Text mining

Извлечение информации и классификация последовательностей

Екатерина Черняк

echernyak@hse.ru

Национальный Исследовательский Университет – Высшая Школа Экономики
НУЛ Интеллектуальных систем и структурного анализа

December 15, 2017

1. Извлечение информации
2. Извлечение информации по словарям, регулярными выражениями и по шаблонам
3. Классификация чанков [chunk]
4. Условные случайные поля
5. Задача классификации последовательности
   - Условные случайные поля
   - Реккурентные нейронные сети
   - Заполнение слотов

Извлечение информации [Information extraction, IE]

Извлечение значимых элементов текста:

- Даты, номера телефонов, адреса
- Именованные сущности [Named entity recognition, NER]
- Отношения, факты и события
- Термины
- Разрешение кореференции [Coreference resolution]
- OpenIE (не сейчас)

Методы извлечения информации

- Словари, регулярные выражения и шаблоны
- Классификация чанков [chunk]
- Классификация последовательностей [sequence labelling]
- Заполнение слотов [slot filling]

Словари

- Wikipedia / DBPedia и другие Wiki
- Словари имен, терминов
- API карт (например, API KudaGo)
- Государственные реестры

Регулярные выражения

Номер телефона

```
re.compile('[+0-9 \- \( \)]8,')
```

email

```
r"(^[a-zA-Z0-9_.+-]+[a-zA-Z0-9-]+\.[a-zA-Z0-9-.]+$)"
```

Римские числа

```
^ M{0,4}(CM|CD|D?C0,3)(XC|XL|L?X{0,3})(IX|IV|V?I{0,3})$
```

Лексические шаблоны

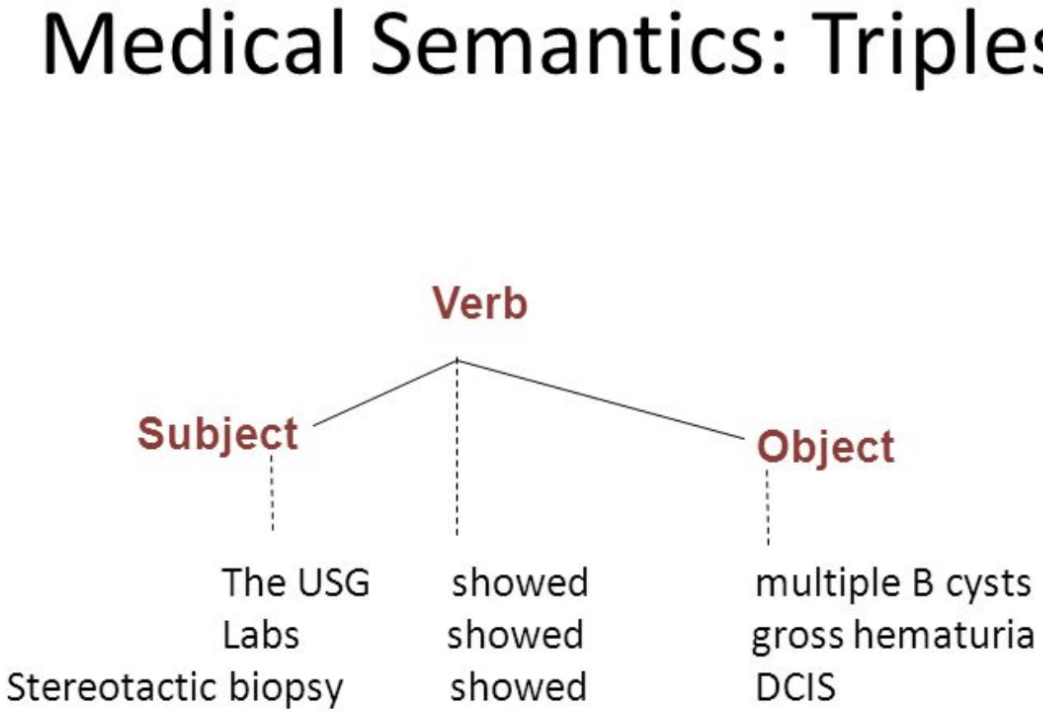
Шаблоны M. Hearst [Hearts, 1992]:

