ИНСТИТУТ СИСТЕМ ИНФОРМАТИКИ
ИМ. А.П. ЕРШОВА СО РАН

Т.В. БАТУРА

УЧЕБНАЯ ПРОГРАММА ДИСЦИПЛИНЫ «МАТЕМАТИЧЕСКАЯ ЛИНГВИСТИКА И ОБРАБОТКА ТЕКСТОВ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ»

основной образовательной программы послевузовского профессионального образования (аспирантура) по специальности 05.13.11 «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей»

НОВОСИБИРСК
ИЗДАТЕЛЬСТВО СИБИРСКОГО ОТДЕЛЕНИЯ
РОССИЙСКОЙ АКАДЕМИИ НАУК
2017

УДК [519.765 + 004.942] (073)
ББК 32.81я7
Б28

Батура Т.В.
Учебная программа дисциплины «Математическая лингвистика и обработка текстов на естественном языке» основной образовательной программы послевузовского профессионального образования (аспирантура) по специальности 05.13.11 «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей» / Т.В. Батура; Рос. акад. наук, Сиб. отдние, Ин-т систем информатики им. А.П. Ершова. – Новосибирск: Изд-во СО РАН, 2017. – 11 с.

Учебная программа соответствует курсу лекций по дисциплине «Математическая лингвистика и обработка текстов на естественном языке», который читается в аспирантуре Института систем информатики им. А.П. Ешова СО РАН. Целью освоения дисциплины является изучение основных методов автоматизированной обработки текстов на естественном языке: формальных методов анализа текстов, алгоритмов семантического поиска и извлечения информации, знакомство с основами теории речевых действий, особенностями построения тезаурусов и основами корпусной лингвистики. Уделяется внимание практическому применению обработки текстовой информации в поисковых системах и системах автоматического определения авторства, при анализе социальных сетей и спам-сообщений.

Учебная программа утверждена к печати
Ученым советом Института систем информатики
им. А.П. Ершова СО РАН
18 сентября 2015 г., протокол № 5-2015

ISBN 978-5-7692-1557-5

1. Цели освоения дисциплины

Целями освоения дисциплины «Математическая лингвистика и обработка текстов на естественном языке» являются ознакомление с основными методами математической логики, теории вероятностей и математической статистики; элементами автоматизированной обработки текстов: формальными методами анализа текстов, алгоритмами семантического поиска и извлечения информации, элементами теории речевых действий; особенностями построения тезаурусов; основами корпусной лингвистики. Также уделяется внимание практическому применению принципов обработки текстовой информации в поисковых системах и системах автоматического определения авторства, при анализе социальных сетей и спам-сообщений, необходимых для самостоятельной работы в научно-исследовательской сфере.

Для достижения поставленной цели выделяются следующие задачи:

- изучение математических основ наиболее интересных и важных для приложений алгоритмов из теории информации, обработки текстов на естественном языке;
- ознакомление с нестандартными методами обработки информации: нейрокомпьютерный подход, методы кластеризации, нечеткая логика Заде;
- ознакомление с методами обработки текстовой информации: алгоритмами морфологического и синтаксического анализа, методами классификации и кластеризации, алгоритмами поиска ключевых слов и др.

2. Место дисциплины в структуре основной профессиональной образовательной программы послевузовского профессионального образования (аспирантура)

Дисциплина «Математическая лингвистика и обработка текстов на естественном языке» (ФД.А.02) относится к группе факультативных дисциплин по специальности 05.13.11.

3. Требования к уровню подготовки аспиранта, завершившего изучение данной дисциплины

Аспиранты, завершившие изучение данной дисциплины, должны:

- знать содержание программы курса, принципы функционирования автоматизированных лингвистических систем, формулировки задач, условия применимости и характеристики рассмотренных в курсе методов;
- уметь применять методы математической лингвистики для анализа текстовой информации на естественном языке;

- владеть методами математической лингвистики для анализа текстовой информации на естественном языке.

4. Объем дисциплины и виды учебной работы

Общая трудоемкость дисциплины составляет 2 зачетных единицы 72 часа.

