Московский государственный университет имени
М. В. Ломоносова Факультет Вычислительной Математики и Кибернетики

Кафедра Алгоритмических Языков

ОТЧЁТ СТУДЕНТОВ 524 ГРУППЫ ПО КУРСУ

«МАТЕМАТИЧЕСКИЕ МОДЕЛИ

АВТОМАТИЧЕСКОГО АНАЛИЗА ТЕКСТОВ»

Извлечение именованных сущностей для текстов на русском языке

Выполнили:

студенты 5 курса 524 группы

Тодуа Антон Романович

Можарова Валерия Александровна

Лукъяненко Светлана Юрьевна

Преподаватель:

Dr Мстислав Масленников

Москва

15 мая 2016 г.

1. Введение

В современном мире задача обработки текстов на естественном языке очень востребована, так как она применяется в информационном поиске, аннотации документов, бизнес-аналитике и так далее. Ее можно разбить на несколько подзадач. Первая – нахождение значимых сущностей в тексте (именованных сущностей). Это могут быть названия организаций, имена людей, и названия географических объектов. Второе – определение отношений между сущностями. Например, должность сотрудника или звание военнослужащего. Третье - выявление событий, связанных с выделенными сущностями. Например, покупка акций, слияние компаний или встреча глав государств.

В данной работе подробно рассматривается первая задача – извлечение именованных сущностей. Работ по этой теме для русского языка очень мало по сравнению с английским или чешским языками, поэтому для нашего языка этот вопрос является актуальным до сих пор.

Именованная сущность – это слово или словосочетание, которое обозначает объект или явление, а также выделяющее этот объект среди прочих схожих объектов. Обычно именованная сущность начинается с большой буквы. Важным признаком именованной сущности является то, что она всегда имеет референта. В данной работе мы рассматриваем три вида именованных сущностей: персоны (PER), организации (ORG), географические объекты (LOC).

Для извлечения именованных сущностей существуют 3 основных метода. Первый основывается на машинном обучении. Для обучения нужна большая размеченная экспертами текстовая коллекция. Второй подход основывается на применении лингвистических правил, которые пишут эксперты. Третий подход – это комбинирование первых двух методов. Сейчас он применяется наиболее часто.

2. Метод и признаки

За основу системы наша команда выбрала CRF-классификатор, так как в ряде работ, посвященных извлечению именованных сущностей на других языках [1, 2, 3, 4], используется именно он. CRF-модель создана специально для

классификации последовательных данных, что играет важную роль в решении нашей задачи. Мы использовали готовую реализацию этого метода машинного обучения - CRF++[1], потому что она имеет открытый код и достаточно быстро обучается на наших данных.

2.1 Предобработка

Перед тем как извлекать признаки для классификатора, наша система прогоняет текст через морфологический анализатор mystem[2], чтобы было легче в дальнейшем создавать признаки. Анализатор делит текст на токены и определяет их морфологические характеристики.

2.2 Признаки

Первый вид признаков, которые мы использовали – это признаки токена:

- Токен
- Лемма
- Регистр букв (BigBig/BigSmall/SmallSmall/Fence)
- Бинарный признак: наличие гласной
- Диапазон длинны токена (1/2-4/5-more)
- Часть речи для слов, для пунктуации — вид знака
- Является ли токен концом предложения (бинарный)

Второй вид признаков – признаки, основанные на словарях. Мы вручную создали ряд словарей (имена, фамилии, столицы и т. д.), а затем для каждого токена в тексте система смотрела его вхождение в тот или иной словарь и проставляла соответствующее значение признака (True/False):

- Список имен
- Список фамилий
- Список отчеств
- Список столиц
- Список государств
- Список названий валют
- Список слов, предшествующих локациям

[1] https://taku910.github.io/crfpp/
[2] https://tech.yandex.ru/mystem/

- Список слов, предшествующих организациям

Следующий вид признаков похож на предыдущий вид признаков, только проверялось не вхождение всего токена в словарь, а лишь его части:

- Список окончаний фамилий
- Список корней организаций

По завершении этого этапа для каждого токена создается одинаковое количество признаков, и все токены с их признаками подаются на вход классификатору, который и проставляет каждому токену соответствующий ему класс именованной сущности (PER, LOC, ORG).

