Олимпиада для студентов и выпускников вузов – 2012 г.

Демонстрационный вариант и методические рекомендации по направлению «Филология»

Профили:
«Компаративистика: русская литература в кросс-культурной перспективе»
«Компьютерная лингвистика»

ДЕМОНСТРАЦИОННЫЙ ВАРИАНТ

Время выполнения задания – 180 минут

В соответствии со своим выбором программы магистерской подготовки выберите и выполните <u>только один</u> блок заданий.

Блок «КОМПАРАТИВИСТИКА: РУССКАЯ ЛИТЕРАТУРА В КРОСС-КУЛЬТУРНОЙ ПЕРСПЕКТИВЕ»

Выберите одну из предложенных тем для написания эссе:

1. Переводима ли поэзия?
2. Тынянов и Бахтин: опыт сверхкраткого сопоставления.
3. Расскажите о Вашем любимом российском или зарубежном филологическом журнале.

Блок «КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА»

Задание 1.

Перед вами он-лайн система “коммуникативный агент”. Ее функция состоит в том, чтобы общаться с пользователем на заданную тему на естественном языке. Система может отвечать на поставленные вопросы и поддерживать разговор. Коммуникативный агент настроен на определенную область: досуг в вашем городе. Агент помогает пользователю сориентироваться в расписании и репертуаре кинотеатров, театров, концертов, выбрать ресторан для встречи с друзьями, место, где можно провести время с детьми. Одним словом, разговор с агентом должен помочь пользователю выбрать, как и где он будет проводить свое свободное время.

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 естественного языка, важны для предлагаемого сервиса и почему? Какие из них абсолютно необходимы, а без каких можно обойтись?

2) С помощью каких запросов эти функции могут быть протестированы?

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?

4) Могут ли недостатки той или иной лингвистической функции быть компенсированы экстралингвистическими (внеязыковыми) средствами? Какими?

Задание 2.

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.

Seeding, transduction, out-of-sample error and the Microsoft approach...

My past master's student Adam Teichert (now at JHU) did some work on inducing part of speech taggers using typological information. We wanted to compare the usefulness of using small amounts of linguistic information with small amounts of lexical information in the form of seeds. (Other papers give seeds different names, like initial dictionaries or prototypes or whatever... it's all the same basic idea.)

The basic result was that if you *don't* use seeds, then typological information can help a lot. If you do you seeds, then your baseline performance jumps from like 5% to about 40% and then using typological information on top of this isn't really that beneficial.

This was a bit frustrating, and led us to think more about the problem. The way we got seeds was to look at the wikipedia page about Portuguese (for instance) and use *their* example list of words for each tag. An alternative popular way is to use labeled data and extract the few most frequent words for each part of speech type. They're not identical, but there is definitely quite a bit of overlap between the words that Wikipedia lists as examples of determiners and the most frequent determiners (this correlation is especially strong for closed-class words).

In terms of end performance, there are two reasons seeds can help. The first, which is the*interesting* case, is that knowing that "the" is a determiner helps you find other determiners (like "a") and perhaps also nouns (for instance, knowing the determiners often precede nouns in Portuguese). The second, which is the *uninteresting* case, is that now every time you see one of your seeds, you pretty much always get it right. In other words, just by specifying seeds, especially by frequency (or approximately by frequency ala Wikipedia), you're basically ensuring that you get 90% accuracy (due to ambiguity) on some large fraction of the corpus (again, especially for closed-class words which have short tails).

This phenomena is mentioned in the text (but not the tables :P), for instance, in Haghighi & Klein's 2006 NAACL paper on prototype-driven POS tagging, wherein they say: "Adding prototypes ... gave an accuracy of 68.8% on all tokens, but only 47.7% on non-prototype occurrences, which is only a marginal improvement over [a baseline system

with no prototypes." Their improved system remedies this and achieves better accuracy on non-prototypes as well as prototypes (aka seeds).

This is very similar to the idea of transductive learning in machine learning land. Transduction is an alternative to semi-supervised learning. The setting is that you get a bunch of data, some of which is labeled and some of which is unlabeled. Your goal is to simply label the unlabeled data. You *need not* "induce" the labeling function (though many approach do, in passing).

The interesting thing is that learning with seeds is very similar to transductive learning, though perhaps with a bit stronger assumption of noise on the "labeled" part. The irony is that in machine learning land, you would *never* report "combined training and test accuracy" -- this would be ridiculous. Yet this is what we seem to like to do in NLP land. This is itself related to an old idea in machine learning wherein you rate yourself only on test example that you *didn't*see at training time. This is your out-of-sample error, and is obviously much harder than your standard generalization error. (The famous no-free-lunch theorems are from an out-of-sample analysis.) The funny thing out of sample error is that sometimes you prefer *not* to get more training examples, because you then know you won't be tested on it! If you were getting it right already, this just hurts you. (Perhaps you should be allowed to see *x* and say "no I don't want to see *y*"?)

