ЛИНГВИСТИЧЕСКИЙ ПРОЦЕССОР КИТАЙСКОГО ЯЗЫКА. ОСОБЕННОСТИ РАЗРАБОТКИ

П. Ю. Довнар, А. В. Воронцов

ИП Инвеншион Машин
Минск, Беларусь
E-mail: pdovnar72@gmail.com

Статья посвящена особенностям разработки лингвистического процессора китайского языка.

Ключевые слова: автоматическая обработка естественного языка, лингвистический процессор.

Современный лингвистический процессор (ЛП) является базовым компонентом практически любой прикладной системы автоматической обработки текстовых документов. Разработка промышленных лингвистических процессоров для обработки естественных языков (ЕЯ) является на сегодняшний день актуальной, перспективной и одновременно трудной задачей. Последнее особенно характерно для языков, использующих иероглифическую письменность. В данном случае речь идет о разработке лингвистического процессора китайского языка, планируемого к внедрению в состав промышленного многоязычного лингвистического процессора системы автоматизации инженерии и управления знаниями Goldfire Innovator компании «Инвеншион Машин». Его функциональность прежде всего направлена на реализацию автоматического лингвистического анализа ЕЯ на всех уровнях их глубины.

По распространенности китайский язык занимает первое место в мире. Только в Китае на нем говорят более 1,3 млрд человек. Ввиду того, что на китайском языке создается огромное количество технической литературы (патенты, инструкции и т. д.), растет потребность в налаживании межъязыковой коммуникации, что делает актуальным создание программ по автоматической обработке китайского языка.

Прежде чем перейти к описанию проблем, связанных с автоматической обработкой китайского языка, перечислим основные особенности данного языка.

Китайский язык пользуется иероглифической письменностью. Каждый иероглиф обозначает отдельный слог и отдельную морфему. Морфемы и простые слова, как правило, односложны. Часть односложных слов из древнекитайского языка употребляется лишь как компонент сложных и производных слов. Доминирует двусложная (двуморфемная) норма слова. В связи с развитием терминологии растет число более чем двусложных слов. Словообразование осуществляется за счет словосложения, аффиксации и конверсии. Формообразование представлено главным образом глагольными видовыми суффиксами. Форма множественного числа присуща существительным, обозначающим лиц, и личным местоимениям. Аффиксы немногочисленны, в ряде случаев факультативны, имеют агглютинативный характер. Агглютинация не служит выражению отношений между словами, и строй китайского языка остается изолирующим.

Синтаксис характеризуется номинативным строем, грамматически значимым порядком слов, определение всегда в препозиции. Предложение с переходным глаголом в качестве сказуемого может иметь форму активной и пассивной конструкции; возможны перестановки слов, не меняющие их синтаксической роли. Китайский язык имеет развитую систему сложных предложений, образуемых союзным и бессоюзным сочинением и подчинением [2].

Автоматическая обработка китайского языка по структуре не отличается от обработки других языков и представляет собой линейный процесс: входной текст проходит последовательные этапы формализации и структуризации, каждый из которых соотносится с определенным уровнем восприятия и реализуется в рамках отдельного модуля обработки текста. Можно выделить следующие основные уровни анализа текста: графематический, лексико-грамматический, синтаксический и семантический [3]. Большинство проблем, с которыми нам пришлось столкнуться, связаны с первыми тремя из перечисленных уровней, так как по результатам синтаксического анализа языковая информация представляется в уже достаточно унифицированном виде и для ее дальнейшей семантической обработки применяются универсальные для других языков подходы и методы. Что же касается начальных уровней анализа, то здесь приходится искать решения, адаптированные под специфику и особенности именно китайского языка.

На уровне графематического анализа текста на китайском языке основная проблема связана с отсутствием в тексте пробелов или других явных признаков, указывающих на границы между словами, в связи с чем остро встает проблема установки критериев определения границ слов. Связь между частями сложного слова в китайском языке осуществляется путем свободного примыкания, не получая формального выражения посредством специальных морфологических показателей. При этом большое количество слов, образованных таким способом, могут быть интерпретированы как цельные понятия и как синтаксические конструкции.

Например, слово «*вулкан*», состоящее из двух иероглифов «*огонь*» и «*гора*», может рассматриваться как развернутая именная группа (главное слово и его атрибут – «*огненная гора*»). Прилагательное «*интересный*» также состоит из двух иероглифов «*иметь*» и «*смысл*» и может быть интерпретировано как глагольная группа (сказуемое и его дополнение –«*имеющий смысл*»).

