О создании параллельного двуязычного корпуса веб-публикаций

Ландэ Д.В., Жигало В.В.
Информационный центр ElVisti, Киев, Украина

Представлен алгоритм создания корпуса параллельных текстов. Алгоритм базируется на использовании «опорных слов» в тестовых документах, а также средствах их автоматического перевода. Опорные слова в рамках данного исследования выделялись с использованием русского и украинского морфологических словарей, а также словарей переводов имен существительных для русского и украинского языков. Кроме того, для вычисления весов терминов в документах использовались эмпирико-статистические правила. Рассматриваемый алгоритм был реализован в виде программного комплекса, интегрированного с системой контент-мониторинга InfoStream. В результате был построен параллельный двуязычный корпус веб-публикаций объемом около 30 тысяч документов

1. Введение

Большое место в теории и практике информационного поиска занимают алгоритмы выделения, так называемых, «опорных слов». Многие алгоритмы выявления опорных слов документа основаны на векторном представлении и используют статистические свойства текстов. В основном при выявлении опорных слов (или основ слов) используются частотные словари на одном или нескольких языках.

В данной статье описывается создание частотного словаря на основе морфологического словаря (МС) с использование тестового массива документов, а также построение алгоритма выявления опорных слов с использованием частотного МС и общеизвестного подхода TF IDF [1].

На основе анализа автоматически выявляемых опорных слов и их перевода на другой язык была реализована процедура выявления дубликатов документов, представленных на разных языках.

Как известно, сегодня актуальна задача создания многоязычных параллельных текстовых корпусов [2-4]. Предложенный подход позволил создать дувуязычный украино-русский параллельный корпус текстов из веб-публикаций на русском и украинском языках. Оцененная экспертами точность предложенного алгоритма составляет 98 %.

2. Описание алгоритма

При реализации средств построения корпуса параллельных текстов использовались следующие процедуры:

- построение МС;
- создание частотных словарей на базе существующих МС;
- создание словарей переводов;
- реализация алгоритма выявления опорных слов в документе;
- перевод опорных слов документа на другой язык;
- реализация алгоритма выявления дубликатов на основе анализа опорных слов и их переводов.

2.1. Построение морфологических словарей

Для русского и украинского языков были взяты свободно доступные электронные словари (ispell с набором слов более 1,102 тыс. словоформ на украинском языке и словарь Зализняка, который насчитывает 93 392 слов в нормальной форме).

Морфологические словари были дополнены названиями известных фирм и известными фамилиями, которых не было в исходных словарях.

2.2. Создание частотного словаря

Для выявления опорных слов из документов необходим частотный словарь, в котором для каждого слова записано количество его появлений в некотором большом информационном массиве, а также количество документов, в которых нашлось это слово.

Для создания частотного словаря взят массив документов за 2007 год, сканируемых из Интернет системой контент-мониторинга InfoStream [5]. Массив состоит из текстов веб-публикаций на украинском (1 344 086 документов) и русском языке (2 399 367 документов).

При машинном обучении частотного словаря из каждого документа выделялись словоформы, которые (с определенной вероятностью) приводились к нормальной форме. При этом подсчитать количество, как словоформ, так и нормальных форм в документах, а также подсчитывалось количество документов, в которых встретилась словоформа и/или нормальная форма.

Для эффективности поиска опорных слов в результирующие словари входили только те слова, которые встретились в массиве документов более двух раз. Также было решено использовать только имена существительные.

2.3. Создание словарей переводов

В рамках данных исследований использовались словари переводов с русского на украинский, и с украинского на русский язык. Исходные данные для построения словарей переводов были получены путем перевода имен существительных в нормальной форме существующими программами перевода текстов.

В случае если одному слову соответствовало несколько переводов, то выбиралось наиболее употребляемое значение в соответствии частотным словарем.

2.4. Алгоритм поиска опорных слов

Для поиска опорных слов использовался стандартный подход TF IDF, а точнее его модификация Okapi BM25 [6]:

$$score(D,Q) = \sum_{i=1}^{n} IDF(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{avgdl}\right)},$$

где $f(q_i, D)$ - частота термина q_i в документе D, $|D|$ - длина документа D (число слов), $avgdl$ - средняя длина документа в массиве, k_1 и b - свободные параметры, обычно выбираемые как k_1 = 2.0 и b = 0.75. $IDF(q_i)$ – инверсная частота документа, которая вычисляется по формуле:

$$IDF(q_i) = \log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5},$$

где N - общее количество документов в массиве, $n(q_i)$ - количество документов, содержащих термин q_i.

Для решения проблемы омонимии слов было принято решение брать ту нормальную форму, которая является наиболее частотной в массиве документов.

Затем, для каждого документа все опорные слова ранжировались, и из них выбирались 12 первых, которые и приписывались документу. Кроме того, опорные слова автоматически переводились и также приписывались рассматриваемому документу.

Для улучшения работы алгоритма также использовались стоп-словари для каждого из языков, которые отсеивали нежелательные слова.

Экспертные оценки показали, что удалось добиться 99% качества при переводе опорных слов.

3. Результаты

В результате выполненных исследований в систему контент-мониторинга InfoStream был интегрирован новый механизм поиска дублей, который позволяет с помощью опорных слов находить дубликаты документов в большом информационном массиве. Для реализации этого механизма требуется вхождение всего лишь 5 опорных слов одного документа, длина которого превышает 1000 символов, в состав 12-и опорных слов (или их переводов) другого документа.

На основании приведенного алгоритма были созданы параллельные украино-русские массивы документов. Исходными данными для построения корпуса были веб-публикации за три месяца полученные с помощью системы InfoStream (3 135 279 документов на русском и 425 293 – на украинском языке).

Были использованы также дополнительные критерии отсеивания не полных дубликатов на разных языках :

- общее количество слов в переведенном варианте не должно отличаться больше чем на 10%;
- количество слов начинающихся с большой буквы (не в начале строки) не должно отличаться больше чем на 3 слова, так как в документ может быть вставлено название другого источника информации;
- количество чисел в документах не должно отличатся больше чем на два;
- найденные числа в документах не должны отличаться более чем на 15 %.

В результате эксперимента был получен корпус параллельных текстов из 29 884 документов различной длины, точность перевода которых по экспертным оценкам составляет 98%.

Отобранные параллельные массивы документов размещены в Интернет по адресу: http://www.infostream.ua/ling. Информация представлена в кодировке KOI8-U, в заархивированном виде (gzip). Общий объем заархивированных массивов – 40 Мбайт.

Использование этого корпуса в научных и учебных целях – свободное.

Литература

1. Salton G, Buckley C., Term-Weighting Approaches // Automatic Text Retrieval. Information Processing and Management. 1988. 24, 5. - pp.513-523.
2. Cysouw M., Wälchli B. Parallel texts: Using translational equivalents in linguistic typology
3. B. Pouliquen, R. Steinberger, A. Ribeiro, C. Ignat. Automatic Identification of Document Translations in Large Multilingual Document Collections. Publication: eprint arXiv:cs/0609060v1
4. P. Resnik., Parallel Strands: A Preliminary Investigation into Mining the Web for Bilingual Text. Publication: eprint arXiv:cmp-lg/9808003v1
5. http:///www.infostream.ua
6. http://www.xapian.org/docs/bm25.html