УДК 800 : 002
ББК 81.1
Т 49
В.Ш. Тлюстен

Об одном подходе к автоматизированному созданию словарей статистически значимой лексики естественных языков

Аннотация:

В работе предлагается формальный подход к моделированию процесса отбора статистически значимой лексики естественных языков из представляющих эти языки репрезентативных текстуальных баз. Обсуждаются факторы, влияющие на эффективность такого подхода, и формулируется ряд задач, которые могли бы быть решены на его основе. Устанавливается факт применимости метода к решению некоторых из указанных задач для адыгейского языка.

Ключевые слова:

Еестественные языки, лексика, адыгейский язык, репрезентативные текстуальные базы, компьютерная лингвистика.

Пусть задан кортеж $T=<a_1, a_2, \ldots, a_n>$ *словоформ*, представляющий объединённое множество образцов литературных текстов в некотором естественном языке (ЕЯ). Назовём этот кортеж *текстуальной базой* рассматриваемого языка. Элементы a_i (i=1,n), некоторые из которых, возможно, совпадают, будем называть *вхождениями слов* в T, а их множество $S = \cup\{a_i\}$ – *словарём* текстуальной базы T.

Зафиксируем некоторую *грамматическую категорию* K (например, категорию глагола), задающую на S унарное отношение $K_S(x)$ принадлежности произвольного элемента $x \in S$ выбранной категории K.

Обозначим $G=\{x \mid K_S(x)\}$. Это множество составлено из всех словоформ выбранной текстуальной базы T, относящихся к данной категории K.

Пусть f – функция такая, что $\forall x \in G$: $f(x)=r$, где r – *основа* (неизменяемая часть) словоформы x.

Обозначим $R=\{r \mid \exists x \in G\ (f(x)=r)\}$ и назовём R полным множеством основ или *лексиконом* текстуальной базы T.

Лексикон R определяет на множестве словоформ G отношение эквивалентности "~", такое, что $\forall x,y \in G$: $x \sim y \Leftrightarrow (\exists r \in R, f(x)=f(y)=r)$.

Другими словами, классов эквивалентности в G столько, сколько основ в R и каждый из классов составляют все *равноосновные* лексемы данной грамматической категории.

Например, *сэкIо, тыкIощт, кIогъагъэх* входят в класс эквивалентности адыгейского глагола с основой *кIо* (идти).

Элементы одного и того же класса эквивалентности различаются входящими в них, помимо основы, словоизменительными аффиксами (префиксами и суффиксами).

При этом, в данном контексте, префиксом мы называем предосновную, а суффиксом – постосновную часть слова (нами не выделяются особо, например, окончания).

Таким образом, словоформа в выбранной нами лингвистической модели имеет, в общем случае, структуру: $\alpha\beta\gamma$, где α – префикс, β – основа, γ- суффикс. Причём, β всегда не пусто, а α и γ, возможно, одновременно, могут быть пустыми.

Вводя такую структуризацию словоформ, мы умышленно (для простоты последующей компьютерной обработки) несколько огрубляем ситуацию, не принимая во внимание возможности видоизменений самой основы, в зависимости от окружающего её префикс-суффиксного контекста.

Одни и те же внутрисловные сочетания аффиксов $<\alpha\text{-}\gamma>$, обрамляя различные основы, могут многократно повторяться в словоформах различных классов эквивалентности, являясь характерными не для самих этих классов, а для выбранной *грамматической категории*.

Любое такое сочетание $<\alpha\text{-}\gamma>$ словоизменительных аффиксов, характерное для той или иной формы фиксированной грамматической категории (глагола, существительного, прилагательного и др.) назовём *элементом основного обрамления* или *сигнатурным элементом* этой категории в данном языке.

Элементами основного обрамления английского глагола, например, являются <-ing>, <-ed>, а адыгейского – <сы-щтыгъэ>, < -гъагъэх>.

В общем случае, с каждой грамматической категорией K в данном языке L связано некоторое множество элементов основного обрамления $\sum(K,L)=\{\sigma_1,\sigma_2, \ldots,\sigma_m\}$, которое мы назовём *сигнатурой* данной грамматической категории.

Так, ранее приведённый пример показывает, что при рассмотрении текстуальной базы адыгейского языка, содержащей, среди прочих, словоформы: *сэкIо, тыкIощт, кIогъагъэх,* в сигнатуру глагола в этом языке следовало бы включить элементы: *<сэ->, <ты – щт>, < – гъагъэх>*.

В терминах введённых выше понятий, сформулируем теперь несколько, имеющих как теоретическое, так и прикладное значение задач.

Пусть задана текстуальная база T некоторого ЕЯ L, фиксирована грамматическая категория K и задана связанная с K в языке L сигнатура $\sum(K,L)=\{\sigma_1,\sigma_2, \ldots,\sigma_m\}$.

Пусть, кроме того, задана эффективно вычисляемая булевская функция $\varphi(a_i,\sigma_j)$, для любой пары $< a_i,\sigma_j > (a_i \in T,$

$\sigma_j \in \sum$, i=1,n; j=1,m) истинная тогда и только тогда, когда словоформа a_i включает в себя сигнатуру σ_j .

