Программа разработана экспертами
Федерального учебно-методического объединения
высшего образования по укрупненной группе
специальностей и направлений подготовки
45.00.00 Языкознание и литературоведение

Утверждена на заседании ФУМО
25 мая 2021 года

Примерная программа учебной дисциплины

ИЗВЛЕЧЕНИЕ ИНФОРМАЦИИ ИЗ ТЕКСТА

Уровень высшего образования:

МАГИСТРАТУРА

Направление подготовки:

45.03.03 «ФУНДАМЕНТАЛЬНАЯ И ПРИКЛАДНАЯ ЛИНГВИСТИКА»

Москва 2021

Раздел 1. Характеристики учебных занятий

1.1 Цели и задачи учебных занятий

Целью данного курса является ознакомление студентов с современными техническими средствами и информационными технологиями, служащими для задач автоматического извлечения информации из текста. Результатом занятий должно стать приобретение студентами навыков работы с основными методами автоматического извлечения информации из текстов.

1.2 Место дисциплины (модуля) в структуре образовательной программы, связь с другими дисциплинами (модулями) программы

Относится к вариативной части ОПОП ВО.

1.3 Требования подготовленности обучающегося к освоению содержания учебных занятий (пререквизиты)

Устанавливаются образовательной организацией.

1.4 Перечень результатов обучения

По окончании курса студент должен знать основные современные технические средства и информационные технологии, служащие для обеспечения лингвистической деятельности; уметь изучать и осваивать современные технические средства и информационные технологии; владеть навыками работы с основными современными техническими средствами и информационными технологиями.
Набор компетенций, соотнесенных с результатами обучения, определяется образовательной организацией.

1.5 Перечень рекомендуемых образовательных технологий

В преподавании дисциплины «Извлечение информации из текста» используются разнообразные образовательные технологии как традиционного, так и инновационного характера, учитывающие смешанный, теоретико- и практикоориентированный характер дисциплины:

- лекции;
- практические занятия;
- дискуссии;
- выступления с докладами и сообщениями;
- аудиторные контрольные работы;
- внеаудиторные контрольные работы;
- тестирование.

Степень необходимости образовательной среды и ее выбор определяется образовательной организацией. Формы текущей аттестации определяются образовательной организацией.

1.6 Объем дисциплины (модуля) в зачетных единицах

2 з.е.

Раздел 2. Организация, структура и содержание учебных занятий

2.1 Организация учебных занятий

Предусмотрены учебные занятия с использованием дистанционных технологий.

2.2 Краткая аннотация содержания дисциплины (модуля)

Наименование темы (раздела, части)	Вид учебных занятий	Кол-во часов
1. Современные подходы к извлечению информации из текста. Типы задач, решаемых автоматическим извлечением информации. Типы и способы разметки данных.	Лекции	2
2. Методы разрешения лексической неоднозначности. Простые методы. Метод PageRank в задаче разрешения лексической многозначности. Методы машинного обучения в задаче автоматического разрешения лексической неоднозначности. Метод Decision List.	Практические занятия	2
3. Задача автоматического распознавания значений. Методы и тестирование. Семантические роли, семантические модели управления. Применение. Семантические ресурсы: FrameNet. FrameBank. PropBank. Автоматическая семантическая разметка ролей. Основные подходы. Проблема формирования обучающей выборки для семантической разметки ролей. Подходы к переносу классификатора ролей на другой язык. BabelNet. Назначение, построение. Подключение новых языков.	Практические занятия	4
4. Задачи и методы извлечения информации из текстов. Подход, основанный на словарях и правилах: этапы работы, проблемы. Подход на основе машинного обучения: этапы работы, проблемы. Подготовка обучающей коллекции. Задача извлечения ключевых слов, многокомпонентной лексики из текстов. Задача извлечения отношений: подходы к подготовке обучающей коллекции. Bootstrapping. Distant supervision. Задача извлечения отношений в открытой области: преимущества и проблемы.	Практические занятия	4