2.3 Сборка именованных сущностей

После того, как классификатор разметил токены, нужно их собрать в цельные именованные сущности. Мы делали это с помощью простых правил:

1. Подряд идущие токены одного типа склеиваются в одну именованную сущность
2. Если между этими токенами встретился знак препинания, то именованная сущность разбивается на несколько. Исключением из этого правила является только точка, не означающая конец предложения (например, «А. В. Сидоров»)
3. Если встретился шаблон [«ORG» имени/им. «PER»], то все токены, вошедшие в шаблон, объединяются в одну именованную сущность типа «ORG»

3.Текстовая коллекция

Для обучения и тестирования мы взяли открытую текстовую коллекцию с соревнования «Dialog: FactRuEval»[3]. В этом корпусе размечены три типа именованных сущностей: персоны, организации и географические объекты.

[3] https://github.com/dialogue-evaluation/factRuEval-2016

Обучающая выборка состоит из 122 новостных документов, а тестовая выборка состоит из 132 документов.

4. Результаты

В качестве целевой метрики мы выбрали метрику F-score, которая является сочетанием точности и полноты. Для ее вычисления на тестовой выборке мы использовали готовый компаратор, который был предложен организаторами соревнования. При комбинировании всех признаков мы получили следующие результаты:

Тип ИС	F-score
PER	0.8044
ORG	0.6198
LOC	0.8393
COMMON	0.7446

5. Заключение

В данной работе мы рассмотрели самые базовые признаки для метода машинного обучения в решении задачи извлечения именованных сущностей. Наша команда рассмотрела такие признаки, как признаки токена и признаки, основанные на словарях. Мы исследовали их влияние на распознавание ИС для текстов на русском языке. Для дальнейшего улучшения F-score нужно увеличивать словари, а также можно попробовать кластеризацию текстов [5, 6, 7] и двухэтапный проход [8, 9, 10].

Литература

1. Marcinczuk M., Stanek M, Piasecki M., Musial A.: Rich Set of Features for Proper Name Recognition in Polish Texts. In: International Joint Conferences, SIIS 2011. pp.332 - 344. Springer Berlin Heidelberg (2012)

2. Antonova A.Y., Soloviev A.N.: Conditional random field models for the processing of Russian. In: International Conference "Dialog 2013", pp. 27 - 44. RGGU (2013)

3.Podobryaev A.V.: Persons recognition using CRF model. In:15th All-Russian Scientific Conference "Digital Libraries: Advanced Methods and Technologies, Digital Collection", RCDL-2013, pp. 255 - 258. Demidov Yaroslavl State University (2013)

4. Gareev R., Tkachenko M., Solovyev V., Simanovsky A., Ivanov V.: Introducing baselines for Russian named entity recognition. In: 14th International Conference, CICLing 2013, pp. 329 - 342. Springer Berlin Heidelberg (2013)

5. Brown P.F., Della Pietra, V.J. Desouza P.V., Lai, J.C., Mercer, R.L.: Class-based n-gram models of natural language. Computational Linguistics. V. 18, 4, pp. 467 - 479 (1992)

6. Chrupala G.: Eficient induction of probabilistic word classes with LDA. In: 5th International Joint Conference on Natural Language Processing, IJCNLP 2011, pp. 363 - 372. Asian Federation of Natural Language Processing (2011)

7. Clark A.: Combining distributional and morphological information for part of speech induction. In: 10th Conference on European Chapter of the Association for Computational Linguistics, EACL 2003, v. 1, pp. 59 - 66. ACL (2003)

8. Finkel J. R., Grenager T., Manning, C.: Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In: 43nd Annual Meeting of the Association for Computational Linguistics, pp. 363 - 370. ACL (2005)

9. Ratinov L., Roth D.: Design challenge sand misconceptions in named entity recognition. In: 13th Conference on Computational Natural Language Learning, CoNLL, pp. 147 - 155. ACL (2009)

10. Strakova J., Straka M., Hajic J.: A New State-Of-The-Art. Czech Named Entity Recognizer. In: 16th International Conference, TSD 2013. pp. 68 - 75. Springer Berlin Heidelberg (2013)