ОСНОВЫ ОБРАБОТКИ ТЕКСТОВ

Лекция №8
Разрешение кореферентности

Лектор: н.с. ИСП РАН Сысоев Андрей Анатольевич

7 ноября 2018

План

- Кореферентность vs анафоричность.
- Подходы к разрешению кореферентности:
  - модель "пара упоминаний" (mention-pair model);
  - модель "сущность - упоминание" (entity-mention model);
  - подходы на основе ранжирования (ranking models).
- Поиск упоминаний в тексте.
- Оценка качества.

Разрешение кореферентности

Упоминание - последовательность слов текста, обозначающая некоторую сущность. Выражение с выделенным главным словом.
Разрешение кореферентности: найти в тексте все упоминания, относящиеся к одной и той же сущности.

сущность = кореферентная цепочка = кластер упоминаний

Шариков злобно покосился на профессора, а тот отправил ему косой взгляд. Через десять минут Шариков уехал в цирк. Филлип Филлипович остался один в своем кабинете. Он начал шагами мерять комнату.

Разрешение кореферентности

Упоминание - последовательность слов текста, обозначающая некоторую сущность. Выражение с выделенным главным словом.
Разрешение кореферентности: найти в тексте все упоминания, относящиеся к одной и той же сущности.

сущность = кореферентная цепочка = кластер упоминаний

Шариков злобно покосился на профессора, а тот отправил ему косой взгляд. Через десять минут Шариков уехал в цирк. Филлип Филлипович остался один в своем кабинете. Он начал шагами мерять комнату.

Разрешение анафоры

Анафор - упоминание, которое соотносится с предшествующим упоминанием (**антецедентом**), заменяя его или указывая на него. Чаще всего в роли анафора выступает местоимение.

Шариков злобно покосился на профессора, а тот отправил ему косой взгляд. Через десять минут Шариков уехал в цирк. Филлип Филлипович остался один в своем кабинете. Он начал шагами мерять комнату.

Разрешение анафоры

Анафор - упоминание, которое соотносится с предшествующим упоминанием (**антецедентом**), заменяя его или указывая на него. Чаще всего в роли анафора выступает местоимение.

Шариков злобно покосился на профессора, а тот отправил ему косой взгляд. Через десять минут Шариков уехал в цирк. Филлип Филлипович остался один в своем кабинете. Он начал шагами мерять комнату.

Кореферентность vs анафоричность

— Кот, — сообразил Борменталь и выскочил за ним вслед.

...

— Где он?! Иван Арнольдович, успокойте, ради бога, пациентов в приемной!

Борменталь, **Иван Арнольдович**

Кореферентность vs анафоричность

— Кот, — сообразил Борменталь и выскочил за ним вслед.

...

— Где он?! Иван Арнольдович, успокойте, ради бога, пациентов в приемной!

Борменталь, **Иван Арнольдович**

- кореферентность есть
- анафоричности нет

Кореферентность vs анафоричность

Я вошел в квартиру: стены белые, потолки высокие.

Кореферентность vs анафоричность

Я вошел в квартиру: стены белые, потолки высокие.

Бриджинг – это вид анафоры, при котором анафорически связанные элементы не кореферентны.

Разрешение кореферентности

Найти в тексте все упоминания, относящиеся к одной и той же сущности.

План

- Кореферентность vs анафоричность.
- Подходы к разрешению кореферентности.
- Поиск упоминаний в тексте.
- Оценка качества.

Классификация

- модель "пара упоминаний" (mention-pair model)
- модель "сущность - упоминание" (entity-mention model)
- ранжирующие модели (ranking models)

V. Ng
Supervised noun phrase coreference research: The first fifteen years

Модель "пара упоминаний"

- Для каждой пары упоминаний определить, входят ли они в одну кореферентную цепочку.
- Использовать классифицированные пары упоминаний для построения финальных кореферентных цепочек.

W. M. Soon, H. T. Ng, D. C. Y. Lim
A Machine Learning Approach to Coreference Resolution of Noun Phrases

Модель "пара упоминаний"

Пример

Шариков злобно покосился на профессора, а тот отправил ему косой взгляд. Через десять минут Шариков уехал в цирк. Филлип Филлипович остался один в своем кабинете. Он начал шагами мерять комнату.

Модель "пара упоминаний"

Пример

Шариков злобно покосился на профессора, а тот отправил ему косой взгляд. Через десять минут Шариков уехал в цирк. Филлип Филлипович остался один в своем кабинете. Он начал шагами мерять комнату.

Модель "пара упоминаний"

Пример

Шариков злобно покосился на **профессора**, а тот отправил ему косой взгляд. Через десять минут Шариков уехал в цирк. Филлип Филлипович остался один в своем кабинете. Он начал шагами мерять комнату.

