АВТОМАТИЧЕСКАЯ ИДЕНТИФИКАЦИЯ ЯЗЫКА ДОКУМЕНТА ДЛЯ ПОСЛЕДУЮЩЕГО CROSS-LANGUAGE АНАЛИЗА

Бредихин Юрий Алексеевич
магистрант, кафедра информатики КСиС БГУИР,
Беларусь, г.Минск
E-mail: bredihinu@gmail.com

Калугина Марина Алексеевна
канд. физ.-мат. наук, доцент БГУИР
Беларусь, г.Минск

Введение

Определение языка является важной стадией работы с текстовыми документами, так как от нее зависит эффективность приложений по обработке естественных языков (NLP): информационный поиск (IR), вопросно-ответные системы (QA), автоматическое реферирование (Summarization).

В зависимости от применяемых правил построения поисковых образов и стратегий их сравнения различают следующие основные методы определения языка текстовых документов:

- коротких слов,
- частотных слов,
- N-грамм,
- статистический,
- алфавитный,
- грамматических слов,
- неграмматических слов.

Алгоритм N-грамм

Алгоритм N-грамм для определения языка и кодировки документа по его содержимому, основывается на статистиках документов, для которых язык и кодировка известны заранее. В данном методе подсчитываются частоты N-грамм (сочетаний символов или подстрок, длиной не более N) и предполагается, что примерно 300 самых часто используемых N-грамм сильно зависят от языка.

После этого среди всех тестовых документов находим тот, для которого расстояние от его N-граммной статистики до статистики тестируемого документа минимально. После этого языком тестируемого документа считается язык найденного тестового документа.

Расстояние между статистиками подсчитывается следующим образом: все N-граммы сортируются в порядке убывания частоты их появления, затем для каждой N-граммы вычисляется разница её позиций в отсортированном списке N-грамм тестового и тестируемого документов. Расстояние между статистиками определяется как сумма разностей позиций каждой N-граммы:$l = \sum_{i=1}^{300} |P_i - P_i^{\sim}|$, где P_i , $P_i^{\sim}$ - позиции i-й N-граммы в текстовом и тестируемом документах соответственно[1].

Основными недостатками описанного выше алгоритма являются: неустойчивость определения языка малых текстовых фрагментов — отдельные предложения на похожих языках распознаются неуверенно, и часто с ошибками.

Иногда поступают еще проще. Например, в ABBYY RME, как минимум до четвертой версии, морфологическая машина при словарной лемматизации перебирала все загруженные языковые словари.

Следует также отметить, что процесс морфологической обработки является сравнительно трудоемким, сами словари занимают в памяти от единиц до десятков мегабайт на один язык, поэтому перебор даже по десятку возможных языков становится весьма медленным[2].

Для того, чтобы сделать его работу надежной, сделаем алгоритм двух проходным - первый проход выполняется как описано выше, но каждое предложение помечается самым вероятным языком только в случае уверенного распознавания (либо детектируется только один язык с весовым коэффициентом не ниже определенного порога, либо самый вероятный язык опережает следующего претендента не менее, чем в 2 раза). Оставшиеся же предложения подвергаются обработке на втором проходе - так называемой процедуре контрастирования. Она заключается в том, что языком предложения становится неуверенно распознанный язык, если окружающие его принадлежности текста определенному языку, достаточно будет только превышение весовым коэффициентом определенного порога.

Предварительный разбор текстового документа

На этапе предварительного разбора производится разбор корпуса документов и выявление в нем:

1. Уникальных для языка N-грамм с длиной до 3-х символов включительно (включая и одиночные символы).
2. Часто встречающихся (но не уникальных по всему корпусу) в языке N-грамм с длиной до 3-х символов включительно.

