АВТОМАТИЧЕСКИЙ АНАЛИЗ ТЕКСТОВ НА КИТАЙСКОМ ЯЗЫКЕ. ПРОБЛЕМА ВЫБОРА БАЗОВОЙ ЕДИНИЦЫ.

Загибалов Тарас Евгеньевич

Красноярский госуниверситет,Красноярск

8055@inbox.ru

Автоматический анализ текстов на китайском языке затруднён нерешённостью проблемы делимитации слова. Слово в данном языке не является единицей, которую можно было бы всегда чётко выделить по каким-либо формальным признакам. С другой стороны, нельзя отрицать наличие слова в китайском языке. Автор предлагает подход к анализу текстов, не требующий предварительной сегментации текста на слова. Данный подход реализован автором в виде программы автоматического реферирования и классификации текстов.

В контексте автоматической обработки текстовой информации наиболее очевидной является проблема делимитации слова в китайском языке. Эту проблему можно условно разбить на две составляющих: техническую и собственно лингвистическую.

Технически сложность определения границ слова заключается в том, что в китайской системе письменности не принято отделять слова пробелами, как это принято в, например, европейских языках. Конечно, и в упомянутых европейских языках проблема определения слова не ограничивается нахождением его физических границ, но, как правило, все системы автоматического анализа письменного текста в качестве базовой единицы принимают именно графическое слово – последовательность знаков алфавита, отделённых друг от друга пробелами или знаками препинания. Проблема делимитации слова в китайском языке осложняется также и тем, что слова в китайском языке почти лишены формальных признаков морфологического уровня. В словообразовании доминирует корнесложение, флективные формы встречаются нерегулярно, словоизменение почти отсутствует, что в совокупности делает невозможным выделение хотя бы основных лексических единиц путём определения их границ по каким-либо морфологическим маркерам.

Тем не менее, многие исследователи пытаются решить проблему так называемой сегментации китайского текста (разбиение текста на слова). Для этого используются три основных способа: словарный (обычно с применением алгоритма максимального соответствия), статистический и комбинированный, сочетающий в себе оба предыдущих. По сообщениям авторов им удаётся правильно сегментировать текст на 95-99 процентов, что, казалось бы, позволяет говорить о том, что проблема почти решена. Но, на наш взгляд, предлагаемые способы не решают проблему делимитации слова в тексте, а скорее имитируют её решение. Так, автор одной из систем сегментации текстов на китайском языке, говоря о критериях оценки эффективности работы подобных систем, писал, что “точность определения эффективности работы программы сегментирования текстов может быть внятно определена только при условии наличия общепринятого определения того, что считать словом в китайском языке. На практике же, исследователи в области автоматической сегментации текстов на китайском языке, отмечая неоднозначность определения слова, как правило, принимают свои рабочие определения слова в китайском языке или просто полагаются на субъективные суждения носителей языка» [8]. Таким образом, очевидно, что результаты подобной сегментации весьма условны и базируются на субъективном понимании того, что есть слово в китайском языке.

На наш взгляд, описанная выше «техническая» проблема не может быть исчерпывающе решена. Причина этому кроется в заявленной выше лингвистической составляющей проблемы делимитации слова. Дело в том, что невозможность полного разбиения текста на слова, обусловлена тем, что само слово (в европейском понимании) в китайском языке не является регулярной единицей. Это нашло выражение в определении лингвистами китайских слов как «нечётких» (fuzzy) или «подвижных» (flexible).

Конечно, в китайском языке большое количество слов имеет постоянные формы и может быть легко выделено в тексте. Но вместе с такими, «классическими» словами существуют классы слов, которые не могут быть однозначно определены не только по формальным признакам, но и по семантическим.

Так, например, группа слов, построенных по глагольно-объектной словообразовательной модели (надо отметить, весьма распространенной в современном китайском языке), как правило, обладают свойствами словосочетания, например:

- *1. 睡觉 shuijiao («спать»), букв.: «спать + сон», (в словарях зафиксировано как слово);*
- *2. 睡了三个小时的觉 shuile sange xiaoshide jiao («спал три часа»), букв.: «спать+пр.вр. три часа+атриб. сон»);*

- *3. 睡懒觉 shui lanjiao («валяться в посели»), букв.: «спать ленивый сон»;*
  и т.д.

