Как и зачем мы делаем Открытый корпус

В. В. Бочаров Д. В. Грановский
Mathlingvo

14 мая 2011 г.

Жизненный цикл текста

1. Исходный текст
   - под лицензией, совместимой с CC-BY-SA
   - проходит вычитку
   - делится на абзацы, предложения и токены

Жизненный цикл текста

1. Исходный текст
   - под лицензией, совместимой с CC-BY-SA
   - проходит вычитку
   - делится на абзацы, предложения и токены
2. Морфологические интерпретации
   - словарь на базе словаря проекта АОТ
   - но морфологический стандарт — свой
   - генерируются все возможные гипотезы

Жизненный цикл текста

1. Исходный текст
   - под лицензией, совместимой с CC-BY-SA
   - проходит вычитку
   - делится на абзацы, предложения и токены
2. Морфологические интерпретации
   - словарь на базе словаря проекта АОТ
   - но морфологический стандарт — свой
   - генерируются все возможные гипотезы
3. Полуавтоматика (сейчас её нет)
   - привязка к словарю на основе эвристик
   - снятие простой неоднозначности

Жизненный цикл текста

1. Исходный текст
   - под лицензией, совместимой с CC-BY-SA
   - проходит вычитку
   - делится на абзацы, предложения и токены
2. Морфологические интерпретации
   - словарь на базе словаря проекта АОТ
   - но морфологический стандарт — свой
   - генерируются все возможные гипотезы
3. Полуавтоматика (сейчас её нет)
   - привязка к словарю на основе эвристик
   - снятие простой неоднозначности
4. Ручное снятие неоднозначности пользователями

Жизненный цикл текста

1. Исходный текст
   - под лицензией, совместимой с CC-BY-SA
   - проходит вычитку
   - делится на абзацы, предложения и токены
2. Морфологические интерпретации
   - словарь на базе словаря проекта АОТ
   - но морфологический стандарт — свой
   - генерируются все возможные гипотезы
3. Полуавтоматика (сейчас её нет)
   - привязка к словарю на основе эвристик
   - снятие простой неоднозначности
4. Ручное снятие неоднозначности пользователями
5. Разметка доступна для просмотра и скачивания

Уровни текста

Концептуальные уровни:

1. Графематика
2. Морфология
3. Синтаксис (в отдаленных планах)
4. Семантика (в совсем отдаленных планах)
5. Something else?

Уровни текста

Концептуальные уровни:

1. Графематика
2. Морфология
3. Синтаксис (в отдаленных планах)
4. Семантика (в совсем отдаленных планах)
5. Something else?

4 иерархических уровня деления (это графематика):

1. Текст
2. Абзац
3. Предложение
4. Токен*

* некоторая последовательность символов без пробелов

Токенизация

Как разделить текст на эти единицы?

- на абзацы – взять из источника

Токенизация

Как разделить текст на эти единицы?

- на абзацы – взять из источника
- на предложения – пока вручную

Токенизация

Как разделить текст на эти единицы?

- на абзацы – взять из источника
- на предложения – пока вручную
- на токены – полуавтоматически

Токенизация

Токенизация должна быть:

- единообразной
- удобной для морфологии

Проблемы ручной токенизации:

- очень трудоемко
- трудно обеспечить единообразие
- не все отличия видны глазами

Токенизация-2

Используем простое машинное обучение:

- корпус предложений, уже разделенных на токены (внутри текста расставлены границы)

Токенизация-2

Используем простое машинное обучение:

- корпус предложений, уже разделенных на токены (внутри текста расставлены границы)
- набор бинарных характеристических функций (15 шт.)
  F1 = «является ли данный символ пробелом»
  ...
  F7 = «является ли данный символ буквой кириллицы»
  ...
  F15 = «является ли цепочка символов от ближайшего пробела слева до ближайшего пробела справа словарным словом»

Токенизация-2

Используем простое машинное обучение:

- корпус предложений, уже разделенных на токены (внутри текста расставлены границы)
- набор бинарных характеристических функций (15 шт.)
  F1 = «является ли данный символ пробелом»
  ...
  F7 = «является ли данный символ буквой кириллицы»
  ...
  F15 = «является ли цепочка символов от ближайшего пробела слева до ближайшего пробела справа словарным словом»
- вычисляем все эти функции для каждой позиции в предложении

Токенизация-3

Используем простое машинное обучение:

- для каждой позиции получается двоичный вектор
  Позиция 1: 001000010000000
  Позиция 2: 100000100000010
  ...

Токенизация-3

Используем простое машинное обучение:

- для каждой позиции получается двоичный вектор
  Позиция 1: 001000010000000
  Позиция 2: 100000100000010
  ...
- для каждой позиции знаем, проходит ли в ней граница токенов

Токенизация-3

Используем простое машинное обучение:

- для каждой позиции получается двоичный вектор
  Позиция 1: 001000010000000
  Позиция 2: 100000100000010
  ...
- для каждой позиции знаем, проходит ли в ней граница токенов
- для каждого двоичного вектора на корпусе вычисляется вероятность того, что в позиции с таким вектором есть граница токенов

Токенизация-3

Используем простое машинное обучение:

- для каждой позиции получается двоичный вектор
  Позиция 1: 001000010000000
  Позиция 2: 100000100000010
  ...
- для каждой позиции знаем, проходит ли в ней граница токенов
- для каждого двоичного вектора на корпусе вычисляется вероятность того, что в позиции с таким вектором есть граница токенов
- в реальном тексте в каждой позиции тоже вычисляем вектор и смотрим вероятность