1. NP such as NP,* (and|or) NP
2. such NP as NP,* (and|or) NP
3. NP ,NP* , (and|or) other NP
4. NP including NP,* NP (and|or) NP
5. NP especially NP,* (and|or) NP

- Such injuries as bruises, wounds and broken bones
- Мало обратимые длительные психические расстройства, такие как нарушения или задержка умственного развития

Легко описать с помощью Томита-парсера.

SOV-триплеты: subject-of object-of verb



Источник: `http://slideplayer.com/slide/11723219/`

Легко описать в грамматике универсальных зависимостей.

Классификация чанков [chunk]

Чанк [chunk] – последовательность из нескольких токенов, иногда – синтаксическая группа
Именная группа [noun phrase, NP] – Adj* N*
Глагольная группа [verb phrase, VP] – Adv* V
Классификация чанков: данный чанк –

- название города?
- имя человека?
- название компании?
- название группы?

Можно использовать любой бинарный классификатор. Признаки:

- Морфологические тэги
- Регистр
- Есть ли одноименная статья в Википедии
- Слова в левом и правом окнах

Извлечение отношений [relation extraction]

- Родился-В: **кто** родился **где**?
- Супруги: **кто** женат / замужем **на ком** / **за кем**?
- CEO: **кто** руководит какой **компанией**?
- Сделка: **кто** заключил сделку **с кем**?
- Образование: **образование** обнаружено в **каком органе**?
- Побочный эффект: **препарат** вызвал **побочный эффект**?
- Мутация: **мутация** найдена **в каком гене**?

Задача классификации: дано два слова / чанка, c_1, c_2.

- простая задача бинарной классификации: находятся ли они в отношении $R(c_1, c_2)$?
- задача классификации на несколько классов: в каком отношении R находятся (c_1, c_2)?

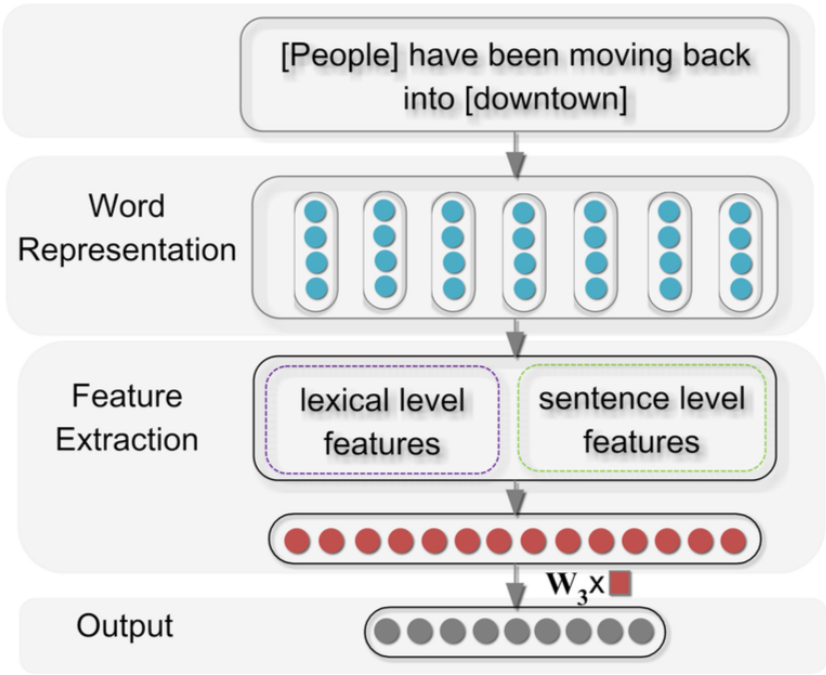
SemEval 2010, Task 8

The most common audits were about waste and recycling .	Message-Topic(e1,e2)	3	6
The company fabricates plastic chairs .	Product-Producer(e2,e1)	1	4
The school master teaches the lesson with a stick .	Instrument-Agency(e2,e1)	2	8

