ОСНОВЫ ОБРАБОТКИ ТЕКСТОВ

Лекция №11
Привязка к базам знаний

Лектор: н.с. ИСП РАН Сысоев Андрей Анатольевич

5 декабря 2018

План

- базы знаний
- задачи викификации и устранения многозначности
- поиск терминов
- алгоритм Milne-Witten
- алгоритм GLOW
- GLOW + интеллектуальный контекст
- оценка качества

Что такое "база знаний"?

В широком смысле

База знаний — база данных, содержащая правила вывода и информацию о человеческом опыте и знаниях в некоторой предметной области.

Что такое "база знаний"?

В широком смысле

База знаний — база данных, содержащая правила вывода и информацию о человеческом опыте и знаниях в некоторой предметной области.

В узком смысле

База знаний — некоторое структурированное описание предметной области.

База знаний. Примеры

WordNet

Тезаурус **WordNet** — словарь + семантические отношения между словами.

WordNet для русского: `https://ruwordnet.ru/ru/`

База знаний. Примеры

WordNet

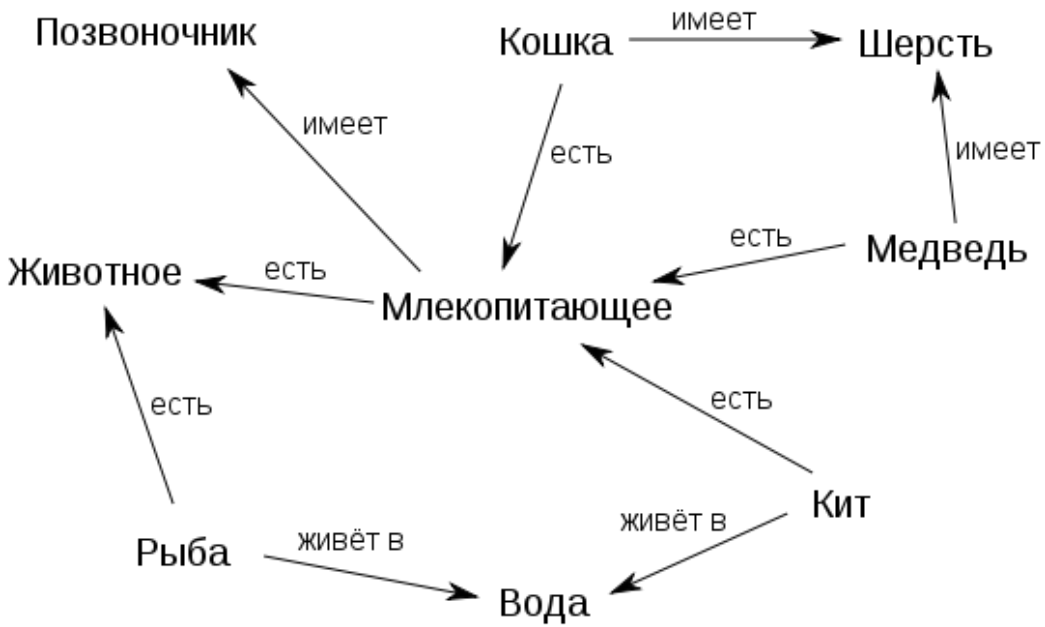
Дом

- дом 1
  - синсет: ДОМ, ЗДАНИЕ
  - гипероним
    - СТРОЕНИЕ, ПОСТРОЙКА, СООРУЖЕНИЕ
    - ВЛАДЕНИЕ, НЕДВИЖИМОСТЬ, ...
  - гипоним
    - НЕЖИЛОЕ ЗДАНИЕ, НЕЖИЛАЯ ПОСТРОЙКА, ...
    - ДВОРЕЦ, ПАЛАЦЦО, ...
    - ДЕПО, ЗДАНИЕ ПОЖАРНОЙ ОХРАНЫ, ...
    - ...
  - часть
    - ЛЕПНОЕ УКРАШЕНИЕ, ЛЕПНИНА
    - ЭТАЖ
    - ЦОКОЛЬ, ЦОКОЛЬ ЗДАНИЯ, ЦОКОЛЬ СТРОЕНИЯ
    - ...
- дом 2
  - синсет: ДОМ, КРОВ, СВОЙ КРОВ, КРЫША НАД ГОЛОВОЙ, ПЕНАТЫ, РОДНОЙ ДОМ, ДОМАШНИЙ ОЧАГ, СЕМЕЙНЫЙ ОЧАГ
  - гипероним: ЖИЛЬЕ, ЖИЛОЙ МОДУЛЬ, ЖИЛОЕ ПОМЕЩЕНИЕ, ЖИЛОЕ ПРОСТРАНСТВО, ЖИЛИЩЕ

База знаний. Примеры

Семантическая сеть

Семантическая сеть — информационная модель предметной области, имеющая вид ориентированного графа, вершины которого соответствуют объектам предметной области, а дуги (рёбра) задают отношения между ними.



---

https://ru.wikipedia.org/wiki/Семантическая_сеть

База знаний. Примеры

Викисловарь

Викисловарь - `https://ru.wiktionary.org`.

Дом

1. архитектурное сооружение, предназначенное для жилья, и имеющее, как правило, стены, дверь и крышу
   Просторный дом. Трёхэтажный дом.
2. место, где кто-либо постоянно проживает
   Здесь мой родной дом.
3. *офиц.* совокупность жилых или производственных корпусов, а также служебных строений, расположенных на одном земельном участке и имеющих один учётный номер
4. *перен.* фирма, предприятие
   Торговый дом. Издательский дом.
5. *спорт.* в кёрлинге — мишень в конце ледовой полосы, образованная четырьмя концентрическими кругами
6. *спорт.* в бейсболе база, с которой начинается и которой заканчивается пробежка игрока
7. *перен.* семья, династия, клан
   Трёхсотлетие дома Романовых. Чума на оба ваши дома.

База знаний. Примеры

Викисловарь

Дом

- Синонимы
  1. здание, корпус
  2. жилище, жильё, жилплощадь, резиденция
  3. —
  4. фирма, предприятие
  5. —
  6. —
  7. семья, династия, клан
- Антонимы
  1. —
  2. —
- Гиперонимы
  1. строение, постройка
  2. —
  3. учреждение
  4. —
- Гипонимы
  1. барак, вилла, коттедж, многоэтажка, небоскрёб, хрущоба
  2. автодом, исправдом

Привязка к базе знаний

В клетке сидит морская свинка.

Привязка к базе знаний

В клетке сидит морская свинка.

1. внутрь чего-либо (предмета или места)
   Пойти в дом.
2. на поверхность чего-либо (предмета)
   Толкнуть в грудь.
3. через, сквозь чего-либо
   Смотреть в окно.

   ...
13. внутри, посреди
    Карандаш лежит в пенале.

    ...
21. образует фразеологизмы и словосочетания с общим значением точного соответствия
    Слово в слово.

Привязка к базе знаний

В клетке сидит морская свинка.

1. внутрь чего-либо (предмета или места)
   Пойти в дом.
2. на поверхность чего-либо (предмета)
   Толкнуть в грудь.
3. через, сквозь чего-либо
   Смотреть в окно.

   ...
13. **внутри, посреди**
    Карандаш лежит в пенале.