№ п/п	Наименование тем и разделов	Всего (часов)	Аудиторные занятия (часов), в том числе			Самостоятельная работа
			Лекции	Семинары	Лабораторные работы	
1	2	3	4	5	6	7
1	Общие принципы построения систем автоматизированной обработки текстов	4	2	2	—	—
2	Синтаксическая структура предложения. Методы синтаксического анализа	8	4	4	—	—
3	Фрагментационный анализ. Взаимодействие синтаксического и фрагментационного анализа	4	2	2	—	—
4	Семантический анализ текстов. Лексические функции. Валентности слов	4	2	2	—	—
5	Теоретико-множественные модели языка	4	2	2	—	—
6	Теория речевых действий. Классификация речевых действий	4	2	2	—	—
7	Системы машинного перевода, электронные словари, тезаурусы, онтологии. Общие принципы построения	4	2	2	—	—

1	2	3	4	5	6	7
8	Представление знаний для компьютерной обработки. Семантические сети. Фреймы. Формальные логические модели	6	3	3	—	—
9	Корпусная лингвистика. Частотные методы в компьютерной лингвистике	4	2	3	—	—
10	Классификация и кластеризация. Иерархические и вероятностные подходы	6	3	3	—	—
11	Автоматические системы извлечения информации. Алгоритмические основы	4	2	2	—	—
12	Формальные методы атрибуции текстов	8	4	4	—	—
13	Анализ социальных сетей. Направления и методы исследований	8	4	4	—	—
14	Методы обнаружения спама	4	2	2	—	—
	ИТОГО:	72	36	36	—	—

5. Разделы дисциплины и виды занятий

№ п/п	Название раздела дисциплины	Объем часов / зачетных единиц					
		Всего аудиторных часов	из них				Самостоятельная работа
			лекции	семинары	практические занятия	КСР	
1	Математическая лингвистика и обработка текстов на естественном языке	36	—	—	—	36	36
2	Математическая лингвистика и обработка текстов на естественном языке	36	—	—	—	36	36

6. Содержание дисциплины

6.1. Новизна курса (научная, содержательная; сравнительный анализ с подобными курсами в России и за рубежом)

Первый раздел служит основой и знакомит студентов с общими принципами и задачами построения систем автоматической обработки текстовой информации. Он подготавливает студентов к дальнейшему обсуждению возникающих проблем и их решению. Во втором разделе рассматриваются довольно сложные вопросы семантического анализа текстов, подходы и методы, недостаточно и разрозненно освещенные в литературе. В третьем разделе изучаются способы практического применения алгоритмов обработки текстов в различных системах поиска и извлечения информации, системах определения авторства, при анализе текстового контента социальных сетей.

6.2. Содержание разделов и тем курса

1. Основные понятия математической лингвистики. Системы автоматизированной обработки текстов. Общие принципы построения. Графематический и морфологический анализ.
2. Методы задания синтаксической структуры предложений. Системы составляющих. Деревья подчинения. Минимальные схемы предложений. Методы синтаксического анализа. Применение морфологического и синтаксического анализа в поисковых системах.
3. Фрагментационный анализ. Взаимодействие синтаксического и фрагментационного анализа.
4. Семантический анализ текстов. Лексические функции. Валентности слов. Меры семантической близости.
5. Теоретико-множественные модели языка. Основные определения: отмеченные последовательности, контексты, дистрибутивные классы и др. Формализация понятий: «часть речи», «синтаксический тип», «грамматический род», «категории падежа».
6. Теория речевых действий. Классификация речевых действий.
7. Системы машинного перевода. Электронные словари, тезаурусы, онтологии. Общие принципы построения.
8. Представление знаний для компьютерной обработки. Семантические сети. Фреймы. Формальные логические модели. Искусственные языки и нотации, применяемые в компьютерной лингвистике.
9. Корпусная лингвистика. Частотные методы в компьютерной лингвистике.