Методы классификации в задаче извлечения отношений

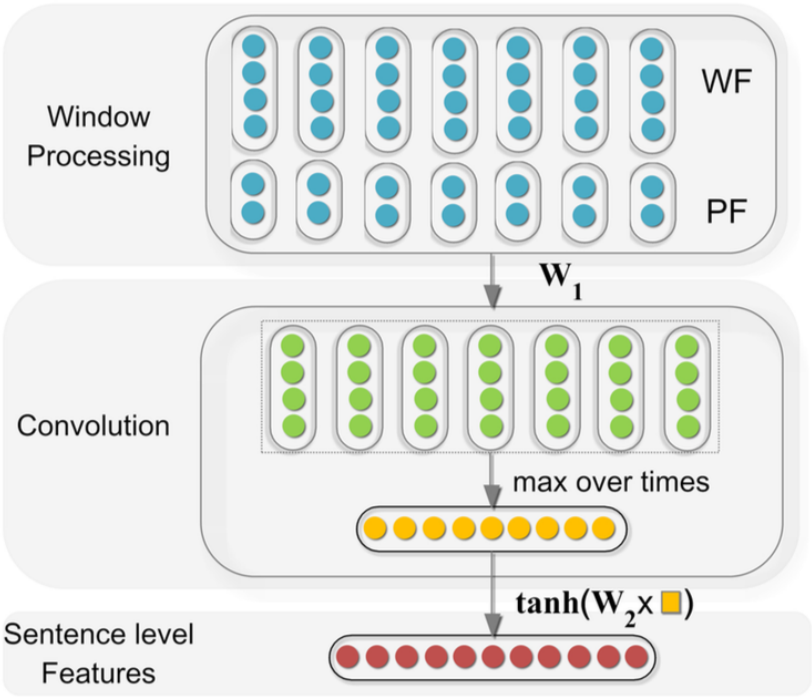
- Любой метод классификации и множество лингвистических признаков: POS-тэги, длина синтаксического пути между чанками, расстояние в WordNet и др.
- CNN и positional embeddings [Zeng et al., 2014, Relation Classification via Convolutional Deep Neural Network]
- Att-BiLSTM и index embeddings [Zhou et al., 2016, Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification]

CNN и positional embeddings [Zeng et al., 2014]



- Вход: слова и word embeddings
- Извлечение признаков: признаки на уровне слова и на уровне предложения
- Выход: число нейронов = количество различных отношение

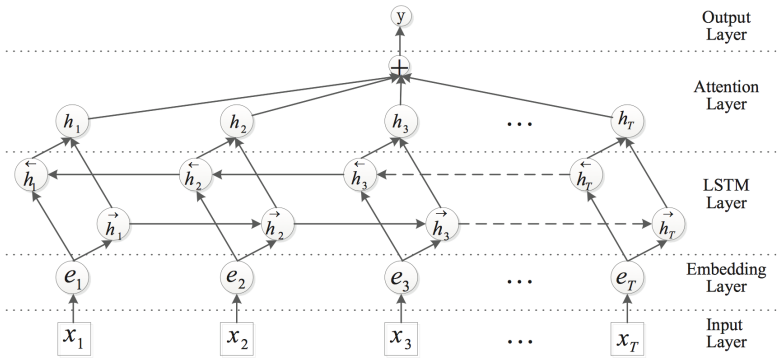
CNN и positional embeddings [Zeng et al., 2014]



Postion embeddings состоят из двух частей: расстояние от текущего слова до e_1 и до e_2. Каждому числу в соответствие ставится свой вектор. Итоговый вектор слова состоит из трех частей: эмбеддинг слова, вектор, кодирующий расстояние до e_1 и вектор, кодирующий расстояние до e_2.