Модель "пара упоминаний"

Пример

Шариков злобно покосился на профессора, а **тот** отправил ему косой взгляд. Через десять минут Шариков уехал в цирк. Филлип Филлипович остался один в своем кабинете. Он начал шагами мерять комнату.

профессора $\longleftarrow$ тот

Модель "пара упоминаний"

Пример

Шариков злобно покосился на профессора, а тот отправил ему косой взгляд. Через десять минут Шариков уехал в цирк. Филлип Филлипович остался один в своем кабинете. Он начал шагами мерять комнату.

профессора $\longleftarrow$ тот

Шариков $\longleftarrow$ ему

Модель "пара упоминаний"

Пример

Шариков злобно покосился на профессора, а тот отправил ему косой взгляд. Через десять минут Шариков уехал в цирк. Филлип Филлипович остался один в своем кабинете. Он начал шагами мерять комнату.

профессора $\longleftarrow$ тот

Шариков $\longleftarrow$ ему

...

{Шариков, ему, Шариков}

{професора, тот, Филлип Филлипович, Он}

{своем кабинете, комнату}

Модель "пара упоминаний"

Подзадачи

- классификатор пар упоминаний
  - признаки
  - алгоритм машинного обучения
- отбор пар упоминаний
- объединение пар упоминаний в кореферентные цепочки

Классификатор пар упоминаний

Классификатор для каждой пары упоминаний должен выдать:

- метку: да / нет
- (опционально) уверенность

Классификатор пар упоминаний

Признаки

- **базовые** (для главных слов упоминаний и контекста):
  - токен
  - лемма
  - часть речи
  - граммемы (род, число, одушевленность)
- **тип именованной сущности**
- **согласованность**
  - род
  - число
  - одушевленность
  - оба упоминания - местоимения
  - тип именованной сущности

Классификатор пар упоминаний

Признаки

- **позиционные**: расстояние между упоминаниями в
  - токенах
  - упоминаниях
  - предложениях
  - параграфах
- **структурные**:
  - размер в токенах
  - взаимодействие с другими упоминаниями:
    - включение
    - вложенность

  его **кабинет**

Классификатор пар упоминаний

Признаки

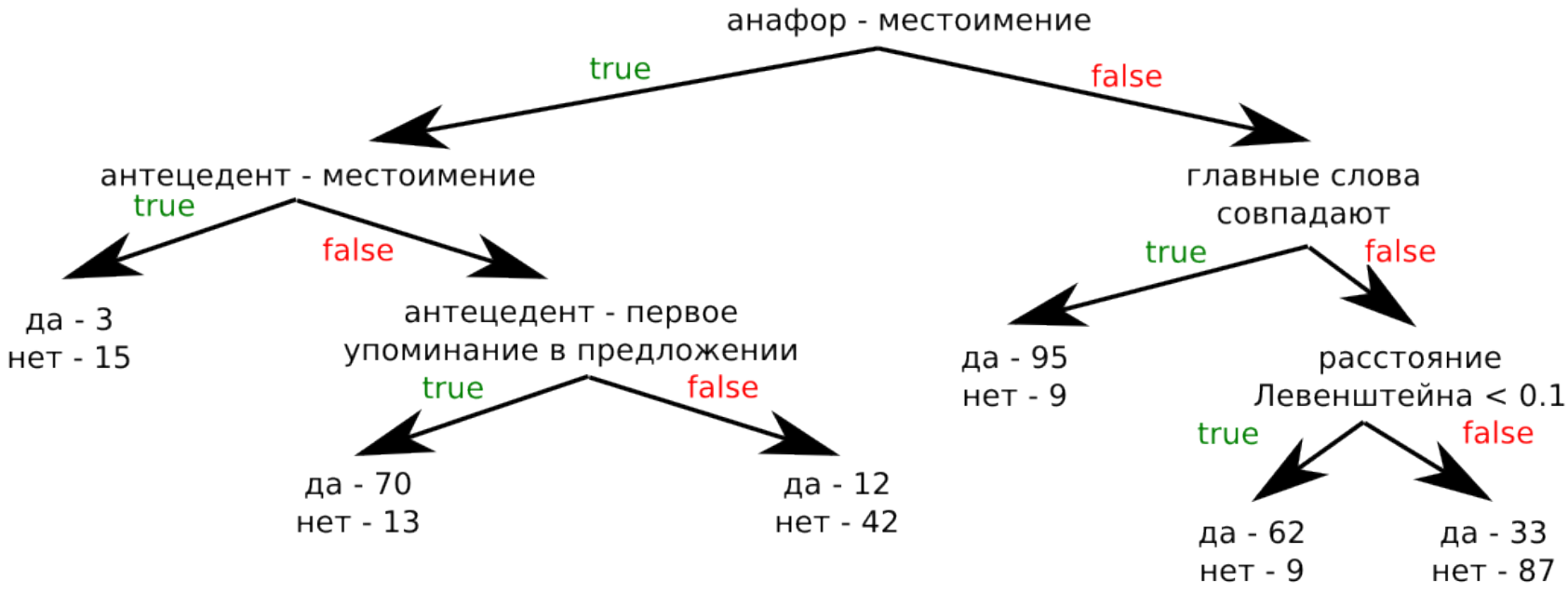
- **согласованность текстовых представлений**:
  - одно упоминание является частью другого:
    профессор *Преображенский* - ***профессор***
  - аббревиатуры:
    нэп - ***н****овая* ***э****кономическая* ***п****олитика*
    домком - ***дом****овой* ***ком****итет*
  - лексикографическая близость: по буквенным n-граммам, расстояние Левенштейна
  - совпадение главных слов упоминаний
- **синтаксические**:
  - роль в предложении
  - принадлежность одной клаузе (простое предложение в рамках сложного)
  - наличие общего родителя в синтаксическом дереве
- **семантические / синонимические**: косинус угла между векторами главных слов упоминаний