Требуется предложить эффективные процедуры:

– построения множества G всех содержащихся в T словоформ категории K;

– воссоздания неявно представленного в текстуальной базе T лексикона R, отвечающего категории K в данном языке L;

– $\forall r \in R$, построения системы классов эквивалентности $E(r)=\{x \mid (x \in G) \& (x \sim r)\}$.

Успех решения поставленных выше задач в основном определяется, на наш взгляд, двумя факторами.

1. Внешним фактором – объёмом (влияющим на репрезентативность) используемой текстуальной базы.

2. Внутренним фактором – степенью специфичности выражаемых в соответствующем сигнатурном наборе словоизменительных парадигм для лексических единиц данной грамматической категории.

Оба этих фактора достаточно взаимосвязаны и могут обеспечить успех только при определённом их балансе.

Рассмотрим, например, задачу выявления из некоторой текстуальной базы английского языка множества G всех представленных в ней глагольных словообразований.

Кажется правдоподобным предположение, состоящее в том, что при сканировании очень большой базы англоязычных текстов, *каждый глагол*, в каждой, возможной для него глагольной словоформе, рано или поздно, обязательно встретится. Поэтому, для отбора всех (в данном случае, точнее сказать, *правильных*) глаголов английского языка, казалось бы, достаточно сделать запрос к базе, используя в качестве поискового признака отбираемой лексики сигнатурный элемент <– ed>. В этом случае мы, по-видимому, действительно можем рассчитывать на *полноту* ответа на наш запрос.

Но при этом, будет наблюдаться и удручающая *избыточность* в массиве найденных лексем, т.е. наличие очень большого числа отобранных наряду с глаголами также и нерелевантных запросу (*шумовых)* слов. Так, например, вполне имеет шанс оказаться в числе идентифицированных нами «глаголов» прилагательное *red* и существительное seed.

Ясно, что второй из двух выше указанных факторов, как раз и влияет на степень избыточности ожидаемого отклика на запрос – чем специфичней (и, возможно, длиннее, чтобы уменьшить вероятность случайных совпадений) поисковый сигнатурный элемент, тем меньше “шум”, а значит, выше надёжность отбора словоформ именно требуемой категории.

Вместе с тем, если максимально специфичные и потому предельно информативные сигнатурные элементы относительно редко встречаются в тексте, то возникает опасность существенного “недобора” в общем массиве выявленных из данной текстуальной базы представителей различных классов эквивалентности (т.е. в результирующем списке лексем найдут отражение далеко не все представленные в базе глагольные основы).

И этот негативный эффект будет тем ниже, чем более обширная текстуальная база будет использована.

Отсюда можно сделать общий вывод: для наилучшего (наиболее полного и наименее избыточного) отбора лексем требуемой грамматической категории, эффективней всего использовать максимально специфичные сигнатурные элементы и как можно более обширную текстуальную базу.

Перейдём теперь к вопросу о возможности автоматизированного создания, на базе рассматриваемого сигнатурного подхода, словарей, охватывающих язык в целом.

Хотя все предыдущие рассуждения строятся относительно фиксированной, а следовательно, ограниченной текстуальной базы T, при достаточном объёме последней, они могут стать вполне надёжным фундаментом получения объективных суждений о языке L в целом.

Действительно, при постепенном пополнении текстуальной базы T, по мере её расширения, в некоторый момент, как нам кажется, должно произойти статистически значимое насыщение представленной в T лексики так, что с дальнейшим увеличением n – числа вхождений слов в T, словарь S практически перестанет расширяться.

Минимальное значение $n=n_0$ при котором это произойдёт, назовём *лексически репрезентативным* текстуальным объёмом, а соответствующую текстуальную базу T_0 – *лексически репрезентативной текстуальной базой выбранного языка*.

Конечно, величина n_0, если она существует, определяет не точное значение, а лишь порядок статистически значимого для данного языка и данной грамматической категории объёма текстуальной базы.

Кроме того, n_0 определяется не только объёмом, но и структурой T_0. По разному формируя текстуальную базу (даже, в различной последовательности сочетая составляющие её тексты), можно получать различные значения n_0. Отсюда возникает вопрос о нахождении наименьшего в данном языке (или, возможно, у данного автора) нетривиального значения n_0 и соответствующей текстуальной базы.

Можно поставить вопрос о степени эффективности использования метода сигнатур при решении выше перечисленных задач для различных *языков*. Скорее всего, окажется, что разброс здесь достаточно большой.

Например, для глагола адыгейского языка, обладающего, как известно, чрезвычайно развитой системой словоизменительных парадигм, этот метод вполне мог бы оказаться успешным.

Для проверки этой гипотезы, в лаборатории компьютерных технологий АГУ, создана экспериментальная текстуальная база адыгейского языка и разработан комплекс реализующих выше указанный подход компьютерных программ.

Результаты соответствующих экспериментальных исследований предполагается опубликовать в отдельной работе.

Примечания:

1. Рогава Г.В., Керашева З.И. Грамматика адыгейского языка. – Краснодар-Майкоп: Краснодарское книжное издательство, 1966.