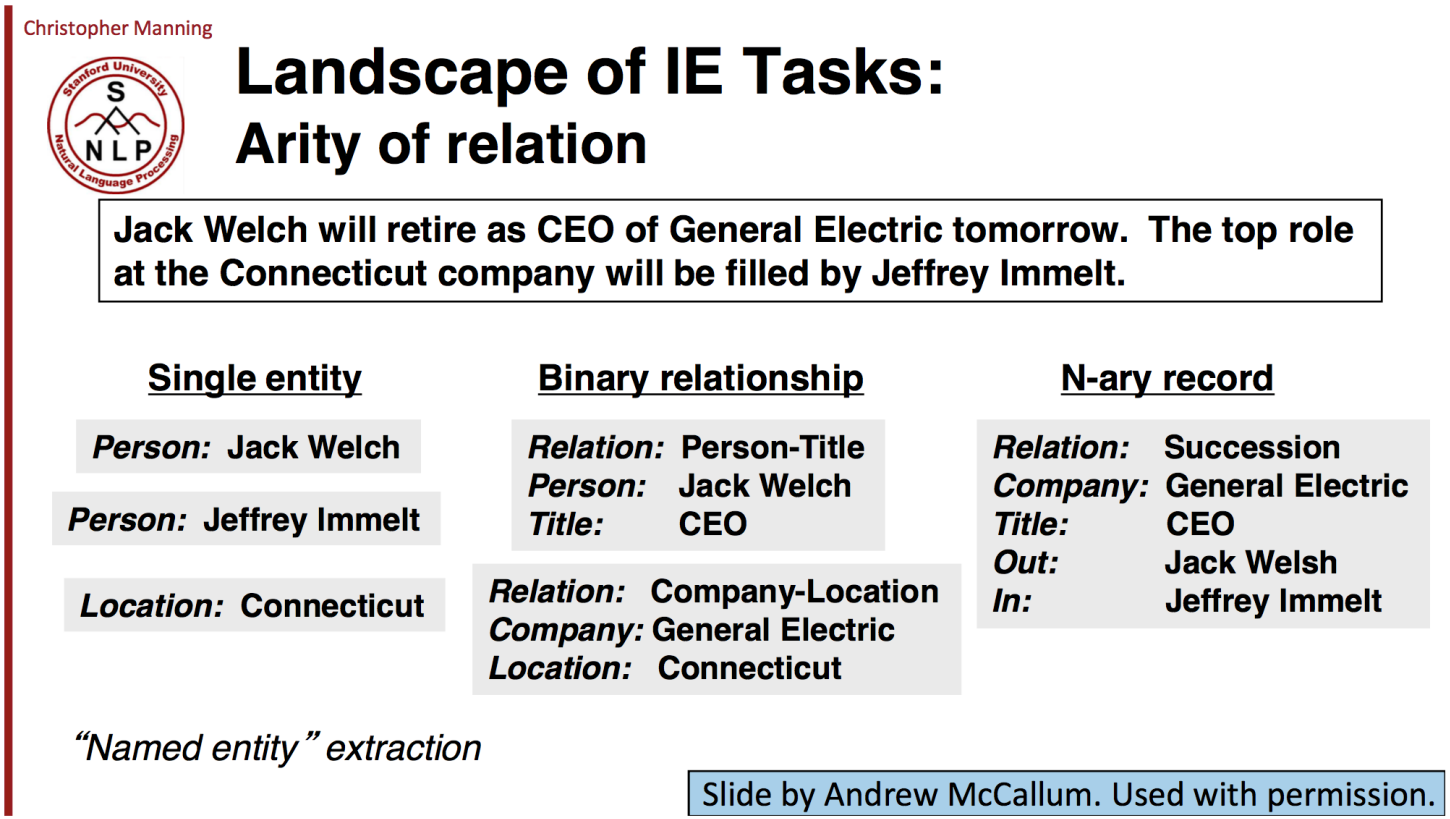
f

Att-BiLSTM и index embeddings [Zhou et al., 2016]



Аналогичная идея: вектор слова состоит из эмбеддинга слова и one-hot кодирования его индекса в словаре.