5. Задачи и методы контент-анализа в компьютерной лингвистике. Задача анализа тональностей. Особенности словарей оценочной лексики. Подходы к автоматическому извлечению словаря оценочной лексики по корпусу. Методы машинного обучения в задаче анализа тональности. Используемые признаки. Подходы к преодолению ограниченности обучающей выборки в задаче анализа тональности: комбинирование словарей и машинного обучения.	Практические занятия	4
6. Задачи и методы автоматического сравнения текстов Методы кластеризации текстов: иерархическая кластеризация, метод k-средних. Векторные представления слов (word embeddings) на основе нейронных сетей. Программа Word2vec и ее применения. Задача агрегации новостей. Задача проверки текстов на плагиат.	Практические занятия	2
ИТОГО		18

Раздел 3. Обеспечение учебных занятий

3.1 Методические указания по освоению дисциплины

Преподавание дисциплины осуществляется в форме лекций и практических занятий. Во время занятий обучающиеся выполняют практические задания, иллюстрирующие основные задачи и методы автоматического извлечения информации из текстов. Для закрепления пройденного материала предлагаются домашние задания по каждой из тем. Успешное овладение содержанием дисциплины «Извлечение информации из текстов» предполагает работу обучающихся в группах в аудитории, а также их самостоятельную работу.

Дополнительные методические указания устанавливаются образовательной организацией.

3.2 Примерный перечень учебно-методического обеспечения самостоятельной работы обучающихся по дисциплине (модулю), в том числе примерный перечень учебной литературы и ресурсов информационно-телекоммуникационной сети «Интернет», необходимых для освоения дисциплины (модуля)

Самостоятельная работа студентов должна включать усвоение теоретического материала, подготовку к практическим занятиям, выполнение творческих заданий, работу с электронным учебно-методическим комплексом, подготовку к текущему контролю знаний, к промежуточной аттестации (зачету).

Список рекомендованной литературы

Dowty D. 1991. Thematic Proto-Roles and Argument Selection. *Language,* Vol. 67, No. 3. 547-619.

Fillmore C. J. 1968. The Case for Case. In E. Bach, & R. T. Harms (Eds.), *Universals in linguistic theory*. New York, NY: Holt, Rinehart, and Winston. 1-88.

Gildea D. & Jurafsky D. 2000. Automatic Labeling of Semantic Roles. *Computational Linguistics*. 28. 245-288. 10.3115/1075218.1075283.

Hatzivassiloglou V. & McKeown K. 1997. Predicting the Semantic Orientation of Adjectives. 10.3115/979617.979640.

Jurafsky D. & Martin J. H. 2000. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition. Upper Saddle River, N.J: Prentice Hall.

Kilgarriff A., & Rosenzweig J. 2000. Framework and results for English SENSEVAL. *Computers and the Humanities*. 34. 15-48. 10.2307/30204788.

Lesk M. 1986. Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation*. 24-26.

Maynard D., Bontcheva K. and I. Augenstein. 2016. Natural Language Processing for the Semantic Web. Morgan & Claypool Publishers.

McCarthy D., Koeling R., Weeds J., and J. Carroll. 2004. Finding predominant senses in untagged text. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, Barcelona, Spain. 279–286

Mintz M. & Bills S. & Snow R. & Jurafsky D. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, Suntec, Singapore, Association for Computational Linguistics. 1003-1011. 10.3115/1690219.1690287.

Navigli R. and Lapata M. 2007. Graph Connectivity Measures for Unsupervised Word Sense Disambiguation. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, Hyderabad, India. 1683-1688

Navigli R., Ponzetto S. P. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*. Volume 193. 217-250

Pang B. & Lee L. & Vaithyanathan S. 2002. Thumbs up? Sentiment Classification Using Machine Learning Techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*. 79–86. 10.3115/1118693.1118704.

