Проектная работа «Крепкие соединения: как мы извлекаем коллокации»

Общий план работы

Парамонова Дарья
dashparamonova@yandex.ru
МГУ им. Ломоносова, ОТиПЛ

15.10.2022 ОЦ «Сириус»

Чем займемся?

I. Подготовим корпус, из которого в дальнейшем будем извлекать коллокации:
   (1) В качестве основы возьмем новостной подкорпус «Тайги»;
   (2) Освоим базовые навыки программирования на ЯП Python и поработаем с т.н. «регулярными выражениями»;
   (3) Разделим новостные тексты на предложения, удалим пунктуацию и приведем к нижнему регистру.
II. Выберем т.н. «node-word», для которого будем извлекать коллокации;
III. Выделим коллокации на основе *поверхностной встречаемости* (surface cooccurrence);
IV. Для каждой гипотетической коллокации рассчитаем частотные характеристики (frequency signatures): O, f_1, f_2, N, а также значение ожидаемой частотности (expected frequency): E;
V. Для каждой гипотетической коллокации рассчитаем ее ассоциативные метрики (association measures): MI, MI^k, local-MI, z-score, t-score, simple-ll и др.;
VI. Оценим приемлемость полученных коллокаций и проанализируем все статистические данные;
VII. Установим, какая из ассоциативных метрик наилучшим образом справилась с извлечением коллокаций.

(Национальный) Корпус (Русского Языка)

Парамонова Дарья
dashparamonova@yandex.ru
МГУ им. Ломоносова, ОТиПЛ

15.10.2022 ОЦ «Сириус»

Что такое корпус текстов и зачем он нужен?

- В лингвистике корпус — подобранная и обработанная по определённым правилам совокупность текстов, используемых в качестве базы для исследования языка. Они используются для статистического анализа и проверки статистических гипотез, подтверждения лингвистических правил в данном языке. Корпус текстов является предметом исследования корпусной лингвистики.

- Корпус — основное понятие и база данных корпусной лингвистики. Анализ и обработка разных типов корпусов являются предметом большинства работ в области компьютерной лингвистики (например, извлечение ключевых слов), распознавания речи и машинного перевода. Корпусы и частотные словари могут быть полезны в обучении иностранным языкам.

Какие у корпуса свойства?

Среди множества определений корпуса можно выделить его главные **свойства**:

- электронный — в современном понимании корпус должен быть в электронном виде;
- репрезентативный — должен хорошо «представлять» объект, который моделирует;
- размеченный — главное отличие корпуса от коллекции текстов;
- прагматически ориентированный — должен быть создан под определённую задачу.

Как классифицируются корпусы?

По критерию ***параллельности***, например, корпусы можно разделить на одноязычные, двуязычные и многоязычные. Многоязычные и двуязычные делятся на два типа:

- параллельные — множество текстов и их переводов на один или несколько языков.

- сопоставимые (псевдопараллельные) — оригинальные тексты на двух или нескольких языках.

Как размечают корпусы?

Разметка заключается в приписывании текстам и их компонентам специальных **тегов**: лингвистических и внешних (экстралингвистических). Выделяют следующие лингвистические типы разметки: морфологическая, семантическая, синтаксическая, анафорическая, просодическая, дискурсная и т. д. К некоторым корпусам применяются дальнейшие структурные уровни анализа. В частности, некоторые небольшие корпусы могут быть полностью синтаксически размечены. Такие корпусы обычно называют *глубоко аннотированными* или *синтаксическими*.

На данный момент в открытом доступе представлены различные программные средства для разметки корпусов. Условно их можно разделить на **обособленные (stand-alone)** и **веб-ориентированные (web-based)**.

Что такое НКРЯ?

Национальный корпус русского языка (НКРЯ) — доступный для поиска электронный онлайн-корпус русских текстов. Открыт 29 апреля 2004 года. Также доступен для поиска исторический корпус церковнославянских, древнерусских (XI—XIV века) и среднерусских (XV — начало XVIII века) текстов.

В корпус входят как письменные тексты, так и записи устных текстов. В корпус также входят подкорпусы поэтических и диалектных текстов, корпусы параллельных текстов, отдельный газетный корпус, церковнославянский корпус, исторический, синтаксический, акцентологический, мультимедийный и обучающий подкорпусы.

С 2010 года в составе исторического подкорпуса Национального корпуса русского языка доступен текстовый корпус берестяных грамот с полной морфологической разметкой

Какие бывают морфологические анализаторы?

- pymorphy2

  Он умеет (при работе использует словарь OpenCorpora):
  - приводить слово к нормальной форме (например, “люди -> человек”, или “гулял -> гулять”).
  - ставить слово в нужную форму. Например, ставить слово во множественное число, менять падеж слова и т.д.
  - возвращать грамматическую информацию о слове (число, род, падеж, часть речи и т.д.);

- mystem — морфологический анализатор русского языка с поддержкой снятия морфологической неоднозначности;

- NLTK (Natural Language ToolKit) — пакет библиотек и программ для символьной и статистической обработки естественного языка, написанных на языке программирования Python. Содержит графические представления и примеры данных.