Извлечение фактов и событий [facts and events extraction]



Slide by Christopher Manning. Used without permission.

Факты и события: N – арные отношения

Факты и события

- Automatic Content Extraction (ACE)
  - Тип события (“смерть”)
  - Аргументы события (кто, где, когда, причина)
  - Триггеры (умер, погиб, убит, смерть, убийство)
  - Атрибуты события (модальность, поляризация, время, и др).
- FactRuEval
  - Тип факта (“должность”)
  - Аргументы факта (кто, должность, компания)

1. Извлечение информации
2. Извлечение информации по словарям, регулярными выражениями и по шаблонам
3. Классификация чанков [chunk]
4. Условные случайные поля
5. Задача классификации последовательности
   - Условные случайные поля
   - Реккурентные нейронные сети
   - Заполнение слотов

Задача классификации последовательности

	Британская	актриса	и	крестница	принца	Чарльза	Тара	Томкинсон
POS	Прил.	Сущ.	Союз	Сущ.	Сущ.	Им.Собств.	Им. Собств.	Им. Собств
IOB (NE)	O	O	O	O	O	B-Per	B-Per	I-Per
IOBES (NE)	O	O	O	O	O	S-Per	B-Per	E-Per
IOBES (R)	O	O	O	O	O	B-Per-1	B-Per-2	I-Per-2

	была	найдена	мертвой	в	ее	квартире	в	Лондоне
POS	Глаг.	Кр. Прич.	Прил.	Пред.	Мест.	Сущ.	Пред.	Им. Собств
IOB (NE)	O	O	O	O	O	O	O	B-Loc
IOBES (NE)	O	O	O	O	O	O	O	S-Loc
IOBES (R)	O	O	O	O	O	O	O	O

	,	сообщает	BBC	.
POS	Пункт.	Глаг.	Им. Собств	Пункт.
IOB (NE)	O	O	B-Org	O
IOBES (NE)	O	O	S-Org	O
IOBES (R)	O	O	O	O

Определение

Обучающие данные:

- $\boldsymbol{x} = x_1, x_2, \ldots, x_n$, $x_i \in V$, V – словарь
- $\boldsymbol{y} = y_1, y_2, \ldots, y_n$, $y_i \in \{1, \ldots, L\}$ – метки
- $\{(\boldsymbol{x}^{(1)}, \boldsymbol{y}^{(1)}), (\boldsymbol{x}^{(2)}, \boldsymbol{y}^{(2)}), \ldots, (\boldsymbol{x}^{(m)}, \boldsymbol{y}^{(m)})\}$ – обучающие данные
- экспоненциальная сложность: если длина входной последовательности $= n$, всего возможно L^n решний

Требуется обучить классификатор: $\boldsymbol{x} \to \boldsymbol{y}$

- y – последовательность
- y – дерево (парсинг)