10. Модели и методы автоматической классификации и кластеризации текстовой информации. Иерархические и вероятностные подходы. Интеллектуальный анализ данных.
11. Автоматические системы извлечения информации. Алгоритмические основы. Принципы обработки неструктурированной и плохо структурированной информации. Тематическая индексация текстов.
12. Формальные методы определения авторства текстов. Лингвостатистические параметры текста. Статистические методы атрибуции. Авторский инвариант и лингвистические спектры. Применение методов кластеризации и классификации для установления авторства текстов.
13. Социальные сети. Направления исследований. Графовые модели анализа социальных сетей. Понятие центральности. Методы обнаружения сообществ и анализ связных подгрупп. Модели динамики сети.
14. Методы обнаружения спама: вероятностные и статистические, байесовский классификатор.

6.3. Перечень примерных контрольных вопросов и заданий для самостоятельной работы

1. Перечислить направления компьютерной лингвистики.
2. Сформулировать общие принципы построения автоматизированных систем обработки текстов.
3. Разъяснить принципы работы графематического и морфологического анализаторов.
4. Перечислить методы задания синтаксической структуры предложений.
5. Разъяснить принципы работы фрагментационного и синтаксического анализаторов. Описать принцип их взаимодействия.
6. Изложить основные идеи подхода И. Мельчука к семантическому анализу.
7. Привести примеры мер семантической близости.
8. Дать определения отмеченных последовательностей, контекста, дистрибутивных классов.
9. Дать формальные определения частей речи, грамматического рода и категории падежа в терминах модели языка, предложенной С. Маркусом.
10. Изложить основные идеи теории речевых действий.
11. Привести классификацию речевых действий.

12. Сформулировать принципы построения систем машинного перевода, электронных словарей, тезаурусов, онтологий.
13. Дать определения семантических сетей, фреймов.
14. Неточные рассуждения. Что такое логика Заде?
15. Привести примеры искусственных языков и нотаций.
16. Что такое корпусная лингвистика?
17. Применение частотных методов в компьютерной лингвистике; перечислить, описать, привести примеры.
18. Назвать отличие классификации текстов от их кластеризации.
19. Перечислить методы классификации и кластеризации текстовой информации; сформулировать основные принципы.
20. Разъяснить принципы работы автоматических систем извлечения информации.
21. Сформулировать принципы обработки неструктурированной и плохо структурированной информации. Индексация текстов.
22. Перечислить формальные методы атрибуции текстов.
23. Дать определения лингвостатистических параметров, авторского инварианта и лингвистических спектров.
24. Привести примеры использования методов кластеризации и классификации для определения авторства текстов.
25. Перечислить основные направления исследований социальных сетей.
26. Дать определения центральностей разного типа.
27. Описать методы анализа социальных сетей.
28. Перечислить основные методы обнаружения спам-сообщений. Привести примеры.
29. Пояснить принцип работы байесовского классификатора.

7. Самостоятельная работа аспирантов

7.1. Изучение основной и дополнительной литературы по вопросам программы.

7.2. Примерная тематика рефератов, курсовых работ:

- аспирантам может быть дано задание написать обзоры по следующим темам: системы автоматического поиска и извлечения информации из текстов, конкретные алгоритмы обработки текстовой информации и их применение в технике, медицине, системах безопасности;

- рефераты предусматриваются в отдельных исключительных случаях, курсовые работы не предусмотрены.

8. Учебно-методическое и информационное обеспечение дисциплины

8.1. Основная и дополнительная литература

а) основная литература:

1. *Батура Т.В.* Методы анализа компьютерных социальных сетей // Вестн. НГУ. Сер.: Информационные технологии. – Новосибисрк, 2012. – Т. 10, вып. 4. – С. 13–28.
2. *Батура Т.В.* Формальные методы определения авторства текстов // Вестн. НГУ. Сер.: Информационные технологии. – Новосибисрк, 2012. – Т. 10, вып. 4. – С. 81–94.
3. *Вежбицка А.* Речевые акты // Новое в зарубежной лингвистике. – М.: Прогресс, 1986. – Вып. 17. – С. 151–169.
4. *Charu C.* Aggarwal Social network data analytics. – 2011. – 520 p.
5. *Захаров В.П.* Корпусная лингвистика: учеб.-метод. пособие. – СПб.: СПбГУ, 2005. – 48 с.
6. *Маркус С.* Теоретико-множественные модели языков. – М.: Наука, 1970. – 332 с.
7. *Мельчук И.А.* Опыт теории лингвистических моделей «Смысл-Текст». – М.: Школа «Языки русской культуры», 1999. – 346 с.
8. *Рубашкин В.Ш.* Представление и анализ смысла в интеллектуальных системах. – М.: Наука, 1989. – 192 с.
9. *Чатуев М.Б., Чеповский А.М.* Частотные методы в компьютерной лингвистике: учеб. пособие. – М.: МГУП, 2011. – 88 с.
10. *Шевелёв О.Г.* Методы автоматической классификации текстов на естественном языке: учеб. пособие. – Томск: ТМЛ-Пресс, 2007. – 144 с.

