Работа была выполнена Денисом Рахманом и Анастасией Пискуновой

1. Англоязычный текст, для которого выделялись ключевые слова, представляет собой отрывок из статьи Википедии ‘cat’. Ниже приведен результат выделения ключевых слов вручную (GS - Golden Standart).

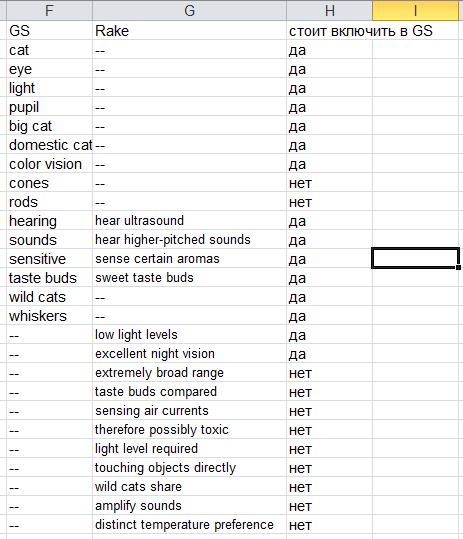


Доля ключевых слов, помеченных соответствующим образом обоими аннотаторами, составляет 0,8125 от общего числа слов, помеченных как ключевые хотя бы одним аннтотатором.

2. На тексте был запущен обычный алгоритм Rake. Вот список ключевых выражений, выделенных им:

hear higher-pitched sounds   
sense certain aromas   
distinct temperature preference   
sensing air currents   
therefore possibly toxic   
sweet taste buds   
taste buds compared   
low light levels   
light level required   
touching objects directly   
excellent night vision   
extremely broad range   
wild cats share   
hear ultrasound   
amplify sounds

Как несложно заметить, наивный алгоритм Rake выдает чрезвычайно отличные от золотого стандарта результаты. Даже если считать верно выделенными такие нулевые выражения, которые хотя бы приблизительно совпадают с одним из ключевых слов золотого стандарта, различие все равно остается весьма значительным. Ниже приведена таблица, сравнивающая выдачу rake и золотой стандарт:



При подобном сопоставлении доля совпадающих решений золотого стандарта и Rake составила около 15%. При этом в подавляющем большинстве случаев выделенные алгоритмом ключевые сочетания хуже (на наш взгляд).

3. Далее алгоритм Rake был улучшен:

* Было введено условие, отбрасывающее потенциальные ключевые выражения, встретившиеся в тексте менее 2х раз. Это позволило существенно сузить круг потенциальных ключевых выражений и оставить только самые частотные.
* Была проведена лемматизация с помощью WordNetLemmatizer. Это позволило рассматривать различные словоформы одной лексемы (напр. *cat* и *cats*) как одно потенциаольное ключевое выражение.

Ниже приведены результаты улучшенного алгоритма в сопоставлении с золотым стандартом.



Как можно заметить, соответствия между ключевыми словами, выделяемыми Rake и золотым стандартом, стали как намного более точными, так и намного более регулярными (совпадение около 35%). Из этого можно сделать вывод, что предложенные нами изменения значительно улучшают качество работы алгоритма.

3. Наивный алгоритм Rake был также применен к русскоязычному тексту сопоставимого размера и содержания, представляющему собой отрывок из статьи Википедии ‘Кошка’. Вот результаты работы алгоритма:

темноте светятся жёлто-зелёным   
котята рождаются голубоглазыми   
условиях слабого освещения   
слабо пигментированным тапетумом   
имеющий максимум поглощения   
вертикально-овальный вплоть   
зрительные поля перекрываются   
это позволяет предположить   
угла падения света   
оттенок меняется градиентом   
пигментацией радужной оболочки   
пигментация тапетума снижается   
тапетума безвозвратно терялась   
уменьшить световой поток   
радужная оболочка сжимается

Легко заметить, что качество работы наивного Rake для русского языка намного ниже, чем для английского -- многие ключевые выражения на самом деле не являются хоть сколько-нибудь синтаксически автономными единицами. Мы думаем, что лемматизация текста для русского языка окажется еще полезнее, чем для английского, так как морфологическая система последнего намного беднее и, следовательно, словоформы лексемы будут с большей вероятностью совпадать вне зависимости от своих грамматических характеристик. Введение минимальной встречаемости ключевого выражения для русского языка также представляется осмысленным.

Таким образом, мы предлагаем использовать для улучшения работы алгоритма Rake на русских текстах те же шаги, что и для улучшения его работы на английских.