Методы классификации последовательности

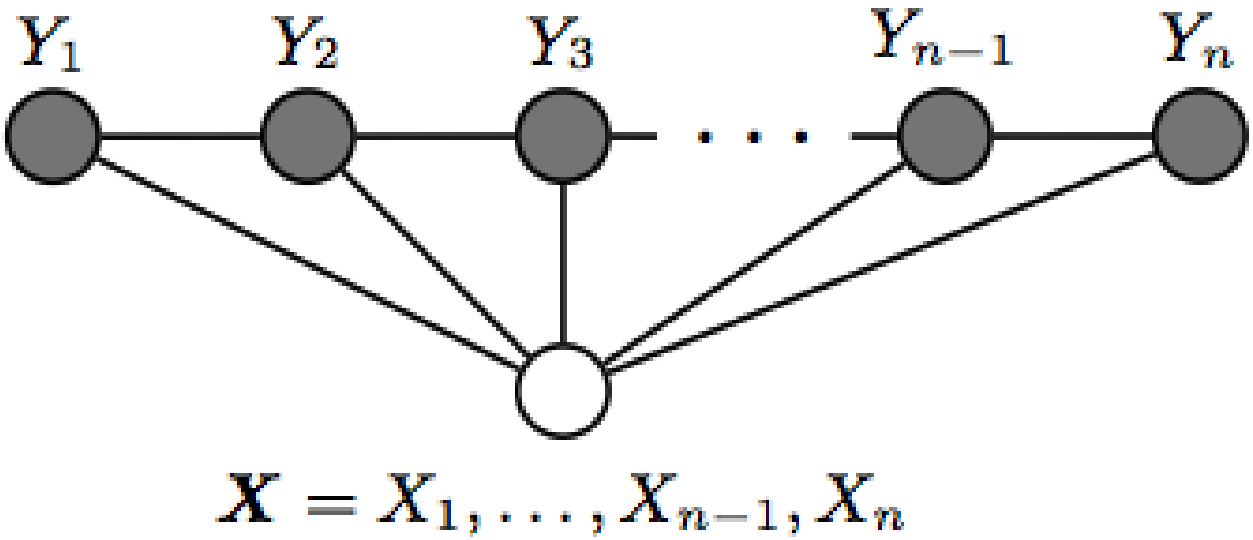
- Sequence labelling
  - Марковские модели максимальной энтропии [Maximum-entropy Markov model, MEMM]
  - Условные случайные поля [Conditional random fields, CRF]
  - Реккурентные нейронные сети (biLSTM)
  - (CNN-)biLSTM-CRF [Ma and Hovy, 2016 End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF]
- Structured prediction
  - SVM^{struct}
  - Structured perceptron
- Slot filing
  - biLSTM-CNN-CRF with attention [Liu and Lane, 2016 Attention-Based Recurrent Neural Network Models for Joint Intent Detection and Slot Filling]

1. Извлечение информации
2. Извлечение информации по словарям, регулярными выражениями и по шаблонам
3. Классификация чанков [chunk]
4. Условные случайные поля
5. Задача классификации последовательности
   - Условные случайные поля
   - Реккурентные нейронные сети
   - Заполнение слотов

Условные случайные поля [Lafferty, 2001]

Условные случайные поля [Conditional random fields]

$$\hat{Y} = \arg\max_{Y} P(Y|X) = \phi(y_i, y_{i-1})\phi(y_i, x_i)$$



Источник: http://davidsbatista.net/blog/2017/11/13/Conditional_Random_Fields/

Условные случайные поля

Вероятность последовательности меток классов для входной последовательности определяется по признакам, которые называются потенциальными функциями. Эти признаки помогают связать класс текущего наблюдения x_i с классами других наблюдений. Для формализации признаков чаще всего используются индикаторные функции. Таким образом, задача обучения сводится к определению весов индикаторных функций.

$$t(y_{i-1}, y_i, x, i) = \begin{cases} 1, & \textit{if } x_i = \text{“}\textit{June}\text{”} \textit{ and } y_{i-1} = \textit{IN} \textit{ and } y_i = \textit{NNP} \\ 0, & \textit{otherwise} \end{cases}$$

$$s(y_{i-1}, x, i) = \begin{cases} 1, & \textit{if } x_i = \text{“}\textit{to}\text{”} \textit{ and } y_i \\ 0, & \textit{otherwise} \end{cases}$$

Условные случайные поля

Для того, чтобы найти вероятность последовательности классов для входной последовательности:

1. извлекаем признаки
2. находим их веса и линейную комбинацию их признаков с найденными весами
3. используем softmax для определения искомых вероятностей.

Обозначим все признаки: $f(y_{i-1}, y_i, x, i)$. Признаки для последовательностей: $F(y, x) = \sum_{i=1}^{n} f(y_{i-1}, y_i, x, i)$. Обозначим веса признаков через λ. Искомая вероятность:

$$p(y|x) = \frac{e^{\sum_{i=1}^{k} \lambda_i F_i(y,x)}}{\sum_{y' \in C^n} e^{\sum_{i=1}^{k} \lambda_i F_i(y',x)}}$$

Пример. NER

Город/O Пушкин/(Per, City) является/O ...