Patwardhan S., Banerjee S., Pedersen T. 2003. Using Measures of Semantic Relatedness for Word Sense Disambiguation. In: Gelbukh A. (eds) *Computational Linguistics and Intelligent Text Processing. CICLing 2003. Lecture Notes in Computer Science,* vol 2588. Springer, Berlin, Heidelberg. 241-257. https://doi.org/10.1007/3-540-36456-0_24

Yarowsky D. 1994. Decision lists for lexical ambiguity resolution: application to accent restoration in Spanish and French. In *Proceedings of ACL '94*. 88-95

Апресян Ю. Д., Богуславский И. М., Иомдин Л. Л. и др. 1981. Лингвистическое обеспечение в системе автоматического перевода ЭТАП-1. *Разработка формальной модели естественного языка*. Новосибирск, 3-28.

Апресян Ю. Д., Богуславский И. М., Иомдин Л. Л., Лазурский А. В., Перцов Н. В., Санников В. З., Цинман Л. Л. 1989. Лингвистическое обеспечение системы ЭТАП-2. М.: Наука.

Баранов А. Н. 2001. Введение в прикладную лингвистику: Учебное пособие. М.: Эдиториал УРСС.

Добров Б. В., Иванов В. В., Лукашевич Н. В., Соловьев В.Д. 2009. Онтологии и тезаурусы: модели, инструменты, приложения. Изд-во ИНТУИТ.

Кибрик А. Е. и др. 2019. Введение в науку о языке. Раздел 5: Прикладная и компьютерная лингвистика. М.: Буки-Веди. 455-534.

Корочков А. В. 2006. О количественной оценке адекватности лингвистических правил (на материале правил чтения для английского языка). *Вопросы Языкознания*. №5, 78-91.

Кузнецов И. П., Сомин Н. В. 2014. Методы автоматического извлечения из текстов семантически значимой информации. М.: ИПИ РАН.

Леонтьева Н. Н. 2006. Автоматическое понимание текстов: системы, модели, ресурсы. М.: Издательский центр "Академия".

Маннинг К. Д и др. 2011. Введение в информационный поиск (пер. с англ.). Издательский дом «Вильямс».

Толдова С. Ю. Извлечение информации из текста // А.Е. Кибрик и др. Введение в науку о языке. М.: Буки-Веди, 2019. С. 527-534.

Описание материально-технической базы, рекомендуемой для осуществления образовательного процесса по дисциплине (модулю)

Учебная аудитория с мультимедийным комплексом.

Описание материально-технической базы (в т.ч. программного обеспечения), рекомендуемой для адаптации электронных и печатных образовательных ресурсов для обучающихся из числа инвалидов и лиц с ОВЗ

Устанавливается образовательной организацией.

3.3 Методика проведения текущего контроля успеваемости и промежуточной аттестации и критерии оценивания

Для контроля усвоения данной дисциплины предусмотрен зачет. Мероприятия по текущему контролю знаний обучающихся проводятся в часы, отведенные для изучения дисциплины.

В течение семестра студентами выполняются практические и контрольные работы.

Порядок проведения зачета определяется ВУЗом.

3.4 Методические материалы для проведения текущего контроля успеваемости и промежуточной аттестации (контрольно-измерительные материалы, оценочные средства)

Примерные вопросы для самоконтроля:

1. Задача автоматического разрешения неоднозначности.
2. Методы разрешения многозначности, основанные на тезаурусах.
3. Методы машинного обучения в задаче автоматического разрешения лексической неоднозначности.
4. Задача автоматического распознавания значений.
5. Семантические роли, семантические модели управления. Применение.
6. Автоматическая семантическая разметка ролей. Основные подходы.
7. Извлечение информации из текстов. Основные подзадачи.
8. Извлечение именованных сущностей на основе машинного обучения.
9. Задача извлечения отношений в открытой области. Сопоставление с другими методами.
10. Задачи автоматического анализа тональности. Проблемы в задаче анализа тональности.
11. Методы машинного обучения в задаче анализа тональности. Используемые признаки.

Примерные практические задания:

1. Оценить работу систем извлечения именованных сущностей:

- https://natasha.github.io/ (основанная на правилах)
- http://demo.ipavlov.ai/#ru (на машинном обучении)

Указать, в чем проявляется различие между подходами (что в какой системе получается хуже, что получается лучше), привести примеры текстов, на которых это видно.

