me.uk/EncLangLing-2006.pdf
11. Ресурсы по статистичекому машинному переводу: Statistical machine translationhttp://www.statmt.org/ Автоматическое извлечение именованных сущностей
12. Nadeau, David and Satoshi Sekine (2007) A survey of named entity recognition and classification. Linguisticae Investigationes
http://nlp.cs.nyu.edu/sekine/papers/li07.pdf
30(1):3–26.
13. Andrew Mccallum, Information Extraction, ACM Queue 2005
14. http://www.cs.umass.edu/~mccallum/papers/acm-queue-ie.pdf
15. Обзор систем, занимающихся извлечением именованныхhttp://pullenti.ru/CompetitorPage.aspx
16. Описание сервиса пресс-портреты Яндекса: http://help.yandex.ru/news/?id=1111171 Примеры логических задач (с решениями): 17. http://www.problems.ru
сущностей:

Список рекомендуемой литературы (цифровые методы в гуманитарных науках):

Основные источники:
Цифровые гуманитарные науки: хрестоматия / под ред. М. Террас, Д. Найхан, Э. Ванхутта, И. Кижнер. — Красноярск: Сиб. федер. ун-т, 2017. — 352 с.
Bod R., Richards L. A New History of the Humanities: The Search for Principles and Patterns from Antiquity to the Present. Oxford: Oxford University Press, 2013.
Schreibman S., Siemens R., Unsworth J. Companion to Digital Humanities (Blackwell Companions to Literature and Culture). Oxford: Blackwell Publishing Professional, 2004

Сборники конференций Альянса организаций цифровых гуманитарных исследований (ADHO):
a. https://dh2018.adho.org/abstracts/
b. https://dh2017.adho.org/program/abstracts/c. http://dh2016.adho.org/abstracts/
d. http://dh2015.org/abstracts/
e. ...

Дополнительные источники:

Лотман Ю. М. Литературоведение должно быть наукой // Вопросы литературы. 1967. No 1. С. 90–100.
Моретти Ф. Дальнее чтение. М., Издательство Института Гайдара, 2016.
Шапир М. И. «Тебе числа и меры нет». О возможностях и границах «точных методов» в гуманитарных науках. // Вопросы языкознания. 2005. No 1. С. 43–62.
Jones S. E. Roberto Busa, S. J., and the Emergence of Humanities Computing: The Priest and the Punched Cards. London: Routledge, 2016
Hoover D.L., Culpeper J., O'Halloran K. Digital Literary Studies: Corpus Approaches to Poetry, Prose, and Drama. New York, London: Taylor & Francis, 2014.
McCarty W. Humanities Computing. London and NY: Palgrave, 2005
Robinson P. Towards a Theory of Digital Editions. // Variants: The Journal
of the European Society for Textual Scholarship, Vol. 10, p. 105, 2013
Schreibman S., Siemens R. A Companion to Digital Literary Studies.
Oxford: Blackwell, 2008.