Признаки:

$$f_1(y_{i-1}, y_i, x, i) = \begin{cases} 1, & \textit{if } x_i = \text{Пушкин } \textit{and } y_{i-1} = O \textit{ and } y_i = Per \\ 0, & \textit{otherwise} \end{cases}$$

$$f_1(y_{i-1}, y_i, x, i) = \begin{cases} 1, & \textit{if } x_i = \text{Пушкин } \textit{and } y_{i-1} = O \textit{ and } y_i = City \\ 0, & \textit{otherwise} \end{cases}$$

$$f_3(y_{i-1}, y_i, x, i) = \begin{cases} 1, & \textit{if } x_i = \text{Пушкин } \textit{and } x_{i+1}[:3] = \text{тся } \textit{and } y_i = Per \\ 0, & \textit{otherwise} \end{cases}$$

$$f_4(y_{i-1}, y_i, x, i) = \begin{cases} 1, & \textit{if } x_i = \text{Пушкин } \textit{and } \ x_{i+1}[:3] = \text{тся } \textit{and } y_i = City \\ 0, & \textit{otherwise} \end{cases}$$

$$f_3(y_{i-1}, y_i, x, i) = \begin{cases} 1, & \textit{if } x_i = \text{Пушкин } \textit{and } x_{i-1} = \text{город } \textit{and } y_i = Per \\ 0, & \textit{otherwise} \end{cases}$$

$$f_4(y_{i-1}, y_i, x, i) = \begin{cases} 1, & \textit{if } x_i = \text{Пушкин } \textit{and } x_{i-1} = \text{город } \textit{and } y_i = City \\ 0, & \textit{otherwise} \end{cases}$$

Веса: $\lambda_1 = 5, \lambda_2 = 2, \lambda_3 = 10, \lambda_4 = 7, \lambda_5 = 7, \lambda_6 = 20.$

Пример. NER

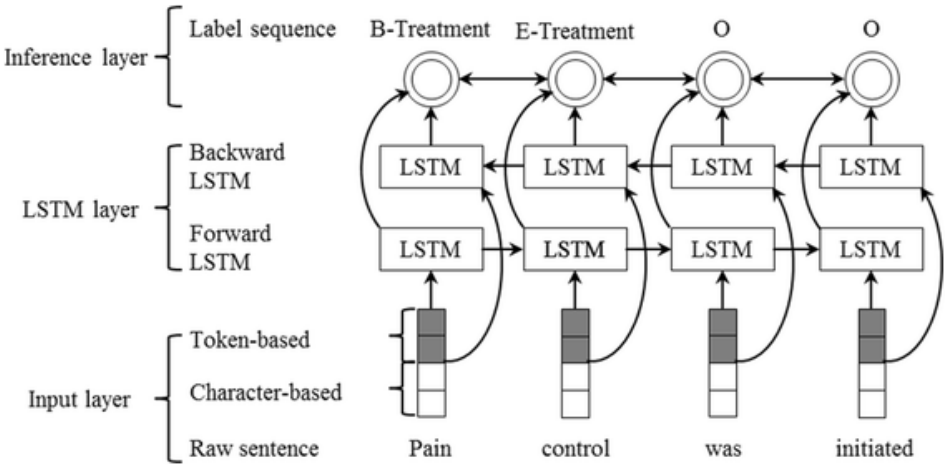
Сравним вероятности $P(\text{O Per O} \mid \text{Город Пушкин является})$ и $P(\text{O City O} \mid \text{Город Пушкин является})$.