Таким образом, мы можем заметить, что приведённая выше единица китайского языка (1), зафиксированная в словарях в качестве слова, воспринимаемая носителями китайского языка как одно слово (фонетически неделимо, вторая часть самостоятельно не используется, в китайской лингвистической традиции такие единицы называются liheci – «слитно-раздельные слова»), тем не менее, распадается на две части, при этом вторая часть может принимать различного рода атрибутивные конструкции. В то же время приведённую единицу нельзя считать и словосочетанием, так как при разделении частей этой единицы происходят семантические изменения, описанные в [5].

Вообще для китайского языка свойственно, что «(…) когда в сложной единице, составленной из самостоятельных по отдельности компонентов, обнаруживается связь аналогичная синтаксической, а компоненты не подвергаются никаким фонетическим видоизменениям, создаётся ситуация неразличения сложного слова и словосочетания» ([1], с 168).

Кроме того, «(…) выделенные из текста в соответствии критериями единицы (…) далеко не всегда (…) являются устойчивыми и воспроизводимыми» [2].

Более того, словоделение, как показали в своих работах Tsai [7] и Hoosain [6], не является необходимым условием понимания текста носителями языка в процессе чтения. Так, сегментированный текст вызвал у читавших его китайцев некоторые затруднения, что выразилось в снижении темпа чтения. Также в [7] описан эксперимент, показавший наличие значительных несовпадений результатов сегментации текста, выполненного китайцами. Это подтверждает тезис о том, что полная разбивка текста на слова является противоречащей реалиям китайского языка и, следственно, субъективной.

В этой связи возникали идеи перейти на уровень иероглифов – zi (цзы). Сразу необходимо пояснить, что цзы (также логограмма) не является прямым аналогом морфемы. Китайская логограмма является минимальной самостоятельной графической единицей китайского языка, формирующей облик языковой системы и в плане формально-графическом, и в плане семиотическом [3]. Более того, и в сознании носителей китайского языка и в китайской лингвистической традиции логограмма всегда была основной единицей, понятие слова (точнее – цы) всегда было вторичным для китайцев (т.н. «слабый ярус» по [4]). Но при этом нельзя учитывать, как минимум, два фактора, не дающих полностью перейти к анализу текстов на основе выделения только данных самим языком графических единиц. Прежде всего, цзы, несёт в себе слишком мало информации, что неизбежно ведёт к неопределённости значения. А кроме этого, нельзя отрицать наличия в китайском языке слов (в составе которых в значительной степени снимается неоднозначность цзы).

В связи с вышеизложенным, представляется разумным при разработке систем автоматического анализа текстов на китайском языке попытаться в максимальной мере учитывать особенность китайского языка, а именно «слабость» слова, его нерегулярность. Это подтолкнуло нас к мысли не применять этап предварительной сегментации текста на слова, так как последние (как показано выше) не могут быть единственной и вполне надёжной единицей деления текста.

Мы попытались воплотить наши выводы в разработанной нами компьютерной программе CLAAS (Chinese Language Automatic Annotation System).

Программа CLAAS предназначена для анализа текстов на китайском языке различной тематики в текстовом формате (plain text).

В качестве результата анализа система создаёт реферат текста (в нашем случае это сжатое изложение анализируемого текста при помощи цитирования самого текста) и список «ключевых слов» (то есть слов, наиболее тесно связанных с тематикой данного текста). Причём уровень «сжатия текста» может быть задан произвольно и обычно находится в пределах от 5% до 50% от объёма исходного текста.

В разработанной нами системе автоматического реферирования текстов на китайском языке (CLAAS) за элементарную единицу предварительного анализа текста мы взяли иероглиф (логограмму).

Именно на основе подсчёта так называемого относительного семантического веса иероглифа определяется семантический вес всех остальных единиц текста (слов, синтагм, предложений, абзацев). На основании подсчёта семантического веса предложений делается вывод о включении того или иного предложения в реферат текста.

Последующий анализ предусматривает поиск повторений окружения (непосредственного контекста) данного иероглифа в данном тексте. Найденные повторяющиеся цепочки иероглифов проверяются на вхождение в них других повторяющихся цепочек, что позволяет находить среди повторяющихся фрагментов словосочетания и обрабатывать их отдельно. Также возможна фильтрация словосочетаний при помощи словаря.