Классификатор пар упоминаний

Алгоритм машинного обучения

- SVM
- дерево решений
- случайный лес (random forest)

Дерево решений



Случайный лес

- строится набор из **N** деревьев решений
- при построении дерева / узла дерева выбор происходит не из всех признаков, а лишь из некоторого случайного подмножества

Построение составных признаков

метка	ант_пер	анаф_пер	лекс_близ >0.5	одуш	ед
да	F	F	T	F	F
нет	T	T	F	F	T
да	T	F	F	F	T
нет	F	F	F	T	F
да	T	F	T	F	T
нет	T	T	F	T	F
да	F	T	T	F	F

M. Segond, C. Borgelt
Item Set Mining Based on Cover Similarity

Построение составных признаков

метка	ант_пер	анаф_пер	лекс_близ >0.5	одуш	ед
да	F	F	T	F	F
нет	T	T	F	F	T
да	T	F	F	F	T
нет	F	F	F	T	F
да	T	F	T	F	T
нет	T	T	F	T	F
да	F	T	T	F	F

M. Segond, C. Borgelt
Item Set Mining Based on Cover Similarity

Построение составных признаков

метка	ант_пер	анаф_пер	лекс_близ >0.5	одуш	ед
да	F	F	T	F	F
нет	**T**	**T**	**F**	F	T
да	T	F	F	F	T
нет	F	F	F	T	F
да	T	F	T	F	T
нет	**T**	**T**	**F**	T	F
да	F	T	T	F	F

Можно добавить новый признак:

$$\textit{ант\_пер} \wedge \textit{анаф\_пер} \wedge \neg(\textit{лекс\_близ} > 0.5)$$

M. Segond, C. Borgelt
Item Set Mining Based on Cover Similarity

Модель "пара упоминаний"

Подзадачи

- классификатор пар упоминаний
  - признаки
  - алгоритм машинного обучения
- отбор пар упоминаний
- объединение пар упоминаний в кореферентные цепочки

Отбор пар упоминаний

На этапе обучения

Для каждого упоминания:

- ближайшее к нему слева упоминание из одной с ним кореферентной цепочки - положительный пример
- упоминания между ними - отрицательные примеры

Отбор пар упоминаний

На этапе обучения

Для каждого упоминания:

- ближайшее к нему слева упоминание из одной с ним кореферентной цепочки - положительный пример
- упоминания между ними - отрицательные примеры

Шариков злобно покосился на профессора, а тот отправил ему косой взгляд. Через десять минут Шариков уехал в цирк. **Филлип Филлипович** остался один в своем кабинете. Он начал шагами мерять комнату.

Отбор пар упоминаний

На этапе обучения

Для каждого упоминания:

- ближайшее к нему слева упоминание из одной с ним кореферентной цепочки - положительный пример
- упоминания между ними - отрицательные примеры

Шариков злобно покосился на профессора, а **тот** отправил **ему** косой взгляд. Через десять минут **Шариков** уехал в цирк. **Филлип Филлипович** остался один в своем кабинете. Он начал шагами мерять комнату.