I think the key question is: what are we trying to do. If we're trying to build good taggers (i.e., we're engineers) then overall accuracy is what we care about and including "seed" performance in our evaluations make sense. But when we're talking about 45% tagging accuracy (like Adam and I were), then this is a pretty pathetic claim. In the case that we're trying to understand learning algorithms and study their performance on real data (i.e., we're scientists) then accuracy on non-seeds is perhaps more interesting. (Please don't jump on me for the engineer/scientist distinction: it's obviously much more subtle than this.)

This also reminds me of something Eric Brill said to me when I was working with him as a summer intern in MLAS at Microsoft (back when MLAS existed and back when Eric was in MLAS....). We were working on web search stuff. His comment was that he really didn't care about doing well on the 1000 most frequent queries. Microsoft could always hire a couple annotators to manually do a good job on these queries. And in fact, this is what is often done. What we care about is the heavy tail, where there are too many somewhat common things to have humans annotate them all. This is precisely the same situation here. I can easily get 1000 seeds for a new language. Do I actually care how well I do on those, or do I care how well I do on the other 20000+ things?

МЕТОДИЧЕСКИЕ РЕКОМЕНДАЦИИ

ПРОФИЛЬ «КОМПАРАТИВИСТИКА: РУССКАЯ ЛИТЕРАТУРА В КРОСС-КУЛЬТУРНОЙ ПЕРСПЕКТИВЕ»

Эссе представляет собой небольшое сочинение творческого характера, содержащее развернутое, аргументированное и законченное изложение одного или нескольких тезисов, самостоятельно выдвинутых автором в связи с выбранной им темой. В качестве такого тезиса может выступить как гипотеза научно-исследовательского характера, так и критическое суждение о том или ином литературном факте. Автор имеет право сузить заданную тему, приведя аргументы для конкретизации в самом тексте эссе и, при необходимости, добавив подзаголовок.

Работая над эссе, участник олимпиады должен продемонстрировать знание художественных текстов и научной литературы, в частности, включенной в рекомендательный список ниже. Выдвигаемые аргументы должны быть подкреплены отсылками к источникам и научной литературе, не допускается цитирование без кавычек и пересказ без указания конкретного автора и его сочинения. При этом не требуется полного оформления библиографической ссылки, достаточно в самом тексте эссе, во внутритекстовой или постраничной сноске указать фамилию и инициалы автора, название сочинения и, по возможности, год выхода.

Объем эссе определяется желанием автора и временем, отведенным на его написание. Работу требуется сдать по истечении 150 минут после начала олимпиады, продление отведенного времени не допускается.

Критерии оценки

В ходе оценки письменных работ во внимание принимаются следующие аспекты:

1. Содержательность и оригинальность выдвинутых автором тезисов.
2. Логичность и последовательность изложения, композиционная и содержательная законченность текста.
3. Способность автора убедительно использовать в ходе аргументации собственные знания в области науки о литературе и других гуманитарных наук.
4. Корректное использование научной терминологии.

5. Способность автора обращаться с «чужим словом», точность передачи идей других авторов, соблюдения правил цитирования.
6. Грамотная русская речь, отсутствие грамматических и стилистических ошибок.

Примерные темы эссе:

1. О моем самом нелюбимом великом русском поэте XX века.
2. Переводима ли поэзия?
3. Рецензия на любое современное прозаическое произведение.
4. Анализ любого поэтического произведения по выбору автора.
5. Тынянов и Бахтин: опыт сверхкраткого сопоставления.
6. Рецензия на любую зарубежную экранизацию любого русского произведения.
7. Зачем поэты и прозаики в своих текстах всё время цитируют друг друга?
8. Зачем филологу, занимающемуся русской литературой, знать иностранные языки?
9. В чем главные заслуги Ю. М. Лотмана как филолога?
10. Расскажите о Вашем любимом российском или зарубежном филологическом журнале.

Список литературы для подготовки к олимпиаде

Список носит рекомендательный характер и является лишь ориентиром для самостоятельной подготовки к написанию эссе. Приветствуется осведомленность участников олимпиады о работах классиков филологической науки и современном состоянии литературоведения.

Автономова Н.С. Открытая структура: Якобсон — Бахтин — Лотман — Гаспаров. М., 2009

Барт Р. Избранные работы. Семиотика. Поэтика. М., 1994.