    ...
21. образует фразеологизмы и словосочетания с общим значением точного соответствия
    Слово в слово.

Привязка к базе знаний

В **клетке** сидит морская свинка.

1. сооружение или помещение, стены (стенки) которого состоят из параллельных друг другу или пересекающихся прутьев, палок и т. п.
   Птичья клетка. Клетка с тигром.
2. отдельная ячейка поверхности, расчерченной рядами пересекающихся линий
3. *биол.* элементарная единица строения живых организмов, обладающая собственной устойчивой структурой, обменом веществ, а также, обычно, способностью к самостоятельному существованию, самовоспроизведению и развитию

Привязка к базе знаний

В **клетке** сидит морская свинка.

1. **сооружение или помещение, стены (стенки) которого состоят из параллельных друг другу или пересекающихся прутьев, палок и т. п.**
   Птичья клетка. Клетка с тигром.
2. отдельная ячейка поверхности, расчерченной рядами пересекающихся линий
3. *биол.* элементарная единица строения живых организмов, обладающая собственной устойчивой структурой, обменом веществ, а также, обычно, способностью к самостоятельному существованию, самовоспроизведению и развитию

Привязка к базе знаний

В клетке **сидит** морская свинка.

1. опираться нижней частью туловища на какую-либо поверхность, держа туловище в положении, близком к вертикальному
   Сидеть на лавочке.
2. *о птице или насекомом* находиться в неподвижности на каком-либо предмете, на какой-либо опоре

   ...
4. *перен., разг.* находиться внутри чего-либо, таиться, содержаться где-либо

   ...
10. *жарг.* регулярно нуждаться, использовать что-либо, зависеть от чего либо

Привязка к базе знаний

В клетке **сидит** морская свинка.

1. **опираться нижней частью туловища на какую-либо поверхность, держа туловище в положении, близком к вертикальному**
   Сидеть на лавочке.
2. *о птице или насекомом* находиться в неподвижности на каком-либо предмете, на какой-либо опоре

   ...
4. *перен., разг.* **находиться внутри чего-либо, таиться, содержаться где-либо**

   ...
10. *жарг.* регулярно нуждаться, использовать что-либо, зависеть от чего либо

Привязка к базе знаний

В клетке сидит **морская** свинка.

1. относящийся к морю, свойственный морю, происходящий на море

Привязка к базе знаний

В клетке сидит морская **свинка**.

1. *уменьш.-ласк.* от свинья, маленький поросёнок.
2. *разг.* эпидемический паротит.

Привязка к базе знаний

В клетке сидит морская **свинка**.

1. *уменьш.-ласк.* от свинья, маленький поросёнок.
2. *разг.* эпидемический паротит.
3. **морская свинка - вид одомашненных грызунов из рода свинок семейства свинковых, популярные домашние животные.**

Привязка к базе знаний

В клетке сидит морская свинка.

Понимание слов предложения $\neq$ понимание предложения

Во время интервью:

— Ваша главная слабость?

— Правильно интерпретирую семантику вопроса, но игнорирую его суть.

— Не могли бы вы привести пример?

— *Мог бы.*

База знаний

Википедия

Вы не представились системе Обсуждение Вклад Создать учётную запись Войти

ВикипедиЯ
Свободная энциклопедия

Статья | Обсуждение | Читать | Текущая версия | Править | Править код | История | Искать в Википедии

- Заглавная страница
- Рубрикация
- Указатель А — Я
- Избранные статьи
- Случайная статья
- Текущие события

Участие

- Сообщить об ошибке
- Сообщество
- Форум
- Свежие правки
- Новые страницы
- Справка
- Пожертвовать

Инструменты

- Ссылки сюда
- Связанные правки
- Служебные страницы
- Постоянная ссылка
- Сведения о странице
- Цитировать страницу

Московский государственный университет

55°42′11″ с. ш. 37°31′50″ в. д. H G Я O

Материал из Википедии — свободной энциклопедии [править | править код]

Текущая версия страницы пока не проверялась опытными участниками и может значительно отличаться от версии, проверенной 21 октября 2018; проверки требуют 7 правок.

Запрос «МГУ» перенаправляется сюда; см. также другие значения.

Моско́вский госуда́рственный университе́т и́мени М. В. Ломоно́сова — один из старейших[4][5] и крупнейших[6][7] классических университетов России, один из центров отечественной науки и культуры, расположенный в Москве.

С 1940 года носит имя Михаила Васильевича Ломоносова.

Полное наименование — Федеральное государственное бюджетное образовательное учреждение высшего образования «Московский государственный университет имени М. В. Ломоносова». Широко используется аббревиатура «МГУ».

Университет включает в себя 15 научно-исследовательских институтов[8], 43 факультета[9][10], более 300 кафедр и 6 филиалов (в их числе пять зарубежных — все в странах СНГ)[11].

С 1992 года ректором МГУ является академик Виктор Антонович Садовничий.

Содержание [скрыть]

Московский государственный университет имени М. В. Ломоносова (МГУ имени М. В. Ломоносова)

Главное здание МГУ имени М. В. Ломоносова, 31 мая 2015 года

База знаний

Википедия

Морская свинка — вид одомашненных грызунов из рода свинок семейства свинковых.

Свинки — род млекопитающих из семейства свинковых.

Свинковые — одно из семейств грызунов.

Морские свиньи — семейство морских млекопитающих подотряда зубатых китов.

Свиньи — семейство нежвачных парнокопытных.

Свинка — река в Московской области России.

Эпидемический паротит (лат. parotitis epidemica: свинка, заушница) — острое инфекционное заболевание.

Варианты использования

Определение тем текста

Последние несколько дней Белинда работала с разработчиками новой системы для **аэропорта**. С тех пор, как мистер Томпкинс принес им украденные в Штатах **спецификации**, они взялись за их изучение, причем начали с одного из основных компонентов — **системы радиоуправления самолетами**. Никто не знал, насколько многоплановой и сложной будет система для моровийского **аэропорта**, однако в любом случае там должна была быть **система радиоуправления**. Вебстер тоже потратил около трех часов на изучение той части **спецификации**, которая касалась этой части системы, а потом отправился на ежевечернее собрание группы разработчиков.

Спецификация · Аэропорт · Радиоуправление · Самолёт

Том ДеМарко "Deadline. Роман об управлении проектами"

Варианты использования

Облегчение понимания текста

— Клевый городок, — говорит Джон, — действительно клевый. Не думал, что такие еще остались. Я утром все осмотрел. У них тут есть бары для скотников — высокие сапоги, пряжки из серебряных долларов, «ливайсы», **стетсоны**, и прочее... И всё — настоящее. Не просто барахло из Торговой Палаты... Сегодня утром в баре через квартал отсюда они заговорили со мной так, будто я прожил здесь всю жизнь.

Роберт Пёрсиг "Дзэн и искусство ухода за мотоциклом"

Варианты использования

Облегчение понимания текста

— Клевый городок, — говорит Джон, — действительно клевый. Не думал, что такие еще остались. Я утром все осмотрел. У них тут есть бары для скотников — высокие сапоги, пряжки из серебряных долларов, «ливайсы», **стетсоны**, и прочее... И всё — настоящее. Не просто барахло из Торговой Палаты... Сегодня утром в баре через квартал отсюда они заговорили со мной так, будто я прожил здесь всю жизнь.