- тот – Филлип Филлипович
- ему – Филлип Филлипович
- Шариков – Филлип Филлипович

Отбор пар упоминаний

На этапе применения

- перебор всех возможных пар

Отбор пар упоминаний

На этапе применения

- перебор всех возможных пар - **неэффективно!**
- для каждого упоминания анализируются **N** упоминаний слева. Выбор **N**:
  - эвристически
  - подбор по валидационному набору
  - по расстояниям в тренировочном наборе

Модель "пара упоминаний"

Подзадачи

- классификатор пар упоминаний
  - признаки
  - алгоритм машинного обучения
- отбор пар упоминаний
- объединение пар упоминаний в кореферентные цепочки

Объединение пар упоминаний в кореферентные цепочки

Все кореферентные пары

- использовать только пары, классифицированные как кореферентные
- отфильтровать по порогу уверенности
- преобразовать пары в кластеры по полученным компонентам связности

Объединение пар упоминаний в кореферентные цепочки

Все кореферентные пары

[0.98] Шариков - Шариков

[0.95] профессора - Филлип Филлипович

[0.91] Шариков - ему

[0.85] профессора - Он

[0.80] порог отсечения

[0.78] профессора - тот

Кореферентные цепочки:

{Шариков, Шариков, ему}

{профессора, Филлип Филлипович, Он}

Объединение пар упоминаний в кореферентные цепочки

Ближайшее кореферентное упоминание

Для каждого упоминания в качестве антецедента выбирается ближайшее к нему слева, классифицированное как кореферентное.

Объединение пар упоминаний в кореферентные цепочки

Ближайшее кореферентное упоминание

Для каждого упоминания в качестве антецедента выбирается ближайшее к нему слева, классифицированное как кореферентное.

[Шариков, профессора, тот] — ему
[Шариков, Филлип Филлипович, своем кабинете] — Он

Объединение пар упоминаний в кореферентные цепочки

Ближайшее кореферентное упоминание

Для каждого упоминания в качестве антецедента выбирается ближайшее к нему слева, классифицированное как кореферентное.

[Шариков, **профессора**, тот] — ему
[Шариков, **Филлип Филлипович**, своем кабинете] — Он

Объединение пар упоминаний в кореферентные цепочки

Лучшее кореферентное упоминание

Для каждого упоминания в качестве антецедента выбирается упоминание, классифицированное как кореферентное с большей уверенностью.

Объединение пар упоминаний в кореферентные цепочки

Лучшее кореферентное упоминание

Для каждого упоминания в качестве антецедента выбирается упоминание, классифицированное как кореферентное с большей уверенностью.

[Шариков$_{[0.8]}$, профессора$_{[0.7]}$, тот] — ему
[Шариков$_{[0.6]}$, Филлип Филлипович$_{[0.9]}$, своем кабинете] — Он

Объединение пар упоминаний в кореферентные цепочки

Лучшее кореферентное упоминание

Для каждого упоминания в качестве антецедента выбирается упоминание, классифицированное как кореферентное с большей уверенностью.

[**Шариков**$_{[0.8]}$, профессора$_{[0.7]}$, тот] — ему
[Шариков$_{[0.6]}$, **Филлип Филлипович**$_{[0.9]}$, своем кабинете] — Он

Объединение пар упоминаний в кореферентные цепочки

Easy-First Mention Pair

- Каждое упоминание образут свой собственный кластер.
- Упорядочить все пары упоминаний (положительные и отрицательные) по убыванию уверенности классификатора.
- Для каждой пары
  - Если пара отрицательная и не противоречит текущей структуре кластеров, то запомнить ее в качестве ограничения. Иначе - проигнорировать.
  - Если пара положительная и не противоречит текущим ограничениям, то объединить соответствующие кластера в один. Иначе - проигнорировать.

O.Uryupina, A. Moschitti
A State-of-the-Art Mention-Pair Model for Coreference Resolution

Объединение пар упоминаний в кореферентные цепочки

Easy-First Mention Pair

Классифицированные пары

профессора - Филлип Филиппович

тот - ему

ему - Шариков

профессора - тот

тот - Филлип Филиппович

ему - Филлип Филиппович

Ограничения

Кластеры

{профессора}, {Филлип Филиппович}, {тот}, {ему}, {Шариков}

Объединение пар упоминаний в кореферентные цепочки

Easy-First Mention Pair

Классифицированные пары

тот - ему
ему - Шариков
профессора - тот
тот - Филлип Филиппович
ему - Филлип Филиппович

