Проверка согласования глагола с подлежащим.

Антон Бузанов, лаборатория учебных корпусов НИУ ВШЭ

1. Описание модели

Модель должна отлавливать ошибки в согласовании глагола с подлежащим(и), то есть, на самом деле, проверять правильность использования или неиспользования окончание *-s* на глаголах в таких предложениях, как:

> Thus, this method work with different parts of speech improperly.
> But the problems of transport always stays the same - for example, as traffic congestions, different kinds of pollution.

В первом примере при подлежащем *this method* использована форма глагола *work*, а во втором примере, наоборот: подлежащее – *problems*, а сказуемое – *stays*.

2. Использование синтаксического парсера

Для корректной работы модели предложения необходимо разметить с помощью зависимостного парсера. Эта модель использует синтаксическую разметку от spaCy. Программа работает на базе самой новой модели – **en_core_web_lg**, эта модель натренирована на OntoNotes с использование GloVe векторов, натренированных на Common Crawl.

3. Работа модели

После того, как spaCy разметил текст, очищенный от некоторых знаков препинания и нормализованный, он выдаёт conll-файл, с которым и производится дальнейшая работа.

В первую очередь conll-файл считывается и конвертируется в словарь Python. Далее программа находит слово, помеченное тэгом ROOT, и принимает его за главный предикат. После этого ищутся слова с тэгами ‘conj’ или ‘ccomp’, которые зависят от главного предиката, а также глаголы, помеченные как ‘acl’ (таким образом, находятся все остальные предикаты).

Следующие шаги проводятся для каждого предиката.

Проверяется, не является ли он одним из тех, что имеют одинаковую форму в прошедшем и в настоящем времени, а значит всегда есть вероятность, что использованное автором время – Past – и тогда отсутствие согласования – это не ошибка.

Если от глагола зависит вспомогательный, то согласование проверяется по нему.

Следующий шаг – поиск субъекта(ов). Проверяются все зависимые от сказуемого ‘nsubj’ и ‘csubj’. Если это не прилагательные, которые иногда ошибочно размечаются, то программа будет проверять согласование по ним.

Если от сказуемого зависит слово с тэгом ‘expl’ (то есть конструкция с *there BE*), то согласование с предикатом проверяется по словам с тэгом ‘attr’.

Каждое подлежащее проверятся на принадлежность к классу существительных, имеющих одинаковые формы единственного и множественного числа (*deer, sheep*). Кроме того, проверяется, относится ли слово к классу собирательных (*police, government*).

Следующий этап – устранение проблем со сложными подлежащими (представляющими не одно слово).

В первую очередь, считаем число в конструкциях типа *a lot of X, plenty of X* как число *X*. Дальше проверяем конструкции с *a number* и *the number*, которые требуют согласование по множественному и по единственному числу, соответственно.

Дальше проверяется, является ли подлежащее сочинённым. Если оно сочинено союзом *and,* то число подлежащего – множественное. Если – *or*, то – число последнего члена однородного ряда. Если связь между подлежащими бессоюзная, то программа считает количество зависимых подлежащих от предиката, и если их больше одного – то число снова множественное.

На каждом из этапов определяется модель согласования личных местоимений: так, всё, что не *he, she* или *it*, должно согласовываться по типу «множественного числа», то есть не лицензировать появление *-s*.

Последнее уточнение, которое делает программа – она проверяет, не находится ли субъект в кавычках (то есть не является ли подлежащее названием, например: *“Three men in a boat…”*), и если так – то согласование будет по единственному числу.

В конце получив число субъекта, программа сравнивает его с согласованием на глаголе и если находит ошибку, то добавляет предложение в список ошибок.

4. Результаты

Сейчас приблизительная точность модели составляет более 70%. Ошибки случаются из-за мисспеллинга (и эту часть планируется исправить программой-спеллчекером), из-за ошибок таггера и некоторых недоработок со стороны создателя модели. Новые случаи ошибок были обговорены на предыдущих семинарах лаборатории и в индивидуальном порядке. Также будут добавлены новые триггеры, которые позволят отловить большее количество предложений.