Стетсон (лунный кратер)
Колин Стетсон (саксофонист)
Стетсонский университет
Стетсон (Ковбойская шляпа)

Роберт Пёрсиг "Дзэн и искусство ухода за мотоциклом"

Варианты использования

Облегчение понимания текста

— Клевый городок, — говорит Джон, — действительно клевый. Не думал, что такие еще остались. Я утром все осмотрел. У них тут есть бары для скотников — высокие сапоги, пряжки из серебряных долларов, «ливайсы», **стетсоны**, и прочее... И всё — настоящее. Не просто барахло из Торговой Палаты... Сегодня утром в баре через квартал отсюда они заговорили со мной так, будто я прожил здесь всю жизнь.

Стетсон (лунный кратер)
Колин Стетсон (саксофонист)
Стетсонский университет
Стетсон (Ковбойская шляпа)

Роберт Пёрсиг "Дзэн и искусство ухода за мотоциклом"

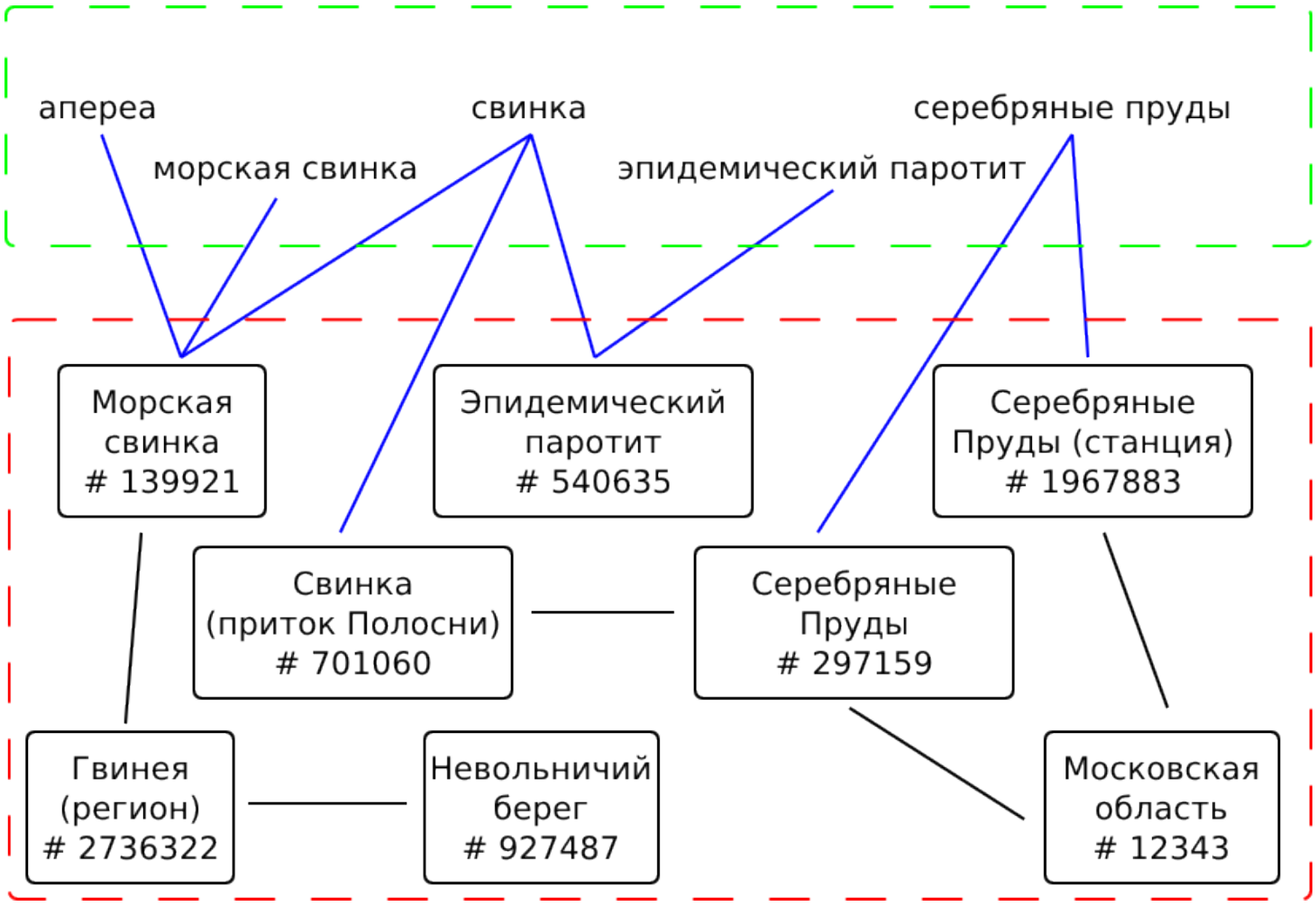
Задачи

- **Викификация** — найти термины и определить их значения из базы знаний (или `NOT_IN_KB`).
- **Устранение многозначности** — определить значения для данных терминов.
- **Связывание именованных сущностей** — викификация именованных сущностей.

База знаний: чаще всего, Википедия, но может быть и любая предметно-специфичная.

База знаний

Википедия



База знаний

Википедия

Связи между концептами - ссылки в Википедии.

Связи между терминами и концептами:

- название статьи
  "Пирсиг, Роберт" # 2089853
- перенаправления
  "Big Blue" $\rightarrow$ "IBM" # 11078
- подписи под ссылками:
  "Были одомашнены инками и использовались ..." $\rightarrow$
  "Империя инков" # 206211

Поиск терминов

Перебор 1—n-грамм, у которых *вероятность ссылки* больше некоторого порога.

Поиск терминов

Перебор 1—n-грамм, у которых *вероятность ссылки* больше некоторого порога.

1—2-граммы

Бестселлер американского писателя Роберта Пёрсига разошёлся на разных языках общим тиражом более 4 миллионов экземпляров.

Поиск терминов

Перебор 1—n-грамм, у которых *вероятность ссылки* больше некоторого порога.

1—2-граммы

Бестселлер американского писателя Роберта Пёрсига разошёлся на разных языках общим тиражом более 4 миллионов экземпляров.

- Бестселлер

Поиск терминов

Перебор 1—n-грамм, у которых *вероятность ссылки* больше некоторого порога.

1—2-граммы

Бестселлер американского писателя Роберта Пёрсига разошёлся на разных языках общим тиражом более 4 миллионов экземпляров.

- Бестселлер

Поиск терминов

Перебор 1—n-грамм, у которых *вероятность ссылки* больше некоторого порога.

1—2-граммы

Бестселлер американского писателя Роберта Пёрсига разошёлся на разных языках общим тиражом более 4 миллионов экземпляров.

- Бестселлер

Поиск терминов

Перебор 1—n-грамм, у которых *вероятность ссылки* больше некоторого порога.

1—2-граммы

Бестселлер американского писателя Роберта Пёрсига разошёлся на разных языках общим тиражом более 4 миллионов экземпляров.

- Бестселлер

Поиск терминов

Перебор 1—n-грамм, у которых *вероятность ссылки* больше некоторого порога.