Ограничения

Кластеры

{профессора, Филлип Филиппович}, {тот}, {ему}, {Шариков}

Объединение пар упоминаний в кореферентные цепочки

Easy-First Mention Pair

Классифицированные пары

ему - Шариков
профессора - тот
тот - Филлип Филиппович
ему - Филлип Филиппович

Ограничения

тот - ему

Кластеры

{профессора, Филлип Филиппович}, {тот}, {ему}, {Шариков}

Объединение пар упоминаний в кореферентные цепочки

Easy-First Mention Pair

Классифицированные пары

профессора - тот
тот - Филлип Филиппович
ему - Филлип Филиппович

Ограничения

тот - ему

Кластеры

{профессора, Филлип Филиппович}, {тот}, {ему, Шариков}

Объединение пар упоминаний в кореферентные цепочки

Easy-First Mention Pair

Классифицированные пары

тот - Филлип Филиппович

ему - Филлип Филиппович

Ограничения

тот - ему

Кластеры

{профессора, Филлип Филиппович, тот}, {ему, Шариков}

Объединение пар упоминаний в кореферентные цепочки

Easy-First Mention Pair

Классифицированные пары

Ограничения

тот - ему

ему - Филлип Филиппович

Кластеры

{профессора, Филлип Филиппович, тот}, {ему, Шариков}

Объединение пар упоминаний в кореферентные цепочки

Easy-First Mention Pair

Классифицированные пары

Ограничения

тот - ему

Кластеры

{профессора, Филлип Филиппович, тот}, {ему, Шариков}

Объединение пар упоминаний в кореферентные цепочки

Общие алгоритмы кластеризации

- Преобразовать список пар в граф (возможно, взвешенный)
- Применить алгоритм кластеризации графа:
  - Chinese whispers
  - Markov clustering
  - кластеризация на основе алгебраической связности

C. Biemann
Chinese Whispers: An Efficient Graph Clustering Algorithm and Its Application to Natural Language Processing Problems

S. van Dongen
Graph Clustering by Flow Simulation

Классификация

- модель "пара упоминаний" (mention-pair model)
- модель "сущность - упоминание" (entity-mention model)
- ранжирующие модели (ranking models)

Модель "сущность - упоминание"

Модель "сущность - упоминание"

- Текст анализируется слева направо.
- Для каждого упоминания нужно определить:
  - принадлежит ли оно одной из уже построенных кореферентных цепочек
  - или начинает новую цепочку

Модель "сущность - упоминание"

Пример

Шариков злобно покосился на профессора, а тот отправил ему косой взгляд. Через десять минут Шариков уехал в цирк. Филлип Филлипович остался один в своем кабинете. Он начал шагами мерять комнату.

{Шариков, ему, Шариков} - Филлип Филлипович
{профессора, тот} - Филлип Филлипович

Модель "сущность - упоминание"

Пример

Шариков злобно покосился на профессора, а тот отправил ему косой взгляд. Через десять минут Шариков уехал в цирк. Филлип Филлипович остался один в своем кабинете. Он начал шагами мерять комнату.

{Шариков, ему, Шариков} - своем кабинете
{профессора, тот, Филлип Филлипович} - своем кабинете

Модель "сущность - упоминание"

Пример

Шариков злобно покосился на профессора, а тот отправил ему косой взгляд. Через десять минут Шариков уехал в цирк. Филлип Филлипович остался один в своем кабинете. Он начал шагами мерять комнату.

Модель "сущность - упоминание"

Признаки

Признаки для пары <кореферентная цепочка, упоминание> получаются из признаков для пар упоминаний с помощью кванторов и агрегирующих функций:

- ALL
- ANY
- MOST
- MIN
- MAX
- AVG
- ...

Ранжирующие модели

Для каждого упоминания в явном виде происходит ранжирование возможных антецедентов / кореферентных цепочек.

Примеры моделей ранжирования:

- модель турнира
- RankSVM

Модель турнира

Бинарный классификатор для тройки:

- анафор
- антецедент-кандидат$_1$
- антецедент-кандидат$_2$

Метка указывает на то, какой антецедент правильный.

Для анафора перебираем все возможные пары антецедентов. В качестве правильного выбираем тот антецедент, который победил большее количество раз.

RankSVM

x - анафор

y^+ - правильный антецедент

$y_1^-, y_2^-, ..., y_n^-$ - неправильные антецеденты

$\varphi(x, y)$ - вектор признаков для пары анафор-антецедент $< x, y >$

Нужно подобрать вектор w:

$$(w, \varphi(x, y^+)) > (w, \varphi(x, y_i^-)), \text{ для } i = 1..n$$

RankSVM

x - анафор

y^+ - правильный антецедент

$y_1^-, y_2^-, ..., y_n^-$ - неправильные антецеденты

$\varphi(x, y)$ - вектор признаков для пары анафор-антецедент $< x, y >$

Нужно подобрать вектор w:

$$(w, \varphi(x, y^+)) > (w, \varphi(x, y_i^-)), \text{ для } i = 1..n$$

То есть

$$(w, \varphi(x, y^+) - \varphi(x, y_i^-)) > 0, \text{ для } i = 1..n$$

RankSVM

x - анафор

y^+ - правильный антецедент

$y_1^-, y_2^-, ..., y_n^-$ - неправильные антецеденты

$\varphi(x, y)$ - вектор признаков для пары анафор-антецедент $< x, y >$

Нужно подобрать вектор w:

