Named Entity Recognition

November 12, 2012

Команда: apelsin

Состав: Хайруллин Егор, Синицин Филлип, Бурмистров Михаил

Github: https://github.com/discobot/NER-MEHMM

Spanish - это мы же, просто описались пару раз. Orange - это не мы. Мандарин - это тоже не мы. Это подражатели. Или фанаты.

Основной язык: испанский. Что происходит в голландском - без понятия, но оно работает.

Вначале скрипт запускался с параметрам -g 2.0. Базовый результат был плохой - 8%. Перечислю наши основные этапы в увеличении качества:

- Добавили униграммы - слова, которые часто встречаются перед NE. Стали считать, что NE - это, обычно, слово с большой буквы. Получили качество около 24%.
- Добавили мега фичу: “слово начинается с маленькой буквы” - качество около 47%. Стоит отметить, что замена этой фичи на противоположную убивало результат до 10% :-)
- Узнали, что -g лучше выставить дефолтным. Получили качество 53%.

Дальше начинается самый сложный этап с структуризацией кода, более аккуратным разбором всех случаев и, конечно, с любимым методом Егора: посмотреть глазами на выходные данные.

Структура кода получилась следующая:

1. Выделение фич, касающихся самого слова.
2. Разбор случая - слово с маленькой буквы (оно может быть частью NE)
3. Разбор случая - слово с большой буквы и вначале предложения
4. Разбор случая - слово с большой буквы и в середине предложения после O
5. Разбор случая - слово с большой буквы после другого NE.

В каждом случае выдаются свои фичи. Максимально простые фичи с максимально сложными и узкими условиями.

Основное что, было сделано для подъема качества с 53% до 72%:

- Реструктуризация, исправление ошибок, упрощение некоторых фич.
- Больший учет контекста для первого слова в предложении: какие слова идут за ним и т.п.
- Частичный учет слов с маленькой буквы внутри NE (их, иногда, может быть два подряд). Рассмотрены не все случаи, увы. Трейдофф: лжесрабатывания, склейка NE когда не нужно.
- Более сильный учет контекста для NE (с большой буквы ли следующее слово, какие части речи у соседей и т.п.) и попытка выделения таких знаковых слов, как "General" и т.п. Учет слов и частей речи до -4 влево. Забавно, но слова на расстоянии 3 и 4 влево стоит "пихать" в одинаковые фичи. Подозреваю, что из-за различных артиклей они могут сдвигаться туда-сюда и важно "примерное" расстояние до этого слова. Такой же учет, но вправо, увы, не работает для испанского.
- Снижение $cutoff$ до 2. Это дало +4%, так как редкие фичи с редкими словами начали все-таки срабатывать.
- Снижение $cutoff$ до 1. Это дало скачок еще на 4% до 71%+. Снижение до 0, увы, ничего не дало :-)
- Прыжок до 72% дало убирание плохих, как выяснилось, фич, которые были добавлены сразу пачкой и не проверены по одиночке.

Основная проблема - это неправильная классификация NE и слияние разных NE. Все-таки, навесить сверху еще какой-нибудь адабуст на решающих деревьях было бы неплохо.

К сожалению, не получилось вытащить какие-либо характерные суффиксы или префиксы. Зато тэги - слабо, но помогали при классификации NE.

В принципе, весь код "мультиязычный" и на голландском также работает. Итоговая версия дает не особо хороший результат - всего 62%, но до "сильной подгонки" под испанский для достижения 72% было 65+%.

Запуск: ./run.sh -f data/spanish.train.txt -c 1 -i 10000 -m spanish -T