**Аннотация**

В данной работе решается задача извлечения метаинформации из документов с научным уклоном, а также реализация клиентского приложения, которое извлекает метаданные из текста и дает возможность опубликовать их в систему управления пространственной информацией – GeoNetwork.

**Введение**

Извлечение метаописаний – это задача автоматического извлечения структурированных данных из источников с неструктурированной или слабо структурированной информацией, которая связана с обработкой данных на естественных языках.

Извлечение метаописаний широко применяется в задачи информационного поиска. Эта процедура помогает более эффективно искать информацию, удовлетворяющую критериям поиска.

В общем случае задача состоит в том, чтобы из предложенного текста, выделить информацию об именованных объектах в тексте (информацию о классах, которые они соответствуют) и выделить ключевые слова. Набор классов фиксируется заранее.

Конкретно в моей задаче – нужно извлечь информацию из текста, предоставить интерфейс, где можно её отредактировать и дополнить, а также опубликовать в систему GeoNetwork.

Одной из подзадач в извлечении метаописаний – это выделение именованных сущностей (Named Entity Recognition, NER). Цель этой задачи выделить в тексте объектов определенного типа, это могут быть имена людей, географические объекты, биологические объекты и т.д.

Эта задача начала активно развиваться в середине 90-х, но и сейчас она остается актуальной, так как с каждым днем появляется потребность классификации, данный с новыми классами.

Под термином именованная сущность (named entity) понимается объект, который имеет такие характеристики как имя или идентификатор.

Сейчас существует много разных стратегий и подходов к решению поставленной задачи. В основном это системы, основанные на словарях, правилах или машинном обучении. Наиболее популярны методы машинного обучения с учителем. Такая система изначально тренируется на множестве примеров. Такую систему можно использовать на произвольных данных. Она имеет преимущество, по сравнению с моделями, основанными на словарях, так как может распознавать данные, которые раньше не встречала. Тем не менее, такая система тоже может ошибаться.

Выделение ключевых слов и словосочетаний

Ещё одна подзадача извлечения данных из текста. Ключевые слова – важные слова и словосочетания, которые дают высокоуровневое описание текста. Выделение таких слов, сделает информационный поиск более эффективным и точным.

Сейчас существуют алгоритмы, основанные на правилах и машинном обучении. Задача сводится к фильтрации текста и далее производится извлечение ключевых слов.

Связывание именованных сущностей с пространственным местоположением.

После извлечения географических объектов, хотелось бы получить информацию о местоположении объекта. Данную задачу можно решить с помощью словаря объектов и их координат и алгоритма неточного сравнения строк.

Эта подзадача не совсем связана с извлечением информации, но она так же актуальна, так как в системе GeoNetwork поиск может осуществляться по пространственным координатам.

Все перечисленное делает задачу извлечения метаописаний достаточно полезной и интересной не только с точки зрения исследователей, но и с точки зрения автоматизации процесса.

**Обзор существующих решений**

На просторах сети интернет существует множество систем, которые так или иначе решают задачу извлечения информации из неструктурированного или слабо структурированного текста.

Приведу некоторые примеры.

**Cremine** – это web сервис с открытым исходным кодом. У него неплохой интерфейс, но на предложенных ему документах он справился не самым лучшим образом.

Этот сервис использует машинное обучение, а именно CRF, SVM и K-means.

У него модульная структура, где для каждой подзадачи используются свои алгоритмы.

**GROBID** – сервис, который предлагается инсталлировать на свою машину и использовать.

Он имеет открытый исходный код и использует SVM для извлечения информации.

**PDF-Tools** – это платная десктопная программа для извлечения информации из PDF файлов.

**Text Razor** сервис для извлечения именованных сущностей из текста.

Несмотря на то, что представленные решения справляются с задачей извлечения информации, в той или иной степени, не одно из них не предоставляет данные в нужном формате, для публикации в GeoNetwork. Также в них нет специфических данных, таких как координаты.