1—2-граммы

Бестселлер американского писателя Роберта Пёрсига разошёлся на разных языках общим тиражом более 4 миллионов экземпляров.

- Бестселлер
- писателя

Поиск терминов

Перебор 1—n-грамм, у которых *вероятность ссылки* больше некоторого порога.

1—2-граммы

Бестселлер американского писателя Роберта Пёрсига разошёлся на разных языках общим тиражом более 4 миллионов экземпляров.

- Бестселлер
- писателя

Поиск терминов

Перебор 1—n-грамм, у которых *вероятность ссылки* больше некоторого порога.

1—2-граммы

Бестселлер американского писателя Роберта Пёрсига разошёлся на разных языках общим тиражом более 4 миллионов экземпляров.

- Бестселлер
- писателя
- Роберта

Поиск терминов

Перебор 1—n-грамм, у которых *вероятность ссылки* больше некоторого порога.

1—2-граммы

Бестселлер американского писателя Роберта Пёрсига разошёлся на разных языках общим тиражом более 4 миллионов экземпляров.

- Бестселлер
- писателя
- Роберта
- Роберта Пёрсига

Поиск терминов

Перебор 1—n-грамм, у которых *вероятность ссылки* больше некоторого порога.

1—2-граммы

Бестселлер американского писателя Роберта Пёрсига разошёлся на разных языках общим тиражом более 4 миллионов экземпляров.

- Бестселлер
- писателя
- Роберта
- Роберта Пёрсига

Вероятность ссылки

$$\frac{\textit{количество терминов как ссылка}}{\textit{количество терминов всего}}$$

"...известный главным образом как автор книги «*Дзэн и искусство ухода за мотоциклом*» (1974), ..."

"В книге «*Дзэн и искусство ухода за мотоциклом*» он описывает себя как далеко не типичного студента."

"В более поздних изданиях книги «*Дзэн и искусство ухода за мотоциклом*» Пирсиг пишет об этом ..."

Вероятность ссылки (*link probability*)

$$\textit{link\_probability}(\textit{Дзэн и искусство ухода за мотоциклом}) = \frac{1}{3}$$

Устранение многозначности

Baseline

Каждому термину назначается наиболее популярный концепт.

Baseline

Каждому термину назначается наиболее популярный концепт.

"лето" ($\sum = \mathbf{1398}$)

- Лето # 58195 **889**
- Летнее время # 22194 **399**
- Лето (мифология) # 54858 **58**
- Лето (фильм, 1976) # 3324176 **16**
- Лето (фильм, 1955) # 4190361 **14**
- Апертура и Клаусура # 937358 **6**
- Летард (граф Фезансака) # 4622310 **6**
- Лето (песня) # 4847903 **5**
- Лета # 88113 **5**

Baseline

Каждому термину назначается наиболее популярный концепт.

"лето" ($\sum = \mathbf{1398}$)

- Лето # 58195 **889**
- Летнее время # 22194 **399**
- Лето (мифология) # 54858 **58**
- Лето (фильм, 1976) # 3324176 **16**
- Лето (фильм, 1955) # 4190361 **14**
- Апертура и Клаусура # 937358 **6**
- Летард (граф Фезансака) # 4622310 **6**
- Лето (песня) # 4847903 **5**
- Лета # 88113 **5**

Популярность значения термина — $commonness(e, m)$

$$commonness(\textit{Лето \# 58195}, \textit{лето}) = \frac{889}{1398} = 0.636$$

Baseline не работает

Нужно использовать контекст

В Дельфах была скульптурная группа, изображающая борьбу: Лето и Артемида успокаивают Аполлона, Афина удерживает Геракла.

- ~~Лето # 58195 **889**~~
- ~~Летнее время # 22194 **399**~~
- **Лето (мифология)** # 54858 **58**
- ...

Алгоритм Milne-Witten

Контекст строится на основе однозначных терминов.

D. Milne and I. H. Witten
Learning to Link with Wikipedia

Алгоритм Milne-Witten

Контекст строится на основе однозначных терминов.

В **Дельфах** была скульптурная группа, изображающая борьбу: Лето и **Артемида** успокаивают **Аполлона**, **Афина** удерживает **Геракла**.

D. Milne and I. H. Witten
Learning to Link with Wikipedia

Алгоритм Milne-Witten

Контекст строится на основе однозначных терминов.

В **Дельфах** была скульптурная группа, изображающая борьбу: Лето и **Артемида** успокаивают **Аполлона**, **Афина** удерживает **Геракла**.

N.B.: все эти термины уже стали многозначными.

D. Milne and I. H. Witten
Learning to Link with Wikipedia

Алгоритм Milne-Witten

Семантическое расстояние между концептами

$$relatedness(e, c) = \frac{log\ max(|E|, |C|) - log|E \cap C|}{log|W| - log\ min(|E|, |C|)}$$

e, c - концепты;

E, C - множества концептов, ссылающихся на e и c соответственно;

W - множество всех концептов.

D. Milne and I. H. Witten
An Effective, Low-Cost Measure of Semantic Relatedness Obtained from Wikipedia Links

Алгоритм Milne-Witten

Семантическое расстояние между концептами

$$relatedness(e, c) = \frac{log\ max(|E|, |C|) - log|E \cap C|}{log|W| - log\ min(|E|, |C|)}$$

e, c - концепты;

E, C - множества концептов, ссылающихся на e и c соответственно;

W - множество всех концептов.

N.B.: на самом деле, это Normalized Google distance

R. Cilibrasi and P. Vitanyi
The Google Similarity Distance

Алгоритм Milne-Witten

Семантическое расстояние от концепта e до контекста S:

$$distance(e, S) = \frac{\sum_{s \in S} w_s \times relatedness(e, s)}{\sum_{s \in S} w_s}$$

Алгоритм Milne-Witten

Семантическое расстояние от концепта e до контекста S:

$$distance(e, S) = \frac{\sum_{s \in S} w_s \times relatedness(e, s)}{\sum_{s \in S} w_s}$$

Вес w_s концепта s:

$$w_s = \frac{1}{2}(link\_probability(term(s)) + \frac{1}{|S| - 1} \sum_{c \in S \setminus \{s\}} relatedness(s, c))$$

Алгоритм Milne-Witten

Семантическое расстояние от концепта e до контекста S:

$$distance(e, S) = \frac{\sum_{s \in S} w_s \times relatedness(e, s)}{\sum_{s \in S} w_s}$$

Вес w_s концепта s:

$$w_s = \frac{1}{2}(link\_probability(term(s)) + \frac{1}{|S| - 1} \sum_{c \in S \setminus \{s\}} relatedness(s, c))$$

- $term(s)$ - т.к. могут быть несколько однозначных терминов с одним и тем же концептом;

Алгоритм Milne-Witten

Семантическое расстояние от концепта e до контекста S:

$$distance(e, S) = \frac{\sum_{s \in S} w_s \times relatedness(e, s)}{\sum_{s \in S} w_s}$$

Вес w_s концепта s:

$$w_s = \frac{1}{2}(link\_probability(term(s)) + \frac{1}{|S| - 1} \sum_{c \in S \setminus \{s\}} relatedness(s, c))$$

- $term(s)$ - т.к. могут быть несколько однозначных терминов с одним и тем же концептом;
- возможно, там должно быть $1 - relatedness(s, c)$.