2. Проанализировать работу системы Text Runner (http://openie.allenai.org/): задать несколько сущностей в разных областях и посмотреть, какие выводятся отношения. Представить результат работы в виде короткого отчета с примерами.

Примерный перечень вопросов к зачету (экзамену) по всему курсу:

1. Типы задач, решаемых автоматическим извлечением информации. Типы и способы разметки данных.
2. Задача автоматического разрешения неоднозначности. Простые методы.
3. Методы разрешения многозначности, основанные на тезаурусах.
4. Метод PageRank в задаче разрешения лексической многозначности.
5. Методы машинного обучения в задаче автоматического разрешения лексической неоднозначности. Метод Decision List.
6. Преодоление проблем создания размеченного корпуса. Автоматизированное создание размеченных корпусов.
7. Задача автоматического распознавания значений. Методы и тестирование.
8. Семантические роли, семантические модели управления. Применение.
9. Семантические ресурсы: FrameNet. FrameBank. PropBank. Представление AMR. Назначение, сравнение.
10. Автоматическая семантическая разметка ролей. Основные подходы.
11. Проблема формирования обучающей выборки для семантической разметки ролей. Подходы к переносу классификатора ролей на другой язык.
12. BabelNet. Назначение, построение. Подключение новых языков.
13. Извлечение информации из текстов. Основные подзадачи. Подход, основанный на словарях и правилах. Этапы работы. Проблемы.
14. Извлечение именованных сущностей на основе машинного обучения. Подготовка обучающей коллекции. Применяемые методы. Признаки.

15. Подходы к подготовке обучающей коллекции в задаче извлечения отношений. Bootstrapping. Distant supervision.
16. Задача извлечения отношений в открытой области. Сопоставление с другими методами. Преимущества и проблемы.
17. Задача извлечения ключевых слов, многокомпонентной лексики из текстов.
18. Задачи автоматического анализа тональности. Проблемы в задаче анализа тональности.
19. Особенности словарей оценочной лексики. Подходы к автоматическому извлечению словаря оценочной лексики по корпусу.
20. Методы машинного обучения в задаче анализа тональности. Используемые признаки.
21. Подходы к преодолению ограниченности обучающей выборки в задаче анализа тональности. Комбинирование словарей и машинного обучения.
22. Задачи автоматического сравнения текстов. Методы кластеризации. Векторные представления слов на основе нейронных сетей.

3.5 Материально-техническое обеспечение

Минимально необходимый для реализации курса перечень материально-технического обеспечения включает лекционные аудитории (с компьютерным и видеопроекционным оборудованием для презентаций, средствами звуковоспроизведения и экраном, с выходом в Интернет). Количество индивидуальных рабочих станций должно соответствовать количество студентов.

3.6 Информационное обеспечение

Рекомендуемая основная литература

Jurafsky D. & Martin J. H. 2000. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition. Upper Saddle River, N.J: Prentice Hall.

Баранов А. Н. 2001. Введение в прикладную лингвистику: Учебное пособие. М.: Эдиториал УРСС.

Добров Б. В., Иванов В. В., Лукашевич Н. В., Соловьев В. Д. 2009. Онтологии и тезаурусы: модели, инструменты, приложения. Изд-во ИНТУИТ.

Маннинг К. Д и др. 2011. Введение в информационный поиск (пер. с англ.). Издательский дом «Вильямс».

Толдова С. Ю. 2019. Извлечение информации из текста // А.Е. Кибрик и др. Введение в науку о языке. М.: Буки-Веди. 527-534.

Рекомендуемая дополнительная литература

Dowty D. 1991. Thematic Proto-Roles and Argument Selection. *Language*, Vol. 67, No. 3. 547-619.

Fillmore C. J. 1968. The Case for Case. In *E. Bach, & R. T. Harms (Eds.), Universals in linguistic theory*. New York, NY: Holt, Rinehart, and Winston. 1-88.

Gildea D. & Jurafsky D. 2000. Automatic Labeling of Semantic Roles. *Computational Linguistics*. 28. 245-288. 10.3115/1075218.1075283.