$$(w, \varphi(x, y^+)) > (w, \varphi(x, y_i^-)), \text{ для } i = 1..n$$

То есть

$$(w, \varphi(x, y^+) - \varphi(x, y_i^-)) > 0, \text{ для } i = 1..n$$

Обучаем SVM на примерах:

$$\varphi(x, y^+) - \varphi(x, y_i^-) \longmapsto 1, \text{ для } i = 1..n$$

$$\varphi(x, y_i^-) - \varphi(x, y^+) \longmapsto -1, \text{ для } i = 1..n$$

RankSVM

x - анафор

y^+ - правильный антецедент

$y_1^-, y_2^-, ..., y_n^-$ - неправильные антецеденты

$\varphi(x, y)$ - вектор признаков для пары анафор-антецедент $< x, y >$

Нужно подобрать вектор w:

$$(w, \varphi(x, y^+)) > (w, \varphi(x, y_i^-)), \text{ для } i = 1..n$$

То есть

$$(w, \varphi(x, y^+) - \varphi(x, y_i^-)) > 0, \text{ для } i = 1..n$$

Обучаем SVM на примерах:

$$\varphi(x, y^+) - \varphi(x, y_i^-) \longmapsto 1, \text{ для } i = 1..n$$

$$\varphi(x, y_i^-) - \varphi(x, y^+) \longmapsto -1, \text{ для } i = 1..n$$

Применение:

$$y^* = \underset{y}{\text{argmax}}(w, \varphi(x, y))$$

План

- Кореферентность vs анафоричность.
- Подходы к разрешению кореферентности.
- Поиск упоминаний в тексте.
- Оценка качества.

Поиск упоминаний в тексте

Где брать упоминания (для разрешения кореферентности)?

- из существующей разметки (**в реальной жизни их не будет!**)
- получать с помощью синтаксического парсера (**много лишних!**)
- обучить алгоритм определения упоминаний

Поиск упоминаний в тексте

Алгоритм определения упоминаний:

- поиск главных слов
- "расширение" главных слов до упоминаний

H. Peng, K.-W. Chang, D. Roth
A Joint Framework for Coreference Resolution and Mention Head Detection

Поиск главных слов

Задача бинарной классификации токенов.

Алгоритм машинного обучения - например, SVM.

Признаки:

- базовые:
  - лемма
  - часть речи
  - граммемы (род, число, одушевленность)
- позиция в предложении
- синтаксические связи
- *базовые* для контекста и родителя в синтаксическом дереве
- количество детей в синтаксическом дереве
- частота токена в документе

"Расширение" главных слов до упоминаний

Итерационное "наращивание" главного слова влево / вправо до полного упоминания.

Задача бинарной классификации пары токенов: добавлять ли текущий токен к упоминанию с соответствующим главным словом.

Алгоритм машинного обучения - например, SVM.

Признаки:

- базовые для анализируемого токена и главного слова:
  - лемма
  - часть речи
- направление
- расстояние
- позиция в предложении
- анализируемый токен и главное слово - части именованной сущности
- наличие связей между анализируемым токеном и главным словом в синтаксическом дереве

"Расширение" главных слов до упоминаний

Пример

Негодяй в грязном колпаке — повар столовой Нормального питания служащих — плеснул кипятком и обварил мне левый бок.

"Расширение" главных слов до упоминаний

Пример

Негодяй в грязном колпаке — повар столовой Нормального питания служащих — плеснул кипятком и обварил мне левый бок.

"Расширение" главных слов до упоминаний

Пример

Негодяй в грязном колпаке — повар столовой Нормального питания служащих — плеснул кипятком и обварил мне левый бок.

"Расширение" главных слов до упоминаний

Пример

Негодяй в грязном колпаке — повар столовой Нормального питания служащих — плеснул кипятком и обварил мне левый бок.

"Расширение" главных слов до упоминаний

Пример

Негодяй в грязном колпаке — повар столовой Нормального питания служащих — плеснул кипятком и обварил мне левый бок.

"Расширение" главных слов до упоминаний

Пример

Негодяй в грязном колпаке — повар столовой Нормального питания служащих — плеснул кипятком и обварил мне левый бок.

План

- Кореферентность vs анафоричность.
- Подходы к разрешению кореферентности.
- Поиск упоминаний в тексте.
- Оценка качества.