Исследования показали, что при решении данной проблемы с точки зрения промышленного характера ЛП целесообразно руководствоваться «прагматическим» принципом: разбивать слова так, чтобы они оптимально подходили для задач последующего анализа и обработки (т. е. для лексико-грамматического и синтаксического анализа), а также для обработки с целью смыслового поиска информации, извлечения знаний, машинного перевода. Другими словами, в этом процессе необходимо найти своеобразную «золотую середину». С одной стороны, деление слов на слишком мелкие части приводит к увеличению многозначности и соответственно трудоемкости обработки. С другой стороны, игнорирование смысловых частей слова, будь то полнозначные слова или отдельные морфемы, на практике приводит к снижению показателя полноты извлекаемой информации, что негативно сказывается на результатах обработки.

В процессе исследования был выведен ряд общих принципов, которыми следует руководствоваться при принятии решений относительно отдельных проблем разбиения конкретных слов на смысловые части. В таблице приводятся некоторые из этих проблем и представлены их возможные решения.

Описание проблемы	Примеры	Решение
Многие глаголы представляют собой сочетание акции и объекта, при этом каждое из составляющих слов свободно используется самостоятельно или в сочетании с другими словами	说话 – говорить: 说 – говорить, 话 – речь. 吃饭 – кушать: 吃– есть, 饭– рис. 唱歌 – петь: 唱– петь, 歌– песню	Если глагол обозначает единое действие и если у объекта нет зависимых слов, то такие слова объединяются в одно
Слова часто образуются способом усечения основ	环境保护 – 环保 защита окружающей среды 世界博览会 – 世博 всемирная выставка	Такие слова распознаются как одно слово и по возможности сохраняются их конвертации к полным вариантам
Существует несколько так называемых «морфем», состоящих из одного иероглифа. Эти «морфемы» используются вместо полнозначных слов и образуют новые слова (эти морфемы можно считать своеобразными суффиксами)	量 обозначает количество чего-либо:零售量 – объем розничной торговли, 成交量 – количество сделок. 率 имеет значение «коэффициент»: 生产率 – производительность (коэффициент производительности), 出生率– рождаемость. 员 придает значение профессии: 教员– преподаватель, 售货员 – продавец. 机 обозначает прибор: 发射机 – излучатель, 录音机 – проигрыватель, 收音机 – радиоприемник	Составлен список таких морфем с указанием их лексических значений; образованные с их помощью слова распознаются как одно; при необходимости объединенные слова могут быть разбиты на этапе семантического анализа и из них может быть извлечена необходимая информация
Термин состоит из нескольких полнозначных слов	淀粉糊精 – амилодекстрин: 淀粉 – крахмал, 糊精 – декстрин. 白血病 – лейкемия:白 – белый, 血 – кровь, 病 – болезнь	При работе с терминами выделяются все значимые части, которые не противоречат общим принципам разбиения слов

Основная проблема на уровне лексико-грамматического анализа текста связана с классификацией слов. Для европейских языков за основу такой классификации принято использовать разделение слов по частям речи [1]. Что касается китайского языка, то ряд исследователей вообще отрицают в нем наличие частей речи, аргументируя это отсутствием словоизменения в европейском понимании. Однако при выделении синтаксических и семантических отношений между словами невозможно обойтись без какой-либо предварительной лексико-грамматической классификации. Поэтому было принято решение о выделении таких традиционных частей речи, как существительное, местоимение, прилагательное, глагол, наречие, числительное. Также были введены лексико-грамматические классы для служебных слов: предлоги, послелоги, союзы, частицы, счетные слова и др. В результате, разработанный лексико-грамматический классификатор включает в себя всего около 50 различных классов слов, а также классы для обозначения знаков пунктуации.

Особенностью разработанного классификатора является то, что он включает в себя только те классы слов, точное автоматическое распознавание которых достижимо с опорой на локальный контекст слова. В случаях, когда для распознавания того или иного класса необходим анализ более широкого контекста, соответствующее решение должно приниматься ЛП на этапе синтаксического анализа. Так, например, в разработанном лексико-грамматическом классификаторе используется один и тот же класс для обозначения предлогов, которые могут вводить после себя как именные, так и глагольные группы.