Бахтин М.М. Эстетика словесного творчества. М., 1979.

Гаспаров Б.М. Литературные лейтмотивы. Очерки по русской литературе XX века. М., 1993.

Гаспаров М.Л. Записи и выписки. М., 2000.

Гаспаров М.Л. Избранные труды: В 3 т. М., 1997.

Гинзбург Л.Я. Записные книжки. Воспоминания. Эссе. СПб, 2002.

Дубин Б. В. Слово — письмо — литература: Очерки по социологии современной культуры. М., 2001.

Женнет Ж. Фигуры: В 2-х томах. М., 1998.

Изер В. Изменение функций литературы; Процесс чтения: феноменологический подход // Современная литературная теория: антология. М., 2004. С.3 – 45, 201 – 225.

Калашникова Е. По-русски с любовью. Беседы с переводчиками. М., 2008

Компаньон А. Демон теории. Литература и здравый смысл. М., 2001.

Лотман Ю.М. Анализ поэтического текста: Структура стиха. Л., 1972.

Лотман Ю.М. Избранные статьи: в 3 т. Таллинн, 1993.

Сегал Д.М. Пути и вехи: Русское литературоведение в двадцатом веке. Пути и вехи: Русское литературоведение в двадцатом веке. М., 2011.

Старобинский Ж. Поэзия и знание. История литературы и культуры: в 2-х томах. М.,2002.

Тынянов Ю.Н. Поэтика. История литературы. Кино. М., 1977.

Цивьян Ю. На подступах к карпалистике. Движение и жест в литературе, искусстве, кино. М., 2010.

Чудакова М.О. Избранные работы: том I. Литература советского прошлого. М., 2001.

Эткинд Е.Г. Поэзия и перевод. М., 1963.

Ю.М. Лотман и тартуско-московская семиотическая школа. М., 1994.

Якобсон Р.О. Работы по поэтике. М., 1987.

Яусс Х.-Р. История литературы как провокация литературоведения // Новое литературное обозрение. 1995. № 12. С.34-84.

Краткая информация о магистерской программе

Магистерская программа: «Компаративистика: русская литература в кросс-культурной перспективе»

Направление: 032700.68 «Филология»

Где читается: факультет филологии

Первый набор на программу: 2012 год

Количество бюджетных мест (бесплатное обучение): 15

Вступительные экзамены: русская литература (устно); иностранный язык (устно).

Руководители магистерской программы:

Осповат Александр Львович – профессор факультета филологии НИУ-ВШЭ, руководитель направления «Филология».

Лекманов Олег Андершанович, доктор филологических наук, профессор факультета филологии НИУ-ВШЭ

Магистерская программа по литературоведческой компаративистике нацелена на фундаментальную подготовку магистров-филологов, свободно ориентирующихся в проблематике не только отечественной, но и зарубежной филологии, способных к междисциплинарным исследованиям с учетом актуальных научных тенденций в других областях гуманитарного знания. **История русской литературы** изучается в широком контексте **взаимопроникновения и взаимоотталкивания национальных традиций;** в центре внимания оказываются механизмы рецепции и культурного обмена, статус текста в инонациональном каноне, история национальных мифологий, проблемы перевода и культурной непереводимости.

Магистерская программа ориентирована на фундаментальную профессиональную подготовку выпускников **в следующих предметных областях**:

- Филологический анализ текстов на русском и иностранных языках в историко-культурной и компаративной перспективе;
- Литературная компаративистика;
- История русской литературы;
- История филологии и гуманитарные науки;

Основу программы составляют авторские курсы ведущих специалистов по литературоведческой компаративистике. Дисциплины, **связанные с изучением европейских культур и текстов, читаются на иностранных языках** (на английском – обязательно, на французском или немецком - по выбору). Обязательное участие в научно-исследовательском семинаре предполагает **включение в коллективные научно-исследовательские проекты** факультета филологии и лаборатории лингвосемиотических исследований, сотрудничество с ведущими

научно-исследовательскими центрами – ИМЛИ РАН, ИРЛИ РАН (Пушкинский дом), РГАЛИ.

Выпускники программы приобретают не только **научно-исследовательский опыт**, позволяющий успешно продолжить академическую карьеру в аспирантуре, но и **опыт организационной и проектной работы**, который в дальнейшем может быть применен в различных профессиональных сферах. Обучение по данной магистерской программе направлено на подготовку к таким видам филологической деятельности, как **преподавательская**, **редакционно-издательская** (в том числе связанная с иностранными языками), **переводческая**, **экспертно-аналитическая**, **международная проектная деятельность** в гуманитарной и образовательной сферах.