Алгоритм Milne-Witten

Когерентность контекста

Оценка "качества" и "согласованности" контекста:

$$quality(S) = \sum_{s \in S} w_s$$

Алгоритм Milne-Witten

Выбор концепта для термина

Признаки:

- популярность значения - $commonness(e, m)$;
- расстояние до контекста - $distance(e, S)$;
- когерентность контекста - $quality(S)$.

Алгоритмы машинного обучения:

- Naïve Bayes;
- дерево решений C4.5;
- метод опорных векторов;
- **дерево решений C4.5 с бэггингом**.

Обучение / применение:

- положительный пример — правильное значение термина;
- отрицательные примеры — все остальные возможные значения термина;
- на этапе применения выбирается концепт, с максимальной уверенностью классифицированный как правильный.

Алгоритм Milne-Witten

Фильтрация

- По правилам Википедии, нужны только наиболее релевантные ссылки.
- Если нет правильного значения, то термин не нужен.

Алгоритм Milne-Witten

Фильтрация

Алгоритм фильтрации концептов на основе признаков:

- вероятность ссылки - $link\_probability(m)$: max, avg;
- расстояние до контекста - $distance(e, S)$;
- уверенность классификатора выбора концепта: max, avg;
- общность — расстояние до корня в графе категорий Википедии;
- положение в тексте;
- частота появления.

Векторная модель документа

Роберт Мейнард Пирсиг — американский писатель и философ. ... В 1946 году Пирсиг поступил на военную службу и был направлен в Корею. ... В 1961—1963 годах Пирсиг лечился в психиатрических клиниках.

роберт			
мейнард			
пирсиг			
американский			
писатель			

...

сша			
в			
и			

...

Векторная модель документа

Роберт Мейнард Пирсиг — американский писатель и философ. ... В 1946 году Пирсиг поступил на военную службу и был направлен в Корею. ... В 1961—1963 годах Пирсиг лечился в психиатрических клиниках.

	Булевский вес		
роберт	1		
мейнард	1		
пирсиг	1		
американский	1		
писатель	1		

...

сша	0		
в	1		
и	1		

...

Векторная модель документа

Роберт Мейнард Пирсиг — американский писатель и философ. ... В 1946 году Пирсиг поступил на военную службу и был направлен в Корею. ... В 1961—1963 годах Пирсиг лечился в психиатрических клиниках.

	Булевский вес	Количество упоминаний	
роберт	1	1	
мейнард	1	1	
пирсиг	1	3	
американский	1	1	
писатель	1	1	

...

сша	0	0	
в	1	4	
и	1	2	

...

Векторная модель документа

Роберт Мейнард Пирсиг — американский писатель и философ. ... В 1946 году Пирсиг поступил на военную службу и был направлен в Корею. ... В 1961—1963 годах Пирсиг лечился в психиатрических клиниках.

	Булевский вес	Количество упоминаний	TF-IDF
роберт	1	1	0.87
мейнард	1	1	1.65
пирсиг	1	3	2.8
американский	1	1	0.002
писатель	1	1	0.03
...			
сша	0	0	0
в	1	4	0.00001
и	1	2	0.00002
...			

TF-IDF

term frequency - inverse document frequency

Пусть $D = \{d\}$ - коллекция документов

d - документ

t - токен

$$tf(t, d) = \frac{count(t, d)}{\sum_{\tilde{t} \in d} count(\tilde{t}, d)}$$

$$idf(t, D) = log \frac{|D|}{|\{d \in D | t \in d\}|}$$

$$TF\text{-}IDF(t, d, D) = tf(t, d) \times idf(t, D)$$

Алгоритм GLOW

Используются два вида контекста:

- локальный (по словам)
- глобальный (по концептам)

L. Ratinov, D. Roth, D. Downey and M. Anderson
Local and Global Algorithms for Disambiguation to Wikipedia

Алгоритм GLOW

Используются два вида контекста:

- локальный (по словам)
- глобальный (по концептам)

Схема работы:

- использовать локальный контекст для вычисления глобального;
- использовать локальный и глобальный контексты для получения финального результата.

L. Ratinov, D. Roth, D. Downey and M. Anderson
Local and Global Algorithms for Disambiguation to Wikipedia

Алгоритм GLOW

Используются два вида контекста:

- локальный (по словам)
- глобальный (по концептам)

Схема работы:

- использовать локальный контекст для вычисления глобального;
- использовать локальный и глобальный контексты для получения финального результата.

На каждом этапе:

- ранжирование концептов термина с помощью RankSVM;
- фильтрация потенциально неверных результатов ранжирования с помощью SVM.

L. Ratinov, D. Roth, D. Downey and M. Anderson
Local and Global Algorithms for Disambiguation to Wikipedia

Ранжирование концептов термина

Ранжирование концептов термина

Признаки, не зависящие от контекста

- Популярность значения термина — $commonness(e, m)$.
- Популярность значения — $P(e)$ — доля всех концептов, ссылающихся на e.

Ранжирование концептов термина

Признаки на основе локального контекста

- $text(m)$ / $text(e)$ — TF-IDF вектор, построенный по всему тексту документа / статьи Википедии.
- $context(m)$ / $context(e)$ — TF-IDF вектор, построенный по окну вокруг термина в документе / всем окнам вокруг всех упоминаний концепта в Википедии.

Ранжирование концептов термина

Признаки на основе локального контекста

$text(e)$

Пирсиг, Роберт
Роберт Мейнард Пирсиг (англ. Robert Maynard Pirsig; 6 сентября 1928, Миннеаполис, штат Миннесота — 24 апреля 2017) — американский писатель и философ, известный главным образом как автор книги «Дзэн и искусство ухода за мотоциклом» (1974), более чем 4 миллиона экземпляров которой было продано по всему миру. ...

$context(e)$

Дзен и искусство ухода за мотоциклом
«Дзен и искусство ухода за мотоциклом» (англ. Zen and the Art of Motorcycle Maintenance: An Inquiry into Values) — бестселлер американского писателя **Роберта Пёрсига** (1974), разошедшийся на разных языках общим тиражом более 4 миллионов экземпляров.

Ранжирование концептов термина

Признаки на основе локального контекста

Косинус угла между векторами (m — термин, e — концепт):

$text(m) \leftrightarrow text(e)$

$text(m) \leftrightarrow context(e)$

$context(m) \leftrightarrow text(e)$

$context(m) \leftrightarrow context(e)$

Ранжирование концептов термина

Признаки на основе локального контекста

Модифицированный TF-IDF:

$$w_{text}(t, e, m) = \frac{text(e)_{[t]}}{\sum_{\tilde{e} \in E_m} text(\tilde{e})_{[t]}}$$

	TF-IDF	книга	роман	писатель	американский	британский	повесть
значение 1	обычный	0.4	0.7	0.3	0.2	0	0
	модифицированный						
значение 2	обычный	0.1	0.6	0.4	0	0.1	0
	модифицированный						
документ	обычный	0.2	0	0.3	0.1	0	0.7

cos	значение 1	значение 2
обычный TF-IDF	0.27	0.24
модифицированный TF-IDF		

Ранжирование концептов термина

Признаки на основе локального контекста

Модифицированный TF-IDF:

$$w_{text}(t, e, m) = \frac{text(e)_{[t]}}{\sum_{\tilde{e} \in E_m} text(\tilde{e})_{[t]}}$$