Опираясь на описанные выше принципы разбиения и классификации слов, был разработан аннотированный лексико-грамматическими классами эталонный корпус текстов. Данный корпус содержит тексты различных жанров и предметных областей в сбалансированной пропорции. Его объем составляет около 50 тыс. предложений (более 1.3 млн словоупотреблений). Эталонный корпус используется в качестве тренировочной и тестировочной базы для реализации графемного и лексико-грамматического анализа текста.

Анализ указанного выше корпуса подтвердил, что в китайском языке показатель омонимии (синтаксической транспозиции) слов достаточно высок, т. е. большое количество слов могут использоваться в функции нескольких частей речи. Достаточно проблемными в этом плане и с точки зрения автоматической обработки текста являются следующие пары классов слов: существительное – глагол, прилагательное – существительное, прилагательное – наречие и предлог – глагол. Важно заметить, что в некоторых случаях частеречная принадлежность слова в определенном контексте в силу его синтаксической позиции, и функции, является однозначной:

- 人类可以**研究**这些动物.

Человечество может **исследовать** этот вид животных.

- **研究**发现两种新的蛋白.

Исследование обнаружило два новых вида белка.

- 该**研究**团队现有教师18人.

В состав данной **исследовательской** группы входят 18 преподавателей.

С другой стороны, встречается много случаев, когда допустимы различные трактования и синтаксической роли слова, и соответственно его частеречной принадлежности:

- **用**光学望远镜观察同区域。

Используя бинокль, исследовать регион.
С помощью бинокля исследовать регион.
• **经**分选清理取得海绵铁 。
Проведя очистку, получить железо.
Получить железо **посредством** очистки.
• **靠**国外公司开发石油。
Развивать нефтепромышленность **за счет** государственных предприятий.
Развивать нефтепромышленность, **опираясь** на государственные компании.

Для того чтобы частично минимизировать указанную многозначность и соответственно повысить точность анализа текста, было принято решение максимально избавиться от омонимии предлог – глагол путем однозначной классификации слов в тот или другой класс.

Исходя из практических соображений, графемный и лексико-грамматический этапы анализа были объединены в рамках единой стадии обработки текста. Такой подход позволяет быстро и эффективно находить среди большого количества всех возможных вариантов разбиения предложения на слова именно те, которые максимально вероятны с точки зрения лексико-грамматического анализа текста. В качестве статистической модели была использована модель Маркова. В дополнение к этому был разработан механизм распознавания незнакомых слов, а также механизм корректировочных правил. При таком подходе точность анализа на сегодняшний день составляет около 93 %.

Синтаксический анализ в целом аналогичен его реализациям для других языков. Особое внимание здесь уделяется порядку слов и наличию в предложениях так называемых рамочных конструкций (т. е. когда, например, косвенный объект оформляется одновременно с помощью предлога и послелога). Из-за отсутствия формальных признаков серьезную проблему представляет правильное распознавание границ именных и предикативных определений, а также предлогов и подчинительных союзов.

Приведем примеры таких трудностей:

- **开**发处**理器的技**术
  Технология, которая разрабатывает процессор
  Разрабатывать технологию процессора
- 应**用技能的学**习
  Применять обучение навыкам
  Обучение, применяющее навыки
- **利用**技术**的**发**展**
  Развитие с использованием техники
  Использовать развитие техники

В качестве тренировочного и тестировочного материала используется разработанный нами эталонный корпус синтаксических деревьев. Его объем корпуса составляет около 4 тыс. предложений.

Проведенное исследование подтвердило, что технология автоматической обработки китайского языка не имеет принципиальных отличий по сравнению с технологией обработки других ЕЯ, а полученный лингвистический процессор китайского языка может быть эффективно внедрен в состав системы автоматизации инженерии и управления знаниями Goldfire Innovator.

ЛИТЕРАТУРА

1. *Белоногов, Г. Г.* Компьютерная лингвистика и перспективные информационные технологии / Г. Г. Белоногов, Ю. П. Калинин, А. А. Хорошилов. Рус. мир, 2004. 203 с.

2. Большой энциклопедический словарь. Языкознание / под ред. В. Н. Ярцевой. М.: БРЭ, 1988. 687 с.

3. *Воронцов, А. В.* Лингвостатистический метод автоматического лексико-грамматического анализа англоязычных текстов: дис. … канд. филол. наук: 10.02.21 / А. В. Воронцов. Минск: МГЛУ, 2008. 173 с.