Кроме того, возможен поиск повторяющихся комбинаций иероглифов с переменной дистанцией между ними в пределах одной синтагмы, что позволяет находить liheci.

Исключение этапа сегментации текста на слова даёт значительный прирост производительности рассматриваемого способа. Это объясняется тем, что для статистического способа сегментации требуется сложная система расчетов для вычисления вероятности того, что данная цепочка знаков является тем или иным

словом. Способы, построенные на использовании словаря, также не могут быть достаточно быстрыми, так как вынуждены производить выборку из словарной базы данных в десятки тысяч слов по довольно сложному алгоритму. А так как слова в китайском языке в основном двух- или трёхсложные (причём один слог записывается одним иероглифом, каждый из которых записывается двухбайтным кодом), то используемая в предлагаемом способе база данных по частотам иероглифов в языке (около 6.000 двухбайтных знаков) оказывается во много раз меньше словарной базы данных (например: словарь в 50.000 слов при средней длине слова 2,5 иероглифа будет иметь 50000*2,5=125000 знаков, вместо 6000 в предлагаемом способе).

Программа CLAAS апробировалась на текстах различной тематики и показала себя надёжным средством сжатия текстовой информации на китайском языке, при условии наличия чётко выраженной тематической направленности текста, его тематической и смысловой однородности. Так, например, попытка анализа дискурса одного из участников Интернет-чата не была вполне успешной, так как его общение не было сосредоточено вокруг одной темы. Тем не менее, удалось найти некоторые интересные особенности дискурса данного лица.

В качестве примера можно привести результаты анализа публицистического текста.

Текст посвящён описанию террористического акта в московском метро 6 февраля 2004 года.

Реферат (27% от общего объёма) оказался следующим (с переводом каждого предложения):

综述：俄大选前莫斯科地铁遭恐怖袭击酿伤亡惨剧

Тема: В России перед президентскими выборами террористический акт в метро привёл к трагедии.

爆炸发生时，列车距站台500**米**。

Во время взрыва поезд находился в 500 метрах от платформы.

俄通社塔斯社报道称，发生爆炸的第二节车箱中共有100名乘客。

Как сообщило российское информационное агентство ТАСС, во время взрыва во втором вагоне находилось около 100 пассажиров.

目击者：这是"**一**场血淋淋的屠杀"

Очевидец (происшествия): «Это было кровавое месиво».

这名目击者称，爆炸发生时他在另一节车箱。

Этот свидетель сказал, что во время взрыва он находился в соседнем вагоне.

他说："**我看到手臂和断腿散落在**车箱中,这是一场血淋淋的屠杀"

Он сказал: «Я видел оторванные руки и ноги, разбросанные по всему вагону, это кровавое убийство»

来自内务部官员的消息称，此次爆炸可能**是一次恐怖**袭击。

По сообщению работника МВД, этот взрыв, вероятно, является терактом.

爆炸系恐怖袭击并非列车故障引发

Взрыв является терактом, а не аварией.

内务部官员认定，此次爆炸可能是一次恐怖袭击。

Чиновник МВД признал, что этот взрыв, вероятно, является терактом.

现场的调查官员称，爆炸的威力相当于2**公斤**TNT，**可能是一名女**"**人**弹"**将炸**药带上地铁列车然后引爆，也可能是恐怖分子将定时爆炸装置放在包里并扔到车箱的某个角落。

Находившийся на месте происшествия следователь заявил, что взрыв по мощности эквивалентен двум килограммам тринитротолуола, и, вероятно, был приведён в действие террористкой-смертнцей, пронёсшей в метро взрывчатку на поясе, также возможно, что взрывное устройство было предварительно заложен террористами в вагоне в сумке, а потом было взорвано часовым механизмом.

因此，只能认为这是一次恐怖袭击。

Поэтому, можно сделать вывод о том, что это был теракт.

莫斯科和俄罗斯各地发生的**自**杀式恐怖袭击多为车臣恐怖分子所为。

Большинство терактов, производимых террористами-смертниками в Москве и по всей России, связывают с чеченскими террористами.