1. O Per O : $5 + 10 + 7 = 12$
2. O City O : $2 + 7 + 20 = 27$

$$P(\text{O Per O} \mid \text{Город Пушкин является}) = \frac{e^{12}}{e^{12}+e^{27}}$$
$$P(\text{O City O} \mid \text{Город Пушкин является}) = \frac{e^{27}}{e^{12}+e^{27}}$$

1. Извлечение информации
2. Извлечение информации по словарям, регулярными выражениями и по шаблонам
3. Классификация чанков [chunk]
4. Условные случайные поля
5. Задача классификации последовательности
   - Условные случайные поля
   - Реккурентные нейронные сети
   - Заполнение слотов

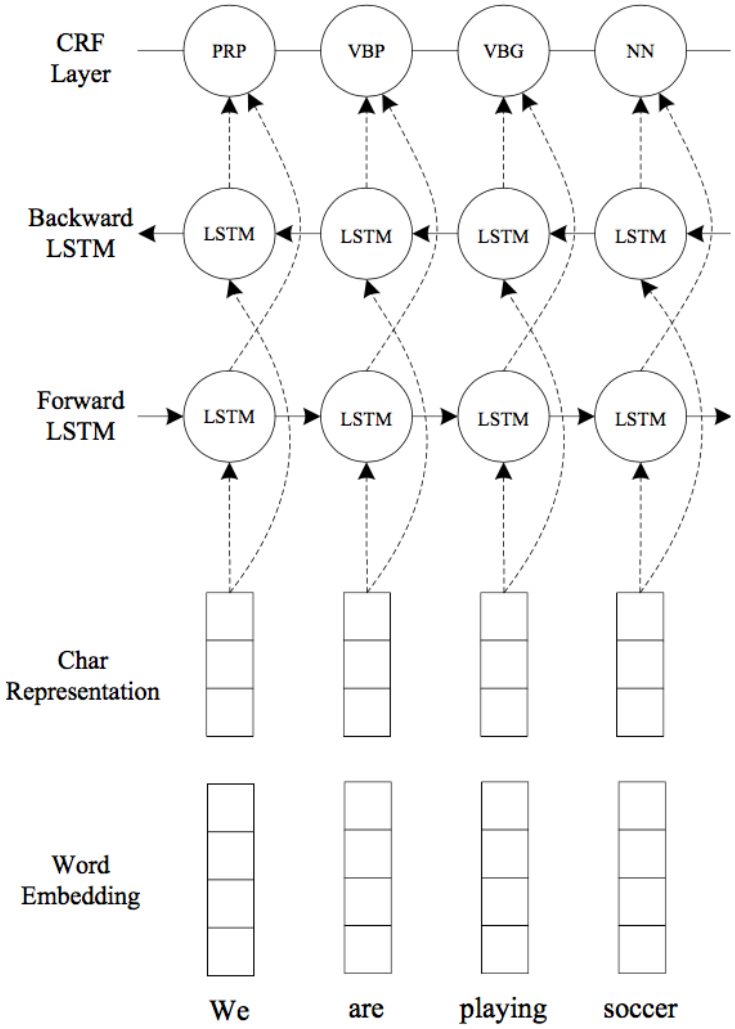
biLSTM для классификации последовательности



Каждый выход нейронной сети решает свою задачу классификации: какую метку приписать входному слову?

Источник: `https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-017-0468-7`

(CNN-)biLSTM-CRF [Ma and Hovy]



Слой с CRF используется для перевзвешивания выхода нейронной сети:

$$\text{score}_{lstm-crf}(x, y) =$$

$$= \sum_{i=0}^{n} W_{y_{i-1}, y_i} \cdot \text{LSTM}(x_i) + b(y_{i-1}, y_i)$$

1. Извлечение информации
2. Извлечение информации по словарям, регулярными выражениями и по шаблонам
3. Классификация чанков [chunk]
4. Условные случайные поля
5. Задача классификации последовательности
   - Условные случайные поля
   - Реккурентные нейронные сети
   - Заполнение слотов

Меры качества классификации последовательностей

1. token-based
   tp – число истинно-положительных токенов, fp – число ложно-положительных токенов, fn – число ложно-отрицательных токенов
2. chunk-based
   чанк – именованная сущность (синтаксическая группа, и др.) целиком
   tp – число истинно-положительных чанков, fp – число ложно-положительных чанков, fn – число ложно-отрицательных чанков