б) дополнительная литература:

1. *Апресян Ю.Д.* Идеи и методы современной структурной лингвистики. – М.: Просвещение, 1966. – 305 с.
2. *Ануреев И.С., Батура Т.В., Боровикова О.И., Загорулько Ю.А., Кононенко И.С., Марчук А.Г., Марчук П.А., Мурзин Ф.А., Сидорова Е.А., Шилов Н.В.* Модели и методы построения информационных систем, основанных на формальных, логических и лингвистических подходах: моногр. / Ин-т систем информатики им. А.П. Ершова СО РАН. – Новосибирск: Изд-во СО РАН, 2009.

3. *Бернштейн Э.С., Лахути Д.Г., Чернявский В.С.* Вопросы теории поисковых систем. – М.: ОВНИИЭМ, 1966. – 130 с.
4. *Batura T., Murzin F., Proskuryakov A., Trelevich J.* Some Approaches to Detection of Spam and Senders of Spam // Перспективы систем информатики: Восьмая междунар. конф. памяти акад. А.П. Ершова, Наукоемкое программное обеспечение: раб. сем. – Новосибирск, 2011. – С. 1–6.
5. *Заде Л.* Понятие лингвистической переменной и его применение к принятию приближеных решений. – М., 1976. – 166 с.
6. *Захаров В.П.* Информационно-поисковые системы: учеб.-метод. пособие. – СПб., 2005. – 48 с.
7. *Леонтьева Н.Н.* Автоматическое понимание текстов: системы, модели, ресурсы. – М.: ИЦ «Академия», 2006. – 304 с.
8. *Батура Т.В., Мурзин Ф.А.* Машинно-ориентированные логические методы отображения семантики текста на естественном языке: моногр. / Ин-т систем информатики им. А.П. Ершова СО РАН. – Новосибирск: Изд-во НГТУ, 2008. – 248 с.
9. *Потапова Р.К.* Новые информационные технологии и лингвистика. – М.: Эдиториал УРСС, 2004. – 320 с.
10. *Рубашкин В.Ш.* Семантический компонент в системах понимания текста // КИИ-2006. Десятая национальная конференция по искусственному интеллекту с международным участием: труды конф. – М.: Физматлит, 2006. – С. 455–463.

в) программное обеспечение и Интернет-ресурсы: компьютеры стандартные типа Pentium, программное обеспечение MS Visual C++, Maple 5.4, Matlab 7.0, Графические редакторы, MS Office.

8.2. Перечень вопросов и заданий (аттестации) и/или тем рефератов

Перечень вопросов совпадает с расшифровкой к пункту 6.3.

9. Материально-техническое обеспечение дисциплины

Персональные компьютеры слушателей курса и преподавателя.

Учебное издание

Батура Татьяна Викторовна

УЧЕБНАЯ ПРОГРАММА ДИСЦИПЛИНЫ «МАТЕМАТИЧЕСКАЯ ЛИНГВИСТИКА И ОБРАБОТКА ТЕКСТОВ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ»

основной образовательной программы
послевузовского профессионального образования (аспирантура)
по специальности 05.13.11 «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей»

Редактор *Ф.Х. Сагалаева*

Подписано в печать 16.08.2017. Формат 60×84 1/16.
Усл.-печ. л. 0,7. Уч.-изд. л. 0,4. Тираж 30 экз. Заказ № 201

Издательство СО РАН
630090, Новосибирск, Морской просп., 2
E-mail: psb@ad-sbras.nsc.ru
Тел. (383) 330-80-50
Отпечатано в Идательстве СО РАН
Интернет-магазин Издательства СО РАН
http://www.sibran.ru