袭击意在破坏俄大选普京谴责

Целью теракта является попытка помешать переизбранию Путина на второй срок.

爆炸发生时，普京总统正在准备接见阿塞拜疆总统阿里耶夫，接到通报后，他立即发表声明对此次地铁爆炸事件进行谴责，并将事件定性为恐怖袭击。

В тот момент, когда произошёл взрыв, Путин готовился к встрече Президента Азербайджана Алиева, получив известие (о теракте), он сразу же выступил с заявлением, в котором осудил теракт, а также выразил уверенность в террористическом характере взрыва.

При большем сжатии текста (около 7%) был получен следующий результат:

综述俄大选前莫斯科地铁遭恐怖袭击酿伤亡惨剧

Тема: В России перед президентскими выборами террористический акт в метро привёл к трагедии.
爆炸*发生时，列车距站台500****米****。*
Во время взрыва поезд находился в 500 метрах от платформы.
爆炸系恐怖*袭击并非列车故障引发*
Взрыв является терактом, а не аварией.
内*务部官员认定，此次爆炸可能是一次恐怖袭击。*
Чиновник МВД признал, что этот взрыв, вероятно, является терактом.
莫斯科和俄*罗斯各地发生的自杀式恐怖袭击多为车臣恐怖分子所为。*
Большинство терактов, производимых террористами-смертниками в Москве и по всей России, связывают с чеченскими террористами.

Как видно из примера, не смотря на некоторые потери в содержании (исчезли описания очевидцев и версия о связи теракта с предстоящими выборами, а также ряд деталей), в целом, по полученному результату можно сделать вполне адекватное заключение о содержании статьи.

В качестве ключевых слов программа выбрала следующие:

Слово	Перевод	Частота
恐怖	террор	18
地▪	метро	17
列▪	поезд	10
普京	Путин	10
▪▪	атака	8
莫斯科	Москва	8
乘客	пассажир(ы)	7
内▪部	МВД	5
大▪	выборы	4
疏散	эвакуация	3
▪▪	президент	3

Как видно из примера, программа в целом показала удовлетворительные результаты, как в реферировании, так и в поиске ключевых слов. Также хотелось бы отметить, что программа сумела безошибочно определить имена собственные, что является весьма сложной задачей при автоматическом анализе текстов на китайском языке.

Конечно, так как программа является лишь прототипом полноценной системы реферирования и не использует ряд методов повышения эффективности работы, используемые в коммерческих продуктах, результат не является идеальным. Но, смеем надеяться, что полученный результат в целом демонстрирует возможность анализа текстов на китайском языке без предварительного разбиения текста на слова.

Список литературы:

1) Солнцев В.М. Введение в теорию изолирующих языков. М.: «Восточная литература», 1995
2) Антонян К.В. Единицы словаря и единицы текста в современном китайском языке //Материалы XII международной конференции «Китайское языкознание. Изолирующие языки».- Институт языкознания РАН, 2004
3) Готлиб О.М. Китайская логограмма в грамматолого-семиотическом аспете // сб. материалов VI международная конференция по языкам Дальнего Востока, Юго-Восточной Азии и Западной Африки: (2001, СПбГУ))
4) Ефремов А.М. Связность китайского текста в сравнительно-типологическом аспекте. Дисс. …канд. Филол.. наук. – М., 1987.
5) Хорошкина А.С. Китайские «пустые дополнения»: к проблеме переходности //Материалы XII международной конференции «Китайское языкознание. Изолирующие языки».- Институт языкознания РАН, 2004
6) Hoosain, R. (1992). Psychological reality of the word in Chinese. In H. C. Chen & O. J. L. Tzeng (Eds.), Language processing in Chinese (pp. 111-130). Elsevier.
7) Tsai, C. H., McConkie, G. W., & Zheng, X. J. (1998, November). Lexical parsing by Chinese readers. Poster session presented at the Advanced Study Institute on Advances in Theoretical Issues and Cognitive Neuroscience Research of the Chinese Language, University of Hong Kong.

8) Nianwen Xue, Chinese Word Segmentation as Character Tagging, Computational Linguistics and Chinese Language Processing; Vol. 8, No. 1, February 2003, pp.29-48 © The Association for Computational Linguistics and Chinese Language Processing