	TF-IDF	книга	роман	писатель	американский	британский	повесть
значение 1	обычный	0.4	0.7	0.3	0.2	0	0
	модифицированный	0.8	0.54	0.57	1	0	0
значение 2	обычный	0.1	0.6	0.4	0	0.1	0
	модифицированный	0.1	0.46	0.43	0	1	0
документ	обычный	0.2	0	0.3	0.1	0	0.7

cos	значение 1	значение 2
обычный TF-IDF	0.27	0.24
модифицированный TF-IDF	0.36	0.18

Ранжирование концептов термина

Признаки на основе глобального контекста

agg_G — агрегируюшая функция:

- max_G
- avg_G

$\mathbb{1}$ — индикатор наличия связи:

- $\mathbb{1}_{e_i - e_j}$ — односторонняя связь
- $\mathbb{1}_{e_i \leftrightarrow e_j}$ — двусторонняя связь

sim — оценка семантической близости

- Pointwise mutual information — PMI
- Normalized Google distance — NGD

$links$ — множество концептов

- in_links — ссылающиеся на концепт
- out_links — на которые концепт ссылается

$$PMI(L_1, L_2) = \frac{|W||L_1 \cap L_2|}{|L_1||L_2|}, PMI' = \frac{PMI}{1 + PMI}$$

Шаблон: $agg_{g \in G}\ \mathbb{1} \cdot sim(links(e), links(g))$

Ранжирование концептов термина

Признаки на основе глобального контекста

Шаблон:

$agg_{g \in G}\ \mathbb{1} \cdot sim(links(e), links(g))$

Пример:

$\max_{g \in G} \mathbb{1}_{e-g} \cdot PMI'(in\_links(e), in\_links(g))$

Дополнительные признаки:

- $max_{g \in G}\ \mathbb{1}_{e \leftrightarrow g}$
- $avg_{g \in G}\ \mathbb{1}_{e \leftrightarrow g}$

Фильтрация потенциально неверных результатов ранжирования

Фильтрация потенциально неверных результатов ранжирования

Признаки

- все признаки, использованные для ранжирования;
- разница оценки между первым и вторым концептом при ранжировании;
- энтропия для значений термина

$$F_{entropy}(m) = -\sum_{e \in E_m} commonness(e, m)\ log\ commonness(e, m)$$

- бинарный признак: термин является именованной сущностью;
- вероятность ссылки — $link\_probability(m)$
- оценка Гуда-Тьюринга того, что правильного значения нет в Википедии:

$$F_{GoodTuring}(m) = \frac{\sum_{e \in E_m} \mathbb{1}_{count(e,m)=1}}{\sum_{e \in E_m} count(e, m)},$$

$count(e, m)$ — сколько раз термин m ссылается на концепт e.

GLOW + интеллектуальный контекст

GLOW + интеллектуальный контекст

Идея

рассматриваем задачу получения контекста (а в идеале - всю задачу викификации) как задачу машинного перевода с естественного языка на "язык концептов"

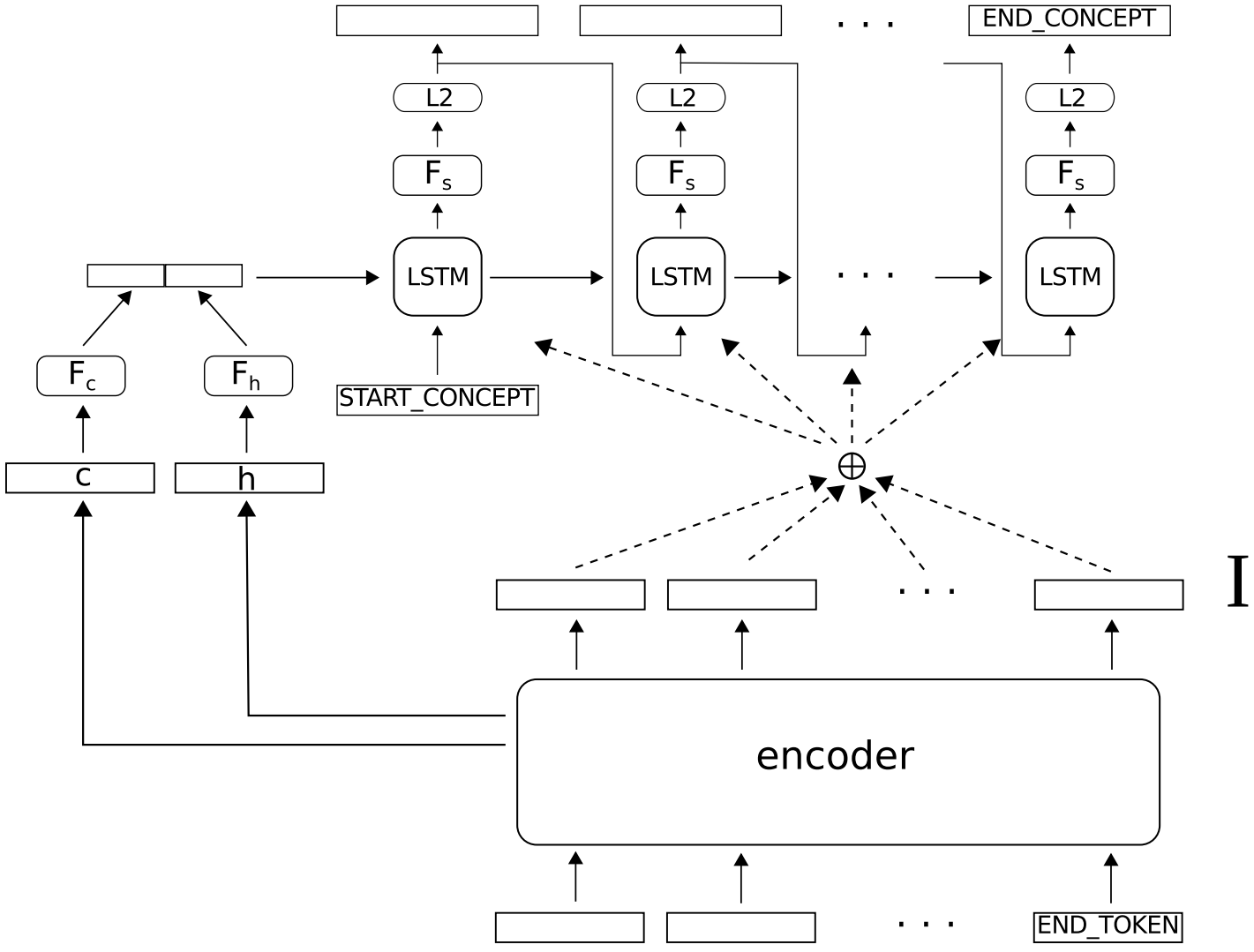
Интеллектуальный контекст

контекст создается с помощью нейронной сети на основе архитектуры кодировщик-декодировщик