Hatzivassiloglou V. & McKeown K. 1997. Predicting the Semantic Orientation of Adjectives. 10.3115/979617.979640.

Kilgarriff A., & Rosenzweig J. 2000. Framework and results for English SENSEVAL. *Computers and the Humanities*. 34. 15-48. 10.2307/30204788.

Lesk. M. 1986. Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation*. 24-26.

Maynard D., Bontcheva K. and I. Augenstein. 2016. Natural Language Processing for the Semantic Web. Morgan & Claypool Publishers.

McCarthy D., Koeling R., Weeds J., and J. Carroll. 2004. Finding predominant senses in untagged text. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, Barcelona, Spain. 279–286

Mintz M. & Bills S. & Snow R. & Jurafsky D. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, Suntec, Singapore, Association for Computational Linguistics. 1003-1011. 10.3115/1690219.1690287.

Navigli R. and Lapata M. 2007. Graph Connectivity Measures for Unsupervised Word Sense Disambiguation. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, Hyderabad, India. 1683-1688

Navigli R., Ponzetto S. P. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*. Volume 193. 217-250

Pang B. & Lee L. & Vaithyanathan S. 2002. Thumbs up? Sentiment Classification Using Machine Learning Techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*. 79–86. 10.3115/1118693.1118704.

Patwardhan S., Banerjee S., Pedersen T. 2003. Using Measures of Semantic Relatedness for Word Sense Disambiguation. In: *Gelbukh A. (eds) Computational Linguistics and Intelligent Text Processing. CICLing 2003. Lecture Notes in Computer Science*, vol 2588. Springer, Berlin, Heidelberg. 241-257. https://doi.org/10.1007/3-540-36456-0_24

Yarowsky D. 1994. Decision lists for lexical ambiguity resolution: application to accent restoration in Spanish and French. In *Proceedings of ACL '94*. 88-95

Апресян Ю. Д., Богуславский И. М., Иомдин Л. Л. и др. 1981. Лингвистическое обеспечение в системе автоматического перевода ЭТАП-1. *Разработка формальной модели естественного языка*. Новосибирск, 3-28.

Апресян Ю. Д., Богуславский И. М., Иомдин Л. Л., Лазурский А. В., Перцов Н. В., Санников В. З., Цинман Л. Л. 1989. Лингвистическое обеспечение системы ЭТАП-2. М.: Наука.

Кибрик А. Е. и др. 2019. Введение в науку о языке. Раздел 5: Прикладная и компьютерная лингвистика. М.: Буки-Веди. 455-534.

Корочков А. В. 2006. О количественной оценке адекватности лингвистических правил (на материале правил чтения для английского языка). *Вопросы Языкознания*. №5, 78-91.

Кузнецов И. П., Сомин Н. В. 2014. Методы автоматического извлечения из текстов семантически значимой информации. М.: ИПИ РАН.

Леонтьева Н. Н. 2006. Автоматическое понимание текстов: системы, модели, ресурсы. М.: Издательский центр "Академия".

Рекомендуемый перечень иных информационных источников

1. Портал материалов по машинному обучению machinelearning.ru
2. https://intuit.ru/studies/courses?service=0&option_id=17&service_path=1
3. http://wordnetweb.princeton.edu/perl/webwn
4. http://project.phil.spbu.ru/RussNet/index_ru.shtml
5. http://www.labinform.ru/pub/ruthes/
6. http://www.ksl.stanford.edu/software/ontolingua/
7. http://www.daml.org/ontologies/
8. http://www.jfsowa.com/ontology/
9. https://ruwordnet.ru/ru
10. https://www.dbpedia.org/
11. https://rusvectores.org/ru/
12. https://framenet.icsi.berkeley.edu/fndrupal/
13. http://openie.allenai.org/

Раздел 4. Разработчики программы

Лукашевич Наталья Валентиновна, доктор технических наук, профессор.
Рабочая группа ФУМО 45.00.00 по проблемам искусственного интеллекта в языкознании и литературоведении.