Оценка качества

Поиск главных слов и упоминаний

$$\textit{точность} = \frac{\textit{количество правильно определенных}}{\textit{всего найденных системой}}$$

$$\textit{полнота} = \frac{\textit{количество правильно определенных}}{\textit{всего правильных}}$$

$$F1 = \frac{2 \times \textit{точность} \times \textit{полнота}}{\textit{точность} + \textit{полнота}}$$

Оценка качества

Разрешение кореферентности

Точность, полнота, F1 по версиям:

- MUC (Message Understanding Conference)
- B-cubed
- CEAF (Constrained Entity-Alignment F-Measure)

$$F1 = \frac{2 \times \textit{точность} \times \textit{полнота}}{\textit{точность} + \textit{полнота}}$$

MUC

S - правильная кореферентная цепочка
$R = R_1, R_2, ..., R_m$ - кореферентные цепочки, вычисленные системой
$P_R(S)$ - разбиение цепочки S в соответствии с ответами системы
$R_1, R_2, ..., R_m$, пересекающимися с S
Каждое упоминание, не входящее в ответы системы, образует кластер из одного элемента.

M. Vilain, J. Burger, J. Aberdeen, D. Connolly, L. Hirschman
A Model-Theoretic Coreference Scoring Scheme

MUC

S - правильная кореферентная цепочка
$R = R_1, R_2, ..., R_m$ - кореферентные цепочки, вычисленные системой
$P_R(S)$ - разбиение цепочки S в соответствии с ответами системы
$R_1, R_2, ..., R_m$, пересекающимися с S
Каждое упоминание, не входящее в ответы системы, образует кластер из одного элемента.

Если $S = \{A, B, C, D\}, R = \{\{A, B\}\}$, то $P_R(S) = \{\{A, B\}, \{C\}, \{D\}\}$

M. Vilain, J. Burger, J. Aberdeen, D. Connolly, L. Hirschman
A Model-Theoretic Coreference Scoring Scheme

MUC

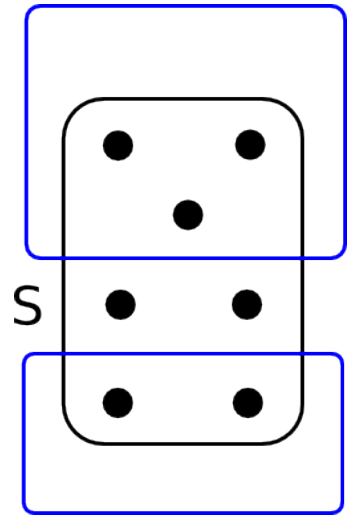
S - правильная кореферентная цепочка
$R = R_1, R_2, ..., R_m$ - кореферентные цепочки, вычисленные системой
$P_R(S)$ - разбиение цепочки S в соответствии с ответами системы
$R_1, R_2, ..., R_m$, пересекающимися с S
Каждое упоминание, не входящее в ответы системы, образует кластер из одного элемента.
$c(S) = |S| - 1$ - минимальное количество связей, необходимых для объединения элементов из S в кореферентную цепочку
$m_R(S) = |P_R(S)| - 1$ - количество связей, которые надо добавить в $P_R(S)$ для получения S

$$полнота_S = \frac{c(S) - m_R(S)}{c(S)} = \frac{|S| - |P_R(S)|}{|S| - 1}$$

M. Vilain, J. Burger, J. Aberdeen, D. Connolly, L. Hirschman
A Model-Theoretic Coreference Scoring Scheme

MUC

Пример



$$полнота_S = \frac{7-4}{7-1} = \frac{1}{2}$$

MUC

$$\textit{полнота} = \frac{\sum_S (|S| - |P_R(S)|)}{\sum_S (|S| - 1)}$$

MUC

$$\textit{полнота} = \frac{\sum_S (|S| - |P_R(S)|)}{\sum_S (|S| - 1)}$$

Для вычисления точности используются те же формулы, но правильные ответы и ответы системы "меняются местами".

B-cubed

M - множество всех упоминаний m

$S(m)$ - правильная цепочка, содержащая упоминание m

$R(m)$ - цепочка, выданная системой, содержащая упоминание m

Для каждого упоминания m:

$$точность_m = \frac{|R(m) \cap S(m)|}{|R(m)|}$$

$$полнота_m = \frac{|R(m) \cap S(m)|}{|S(m)|}$$

A. Bagga, B. Baldwin
Algorithms for scoring coreference chains

B-cubed

M - множество всех упоминаний m

$S(m)$ - правильная цепочка, содержащая упоминание m

$R(m)$ - цепочка, выданная системой, содержащая упоминание m

Для каждого упоминания m:

$$точность_m = \frac{|R(m) \cap S(m)|}{|R(m)|}$$

$$полнота_m = \frac{|R(m) \cap S(m)|}{|S(m)|}$$

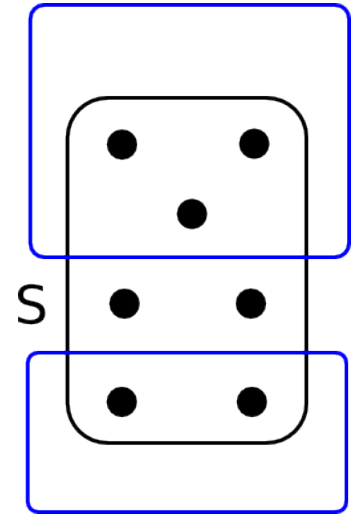
$$\textbf{точность} = \frac{1}{|M|} \sum_{m \in M} точность_m$$

$$\textbf{полнота} = \frac{1}{|M|} \sum_{m \in M} полнота_m$$

A. Bagga, B. Baldwin
Algorithms for scoring coreference chains

B-cubed

Пример



$$\textit{полнота} = \frac{1}{7}\left(\frac{3}{7}+\frac{3}{7}+\frac{3}{7}+\frac{1}{7}+\frac{1}{7}+\frac{2}{7}+\frac{2}{7}\right) = \frac{1}{7}\times\frac{15}{7} = \frac{15}{49}$$

$$\textit{точность} = \frac{1}{7}\left(\frac{3}{3}+\frac{3}{3}+\frac{3}{3}+\frac{1}{1}+\frac{1}{1}+\frac{2}{2}+\frac{2}{2}\right) = \frac{1}{7}\times 7 = 1.0$$

CEAF

$\mathfrak{S} = \{S_1, S_2, ..., S_n\}$ - правильные кореферентные цепочки
$\mathfrak{R} = \{R_1, R_2, ..., R_m\}$ - кореферентные цепочки, вычисленные системой
$\varphi(S, R) \geq 0$ - оценка близости множеств S и R

$\varphi(S, R) = 0$, если $S = \emptyset$ или $R = \emptyset$. Дальше можно считать, что $n = m$ - меньшее множество дополняем с помощью $\emptyset$ до большего.

$g : \mathfrak{S} \mapsto \mathfrak{R}$ - взаимнооднозначное соответствие между $\mathfrak{S}$ и $\mathfrak{R}$
$\Phi(g) = \sum_{i=1}^{n} \varphi(S_i, g(S_i))$ - суммарная близость
$g^* = \text{argmax}_g \Phi(g)$

$$\textit{точность}_{CEAF} = \frac{\Phi(g^*)}{\sum_{i=1}^{n} \varphi(R_i, R_i)}$$

$$\textit{полнота}_{CEAF} = \frac{\Phi(g^*)}{\sum_{i=1}^{n} \varphi(S_i, S_i)}$$

X. Luo
On Coreference Resolution Performance Metrics

CEAF

$\text{CEAF}_{\textit{упоминание}}$ (CEAF_m):

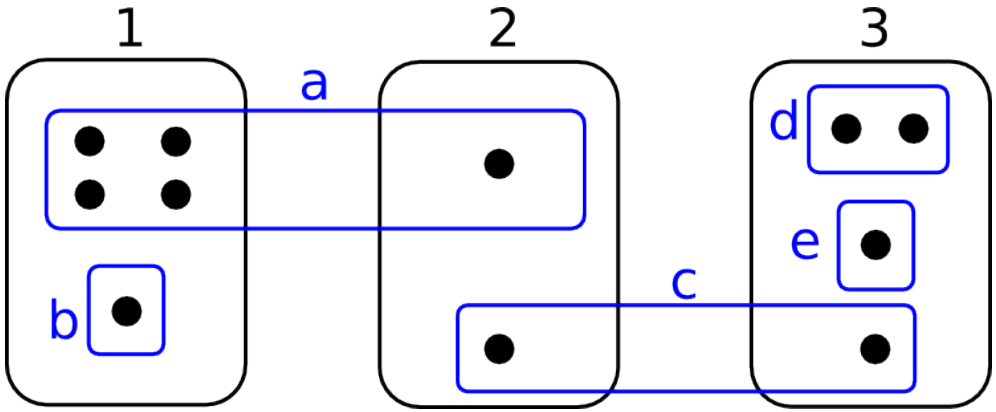
$$\varphi_m(S, R) = |S \cap R|$$

$\text{CEAF}_{\textit{сущность}}$ (CEAF_e):

$$\varphi_e(S, R) = \frac{2|S \cap R|}{|S| + |R|}$$

CEAF

Пример



	1	**2**	**3**	∅	∅
a	4	1	0	0	0
b	1	0	0	0	0
c	0	1	1	0	0
d	0	0	2	0	0
e	0	0	1	0	0

$$точность_{CEAF_m} = полнота_{CEAF_m} = \frac{7}{11}$$

Следующая лекция

Другие задачи обработки текстов

лектор: Майоров Владимир Дмитриевич