A. Sysoev and I. Nikishina
Smart Context Generation for Disambiguation to Wikipedia

GLOW + интеллектуальный контекст

Общая архитектура



GLOW + интеллектуальный контекст

Векторное представление концептов

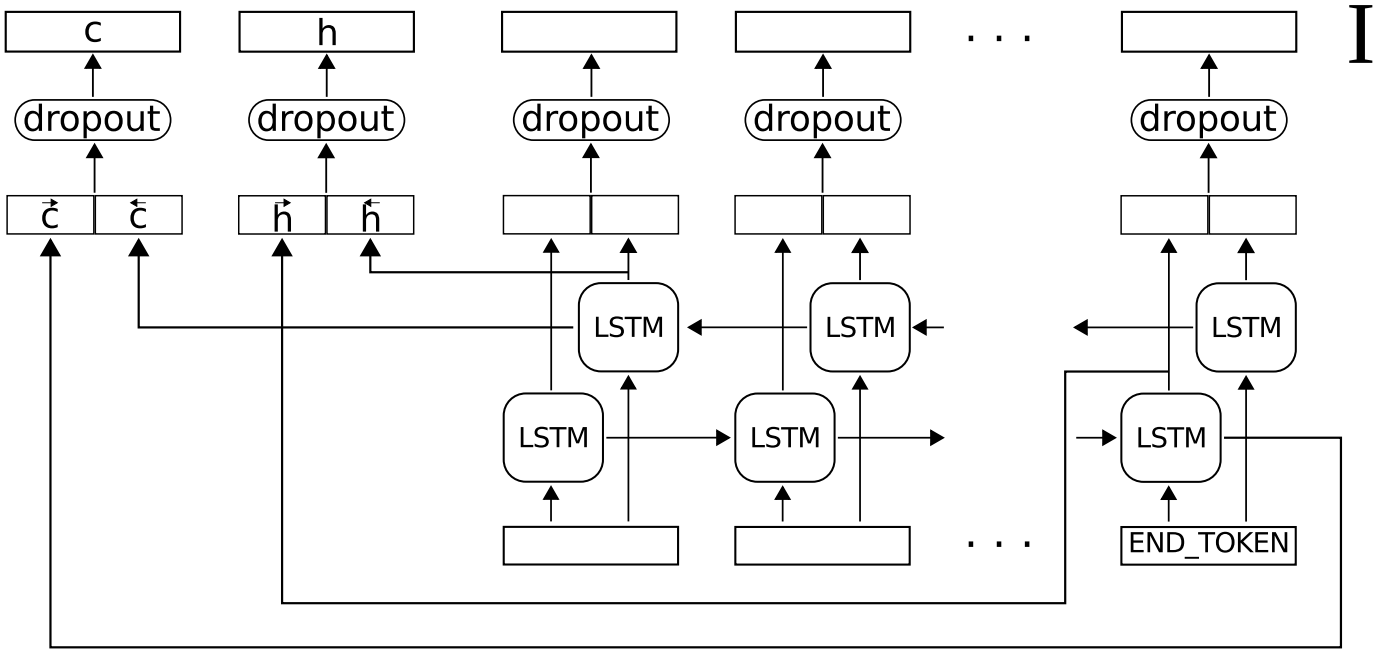
Дзэн, дзен — одна из важнейших школ китайского и всего восточно-азиатского буддизма, окончательно сформировавшаяся в Китае в V—VI веках под большим влиянием даосизма и являющаяся доминирующей монашеской формой буддизма Махаяны в Китае, Вьетнаме и Корее.

word2vec по "токенам":

Буддизм_в_Китае_#101067
Буддизм#306
Китайская_Народная_Республика#1134
Даосизм#469
Махаяна#9644
Китайская_Народная_Республика#1134
Вьетнам#4056
Корея#7472