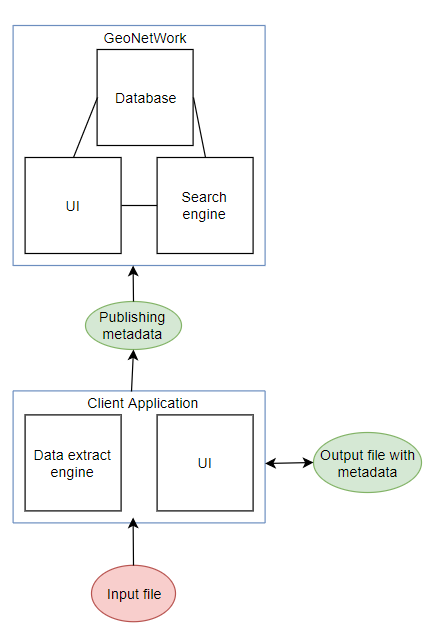
Следовательно, возникает необходимость разработать свою систему для извлечения метаописаний.

**Постановка задачи**

Задача данной работы состоит в разработке программного продукта для извлечения метаописаний из документов с последующим редактированием и публикацией в системе GeoNetwork. Для этого требуется:

1. Исследовать существующие методы извлечения информации из текстов;
2. Разработать методы извлечения информации из текстов;
3. Выполнить программную реализацию разработанных методов;
4. Реализовать клиентское приложение для системы GeoNetwork, в которое будет загружаться текстовый документ и извлекать метаданные из него. Также должна быть возможность отредактировать и дополнить эти данные, а затем опубликовать их.
5. Провести тестирование разработанной системы.

Примерная архитектура системы:



**Описание метода**

**Извлечение именованных сущностей**

Итак, мы имеем неструктурированный текст, и наша задача получить теги для слов из этого текста (I-PER, O, B-LOC и т. д.).

Пример:

Inna V. Stonik lives in Vladivostok.

I-PER I-PER I-PER O O B-LOC

В качестве обучающего корпуса был взят CoNLL2003, в котором имеются следующие виды именованных сущностей: места, имена людей, организации, разное.

Данные для обучения представлены в следующем виде:

-DOCSTART- -X- O O

Phil NNP I-NP I-PER

Simmons NNP I-NP I-PER

took VBD I-VP O

four CD I-NP O

for IN I-PP O

38 CD I-NP O

on IN I-PP O

Friday NNP I-NP O

as IN I-PP O

Leicestershire NNP I-NP I-ORG

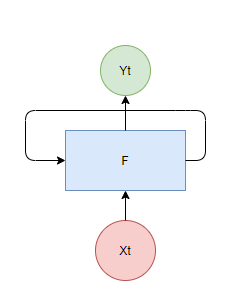
beat VBD I-VP O

Somerset NNP I-NP I-ORG

…

Проблема в задачи определения тега слова в том, что нельзя рассматривать каждое слово в отдельности. Например, Russian Federation – имеет тег ‘место’ и состоит из двух слов. Следовательно, нужно, чтобы модель имела память, и при встрече слова ‘Federation’ понимала, что ‘Russian’ тоже является местом. С этой задачей справляются рекуррентные нейронные сети.

Рекуррентные нейронные сети - сети, содержащие обратные связи и позволяющие сохранять информацию. Наличие этой связи дает возможность передавать информацию от одного шага сети другому.



Здесь сеть принимает Xt и возвращает Yt.

Конкретно в используемой модели используется модификация РНС – LSTM.

Долгая краткосрочная память или Long short-term memory (LSTM) – разновидность рекуррентных нейронных сетей, способная к обучению долговременным зависимостям.

Её особенность в том, что что она может хранить состояния. В ходе работы LSTM её состояние может меняется.

Нейронная сеть использует для обучения числовые данные, поэтому нужно слова преобразовать в вектора. Для этого используется алгоритм GloVe.