Из тонкостей разбора можно отметить следующие моменты:

1. Тексты предварительно разбиваются на слова (монолитные последовательности из символов, соответствующих данному языку).
2. Все слова приводятся к одному регистру (например, к верхнему).
3. Сочетания кодируются следующим образом – если мощность N-граммы меньше, чем 3, то отсутствующие символы заменяются на символ с кодом 0.
4. К словам слева и справа добавляется по пробелу (или любой другой символ, гарантировано не являющийся буквой из какого-то алфавита). Это позволяет в дальнейшем отличить сочетание с мощностью до 2-х включительно, стоящее в конце, от такового же, стоящего в середине и в конце слова. Такая мелочь позволила увеличить точность распознавания языка примерно в 1.5 раза и более[3].
5. Уникальные N-граммы для того, чтобы попасть в окончательный список N-грамм, должны превысить определенный порог встречаемости (на практике использовался уровень в 10-15 раз). Это позволяет убрать случайные сочетания, нехарактерные для данного языка.
6. Часто встречающиеся сочетания сортировались по частоте, и выбиралось определенное их количество - N_{freq}. Эта процедура выполнялась отдельно для сочетаний разной длины. Т.е., при N_{freq} =16, в конечную таблицу сочетаний попадало 16 триграмм, 16 биграмм. Для одиночных символов сделано исключение - для них количество выбирается как меньшее значение из N_{freq}, либо 1/8 от размера алфавита (учитывается количество символов одного регистра). Эксперименты показали, что оптимальное значение N_{freq} (для разных текстов и языковых наборов), лежит, как правило, в промежутке 64.. .128.

Подстройка весовых коэффициентов

Для выявления языка документа, к нему применялся разбор, аналогичный тому, который велся на предварительном этапе, только теперь никакие слова не отфильтровывались, а выявляемые N-граммы сравнивались с таблицей (в практической реализации использовалась структура, известная как хэш-таблица) N-грамм, построенной на предварительном этапе. Если находилось соответствие, то весовой коэффициент языка, которому принадлежала табличная N-грамма увеличивался по следующему правилу.

Если найденная N-грамма являлась уникальной по отношению ко всем остальным языкам, то весовой коэффициент увеличивался на W_U. В текущей реализации W_U=10. Если N-грамма просто являлась часто встречающейся в соответствующем языке, то весовой коэффициент увеличивался на a_N. В текущей реализации было принято, что a_N=N. Данный выбор весовых коэффициентов является эмпирическим, и на практике он показал себя надежно.

Было произведено исследование 3-х вариантов алгоритма оценки текстов:

1. Учитываются все найденные сочетания - как уникальные, так и неуникальные.

2. Учитываются все уникальные сочетания. Неуникальные учитываются только в том случае, если в других языках такие сочетания не прошли в часто встречающиеся (т.е., остались за пределами, которые очерчиваются параметром N_{freq}).

3. Учитываются только уникальные сочетания.

Эксперименты показали, что результаты 1-го алгоритма, учитывающего все, как уникальные, так и часто встречающиеся сочетания, самые неубедительные.

Лучше всего проявляет себя алгоритм 3, который работает только с уникальными сочетаниями. Однако у него имеется один изъян - из-за очень большой требовательности к используемым данным, на коротких последовательностях он может просто не успеть выбрать главный язык. Т.е., если встречается предложение из двух-трех слов, то очень часто все языки имеют весовой коэффициент 0. Кроме того, исследование показало, что имеются языки, имеющие мало уникальных сочетаний (например, испанский), что также затрудняет его применение.

В связи с этим наиболее надежной видится следующая схема работы алгоритма распознавания языка - на первом этапе текст обрабатывается по 3-му алгоритму. Если отношение коэффициентов 1-го и 2-го по значимости языков больше, чем 2, а абсолютные значения весовых коэффициентов выше некоторого порога (например, 30-50), то на этом и останавливаемся. В противном случае, обрабатываем текст повторно, используя алгоритм 2. В этом случае, для принятия решения о провале на каком-то языке попадают отдельные слова на другом, их вполне можно считать инвариантами для основного языка.

Построение модуля приложения по заданному алгоритму позволит на высоком уровне решить задачу автоматической идентификации языка текстового документа.

Список литературы:

1. Крапивин, Ю.Б. Автоматическое определение языка текстового документа для основных европейских языков / Ю.Б. Крапивин / Информатика. - 2011.
2. Cowie, J. Language recognition for mono- and multilingual documents / J. Cowie, Y. Ludovic, R. Zacharski // Proc. of the Vextal Conference. - Venice, 1999.
3. Natural Language Identification using Corpus-based Models / C. Souter [et al.] // Hermes Journal of Linguistics. - 1994. - № 13. - P. 183-203.\