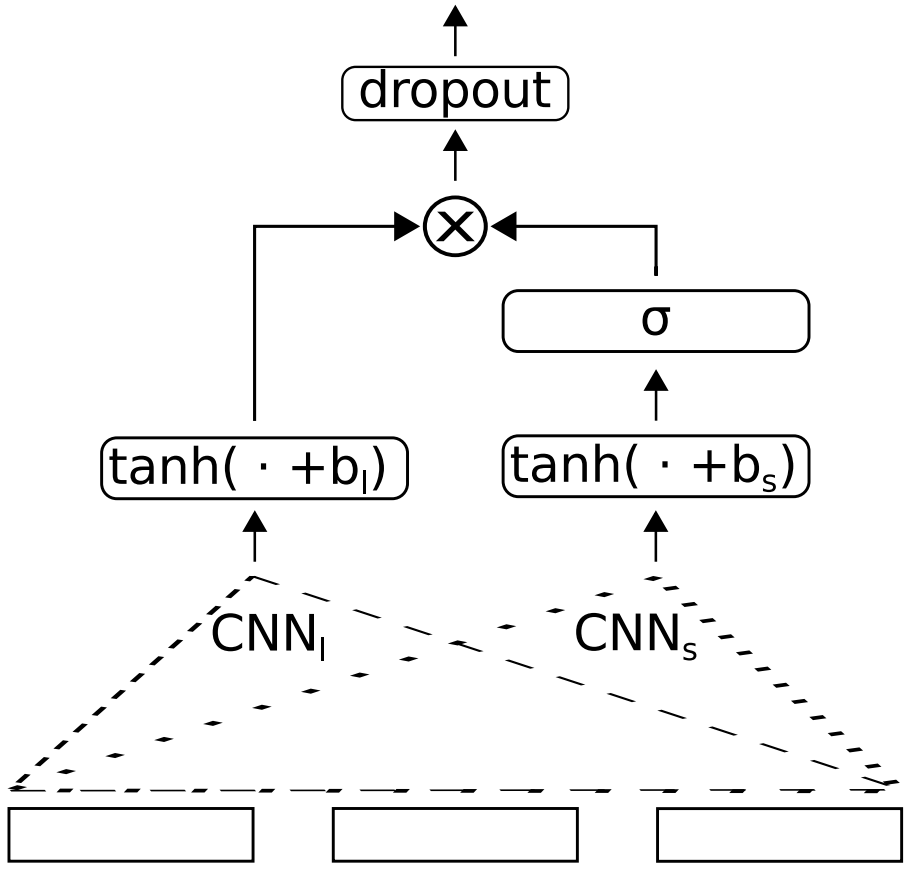
GLOW + интеллектуальный контекст

BiLSTM-кодировщик



GLOW + интеллектуальный контекст

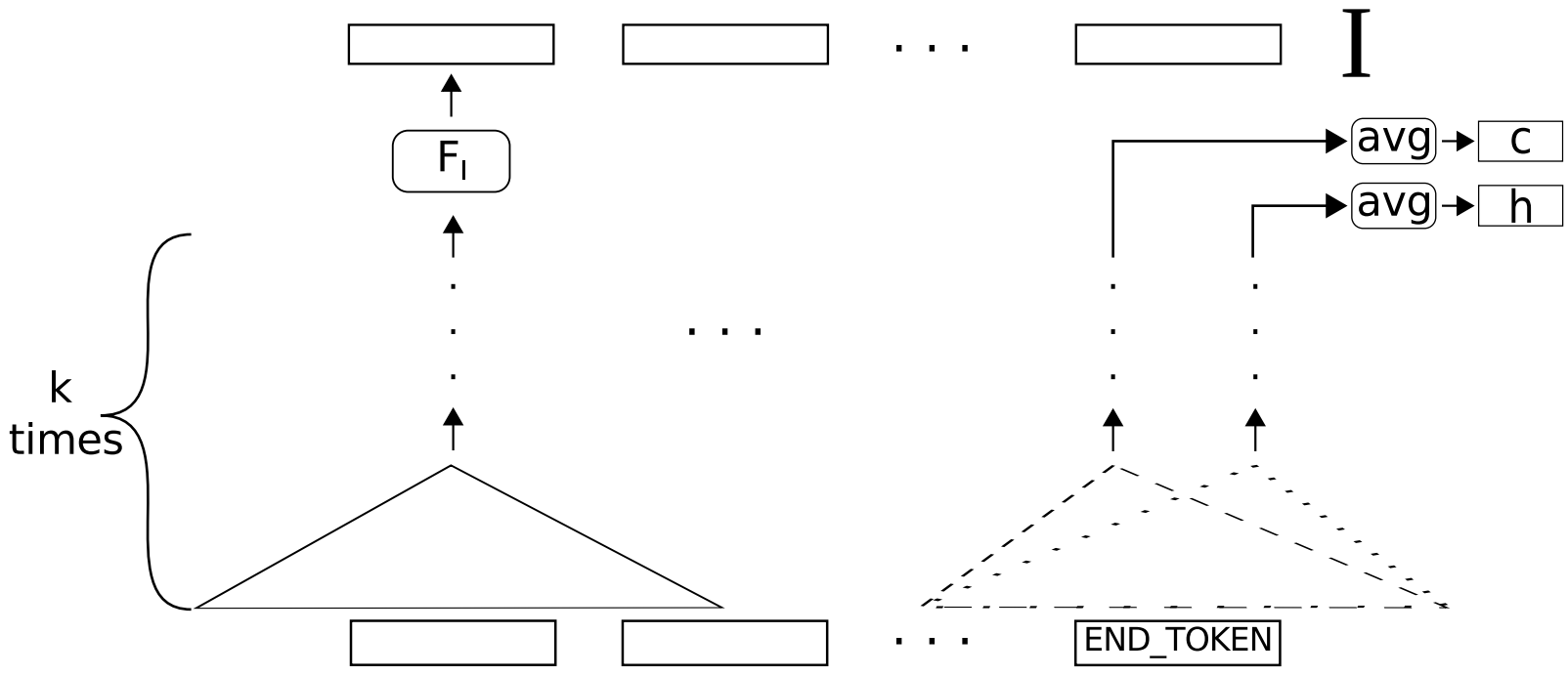
Gated unit



$$tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

GLOW + интеллектуальный контекст

CNN-кодировщик



GLOW + интеллектуальный контекст

Близость к контексту

$$F_S^{max}(e, m) = \max_{s \in S} cos(embedding(e), s)$$

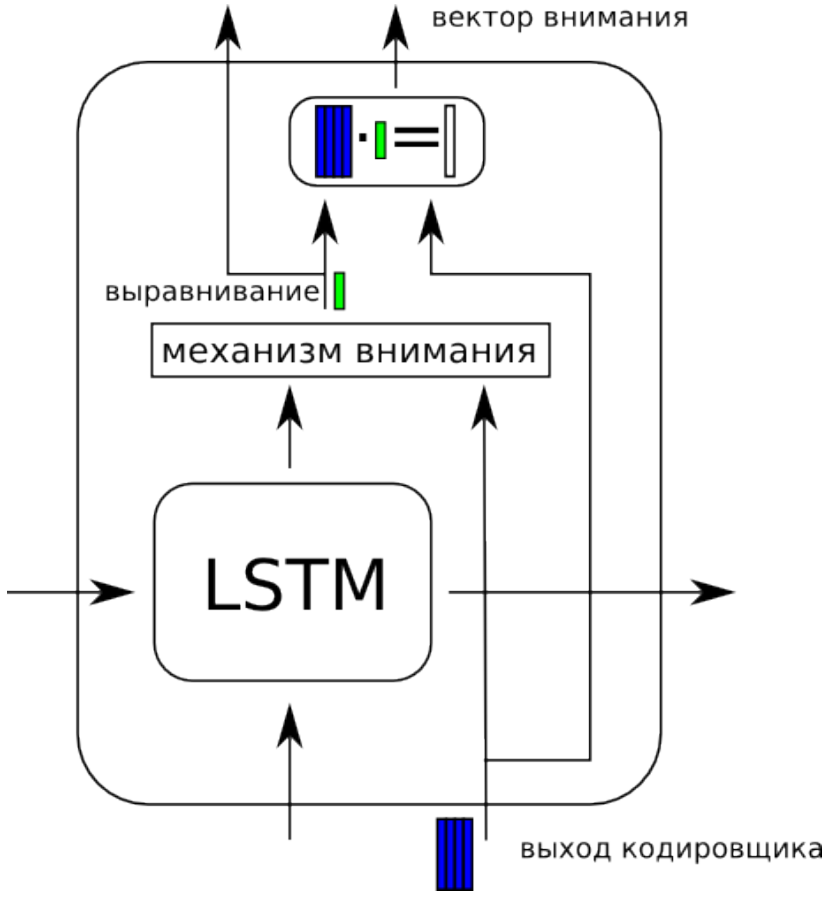
$$F_S^{avg}(e, m) = \frac{1}{|S|} \sum_{s \in S} cos(embedding(e), s)$$

e — концепт, одно из возможных значений термина m

S — интеллектуальный контекст

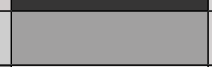
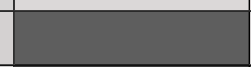
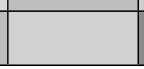
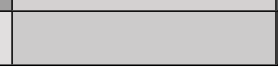
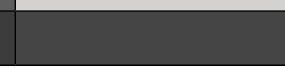
GLOW + интеллектуальный контекст

Внимание в декодировщике



GLOW + интеллектуальный контекст

Выравнивания

В	1974	году	Пирсигу	присуждена	стипендия	Гуггенхайма	
							0.17 0.21 -0.33 ...
							-0.53 0.42 0.87 ...
							0.09 0.13 -0.05 ...

GLOW + интеллектуальный контекст

Близость к контексту

$$F_S^{attention_{max}}(e, m) = \max_{s \in S} \frac{\alpha(s, m)}{\sum_{s \in S} \alpha(s, m)} cos(embedding(e), s),$$

$$F_S^{attention_{avg}}(e, m) = \frac{1}{\sum_{s \in S} \alpha(s, m)} \sum_{s \in S} \alpha(s, m) cos(embedding(e), s),$$

$$\alpha(s, m) = \sum_{t:\ t\ intersects\ m} alignment(t, s),$$

e — концепт, одно из возможных значений термина m

S — интеллектуальный контекст

t — токен

$alignment(t, s)$ — вес токена t в выравнивании при генерации векторного представления концепта s

Оценка качества

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

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

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

Варианты:

- должны совпадать и границы терминов, и присвоенные им концепты;
- должны совпадать только концепты, а границы полностью игнорируются (т.е. документ преобразуется в множество концептов).

Следующая лекция

Перенос знаний, совместное обучение

лектор: Андрианов Иван Алексеевич