Токенизация-4

Так выглядит обучение:

```
576     0.000   8       0000010010000000
768     0.000   3       0000011000000000
1024    0.026   39      0000100000000000
1030    0.000   2       0000100000000110
1056    0.000   11      0000100000100000
1058    0.000   4       0000100000100010
1152    0.000   2       0000100100000000
1536    0.000   433     0000110000000000
2048    0.366   41      0001000000000000
2056    1.000   11      0001000000001000
2058    0.000   10      0001000000001010
2064    1.000   27      0001000000010000
2072    1.000   12      0001000000011000
2074    0.000   11      0001000000011010
```

Токенизация-5

Получаемое деление – вероятностное, поэтому его нужно проверять глазами:

I. 1. Федеральные министры и заместители председателя правительства до 1 июля 2011 г. должны быть выведены из состава советов директоров (наблюдательных советов) 17 компаний с государственным участием для улучшения инвестиционного климата в России . Полный перечень компаний опубликован 2 апреля на официальном сайте президента РФ .
внести исправления

Токенизация-5

Получаемое деление – вероятностное, поэтому его нужно проверять глазами:

I. 1. Федеральные министры и заместители председателя правительства до 1 июля 2011 г. должны быть выведены из состава советов директоров (наблюдательных советов) 17 компаний с государственным участием для улучшения инвестиционного климата в России . Полный перечень компаний опубликован 2 апреля на официальном сайте президента РФ .
внести исправления

1. Основную сложность представляют слова , которые пишутся через дефис ; в каких-то случаях дефис по-хорошему нужно оставлять внутри токена , в некоторых непонятно (а они-то самые частые) , а в случаях-когда -автор -специально -пишет -много -слов -через -дефис и вовсе не надо ; еще мы очень любим формулы , типа C 2 H 5 O H , и имена собственные (Яндекс . Деньги , АК - 47) .

Вопросы про токенизацию?

Морфология

Суть морфологического уровня:

- связать токен с морфологическим словарем
- или обозначить, что токен не является словом

Морфология

Суть морфологического уровня:

- связать токен с морфологическим словарем
- или обозначить, что токен не является словом

Зачем нужен морфологический словарь?

- можно изменить конкретный разбор конкретной словоформы во всем корпусе сразу
- легче находить опечатки
- в будущем можно будет добавлять лексико-семантическую информацию, почти не меняя разметку

Морфология-2

- за основу взят словарь группы АОТ

Морфология-2

- за основу взят словарь группы АОТ
- описание слова = лемма + набор форм (парадигма) + набор граммем леммы

Морфология-2

- за основу взят словарь группы АОТ
- описание слова = лемма + набор форм (парадигма) + набор граммем леммы
- описание формы = текст + набор граммем формы

Морфология-2

- за основу взят словарь группы АОТ
- описание слова = лемма + набор форм (парадигма) + набор граммем леммы
- описание формы = текст + набор граммем формы
- леммы связаны между собой связями

Морфология-2

- за основу взят словарь группы АОТ
- описание слова = лемма + набор форм (парадигма) + набор граммем леммы
- описание формы = текст + набор граммем формы
- леммы связаны между собой связями
- словарь можно редактировать

Морфология-2

- за основу взят словарь группы АОТ
- описание слова = лемма + набор форм (парадигма) + набор граммем леммы
- описание формы = текст + набор граммем формы
- леммы связаны между собой связями
- словарь можно редактировать
- сочетаемость граммем регулируется моделью морфологии, которая выражена в виде набора правил

Морфология-2

- за основу взят словарь группы АОТ
- описание слова = лемма + набор форм (парадигма) + набор граммем леммы
- описание формы = текст + набор граммем формы
- леммы связаны между собой связями
- словарь можно редактировать
- сочетаемость граммем регулируется моделью морфологии, которая выражена в виде набора правил
- каждое правило имеет вид:
  «Если у [леммы/формы] есть граммема А, то у [леммы/формы] [должна быть/не должно быть/может быть] граммема Б»

Морфология-3

Модель нужна, чтобы отслеживать ошибки, присутствующие в словаре изначально или вносимые редакторами.
Примеры правил:

- NOUN -> NMbr (лемма -> форма, обязательно)
- VERB -> ASpc (лемма -> лемма, обязательно)
- indc -> TEns (форма -> форма, обязательно)
- VERB -> Impe (лемма -> лемма, возможно)
- Impe -> PErs (лемма -> форма, запрещено)

Всего сейчас 107 граммем и 127 правил + 218 автоматически выведенных.

Морфология-4

Итого в словаре бывает 5 типов ошибок:

1. неизвестная граммема
2. несовместимые граммемы
3. явно не разрешенная правилами граммема
4. отсутствует обязательная граммема
5. две формы в рамках парадигмы имеют полностью совпадающие наборы граммем

Вопросы про морфологию?

Разрешение неоднозначности

2 этапа: (полу)автоматический (сейчас нет), ручной.

Разрешение неоднозначности

2 этапа: (полу)автоматический (сейчас нет), ручной.
Ручное разрешение морфологической неоднозначности – основная задача, для которой мы хотим привлекать пользователей-разметчиков.

Разрешение неоднозначности

2 этапа: (полу)автоматический (сейчас нет), ручной.
Ручное разрешение морфологической неоднозначности – основная задача, для которой мы хотим привлекать пользователей-разметчиков.
От разметчика требуется:

- исключить неверные разборы, в идеальном случае – выбрать один
- или отметить, что верный разбор отсутствует

Интерфейс разрешения неоднозначности

(Here be live demonstration)

Contacts

Берем студентов на практику!

http://opencorpora.org

granovsky@opencorpora.org
bocharov@opencorpora.org