Итак, каждое слово нужно преобразовать в вектор, где каждое такое слово – конкатенация из векторного представления GloVe и вектора, определённого из символов слова. (Буквы помогут понять, например, что слово начинается с заглавной буквы.)

Далее запускаем bi-LSTM (LSTM с двумя скрытыми слоями) на полученном пространстве векторов и получаем новое пространство векторов .

Теперь, каждое слово ассоциируется с вектором .

Чтобы сделать предсказание, выполним:

, где каждая компонента вектора – это вероятность в пользу того, что слово относится к типу .

Теперь для окончательного предсказания нужно оценить вероятность последовательности слов.

Для этого используется алгоритм условно случайных полей (CRF).

Нужно оценить

– матрица перехода, – вектора оценок, которые фиксируют стоимость начала или конца заданного тега.

Таким образом, вектора тегов между оцениваются так, чтобы была максимальна.

В итоге общая вероятность определяется так: , где – нормирующий коэффициент .

**Извлечение библиографических ссылок**

Так как библиографическая ссылка имеет стандартизированный формат решено извлекать их с помощью регулярных выражений.

Каждую библиографическую ссылку можно разбить на составляющие:

список имен, заголовок, издание, год, количество страниц, серия.

Таким образом, можно составить регулярное выражение для каждой части, и их конкатенация будет давать полное регулярное выражение.

Полученные выражения:

name = "[A-Z]+[A-Za-z\\-]+,([\\s\*)([A-Za-z\\.]\*)](file:///\\s*)([A-Za-z\\.%5d*))"

names = "(" + self.regexName + "(\\s\*)[and,]\*(\\s\*)){1,}"

eds = "(\\(Ed\\.\\))|(\\(Eds\\.\\))"

names+eds = names + "(\\s\*)" + "("+eds+")\*"

year = "[0-9]\*\\.";

names+eds+year = names + "(\\s\*)" + "("+eds+")\*" + "(\\s\*)" +"("+year+")\*"

title+pages = "[A-Za-z0-9/\\.,\\s();\\-:]{1,300}[0-9]+(([\\-:][0-9]+\\.)|((\\s\*)pp\\.))"

reference = names+eds+year + "(\\s\*)" + title+pages

**Ссылки**

1. VanderPlas, J. Python Data Science Handbook. — 1-е изд., 2016.— 529 с.
2. Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. Пособие — М.: МИЭМ, 2011. — 272 с.
3. Klinger, R., Friedrich, C.M. Feature Subset Selection in Conditional Random Fields for Named Entity Recognition / Klinger, R, Friedrich, Ch. // Fraunhofer Institute for Algorithms and Scientiﬁc Computing (SCAI). — 2009. — С. 185–191.
4. Лабутин, И.А., Фирсов, А.Н., Чуприна, С.И. Распознование именованных сущностей в текстах на естественном языке с использованием метода пробросо-цепочных условных случайных полей / Лабутин, И.А., Фирсов, А.Н., Чуприна // Пермский государственный национальный исследовательский университет. – 8 с.
5. McCallum, A., Li, W. Named Entity Recognition with Conditional Random Fields, Feature Inductionand Web-Enhanced Lexicons / McCallum, A., Li, W. // University of Massachusetts Amherst. – 4 c.
6. Антонова, А.Ю., Соловьев, А.Н. Метод условно случайных полей в задачах обработки русскоязычных текстов / Антонова, А.Ю., Соловьев, А.Н. – С. 321-325.
7. Zhu, X Advanced NLP. Conditional Random Fields. 2007. - 6 с.
8. Колмогорцев, С.В., Сараев, П.В. Извлечение библиографии из текстов регулярными выражениями // [Новые информационные технологии в автоматизированных системах](https://cyberleninka.ru/journal/n/novye-informatsionnye-tehnologii-v-avtomatizirovannyh-sistemah), 2017.