Поступающим

Студентами магистерской программы могут стать выпускники бакалавриата или специалитета, имеющие гуманитарное образование: филологическое, философское, лингвистическое, культурологическое, историческое. Для магистрантов, успешно сдавших вступительные экзамены, но не имеющих базового филологического образования, предусмотрены адаптационные курсы, помогающие быстро достичь необходимого уровня для успешного освоения программы.

Обязательным условием обучения по магистерской программе является знание английского и еще одного европейского языка. Магистрантам с недостаточным уровнем знания второго языка будут предложены обязательные дополнительные занятия.

ПРОФИЛЬ «КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА»

Олимпиада по направлению «Компьютерная лингвистика» проводится с целью конкурсного отбора на магистерскую программу «Компьютерная лингвистика». Основная цель программы – подготовка специалистов, владеющих базовыми теоретическими знаниями по методам и алгоритмам решения задач в области автоматической обработки, способных участвовать в самых современных проектах, связанных с языковыми технологиями, а также самим их создавать, формулировать новые задачи и предлагать алгоритмы их решения.

Председатель конкурсной комиссии – заместитель декана филологического факультета по направлению “Лингвистика”, доктор филологических наук Е.В. Рахилина.

Олимпиада предполагает два задания: творческая работа в формате кейс-стади и краткое резюме на русском языке текста по специальности на русском языке. Время написания работы – три часа.

Магистерская программа по компьютерной лингвистике носит междисциплинарный характер и имеет целью привлечь абитуриентов, имеющих базовое образование в области математики, лингвистики и информационных технологий и интересующихся современными методами обработки языковых данных. Основным предметом оценки для первого задания будет являться творческий подход, проявленный участниками олимпиады, умение формулировать задачи и находить их решения, умение ясно и аргументированно изложить свои мысли, построить текст, не содержащий внутренних противоречий. Объем первого задания не ограничен. Во втором задании будет оцениваться точность и корректность резюме английского текста, имеющего специальную проблематику. Объем второго задания – не более 1000 знаков.

Основные критерии оценки

Оцениваемые навыки	**Критерии оценки**	**Баллы**
	задание 1	
Аргументированность и обоснованность изложения	Полнота, аргументированность, непротиворечивость, структурированность изложения, умение критически мыслить, убедительность приведенных примеров, обоснованность выводов	0-25
Понимание проблематики, творческий подход	Понимание круга проблем кейса, умение их структурировать, оригинальность решений,	0-20

	представление о процессе автоматического анализа языковых данных в целом	
Содержательная полнота текста	Полнота ответов на ключевые вопросы, наличие внутренней связи между ответами на вопросы, убедительность внутренней структуры работы	0-20
Стилистика и грамотность	Отсутствие грубых стилистических, грамматических и орфографических ошибок	0-10
задание 2		
Содержательная точность	Точность изложения содержания текста	0-20
Качество резюме	Логичность и краткость изложения, отсутствие грубых стилистических, грамматических и орфографических ошибок	0-5
Итого: максимально 100 баллов		

Литература

Леонтьева Н.Н. Автоматическое понимание текстов: Системы, модели, ресурсы. Академия, 2006

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

Болдасов М.В., Соколова Е.Г. Генерация текстов на естественном языке – состояние вопроса и прикладные системы // НТИ, Серия 2, №10, 2005, с.12-22.

Болдасов М.В., Соколова Е.Г. Генерация текстов на естественном языке – теории, методы, технологии// НТИ. Сер. 2. Информационные процессы и системы. 2006.

Cécile L. Paris, William R. Swartout, William C. Mann Natural language generation in artificial intelligence and computational linguistics. Springer, 1991

Ресурсы, связанные с автоматической генерацией текстов в сети: ACL Special Interest Group on Natural Language Generation (SIGGEN) http://www.siggen.org/

Машинный перевод

Хроменков П.Н. Современные системы машинного перевода. М., 2005.

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

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

Ресурсы по статистичекому машинному переводу: Statistical machine translation http://www.statmt.org/

Автоматическое извлечение именованных сущностей

Nadeau, David and Satoshi Sekine (2007) A survey of named entity recognition and classification. Linguisticae Investigationes 30(1):3–26. http://nlp.cs.nyu.edu/sekine/papers/li07.pdf

Andrew Mccallum, Information Extraction, ACM Queue 2005

http://www.cs.umass.edu/~mccallum/papers/acm-queue-ie.pdf

Обзор систем, занимающихся извлечением именованных сущностей: http://pullenti.ru/CompetitorPage.aspx

Описание сервиса пресс-портреты Яндекса http://help.yandex.ru/news/?id=1111171