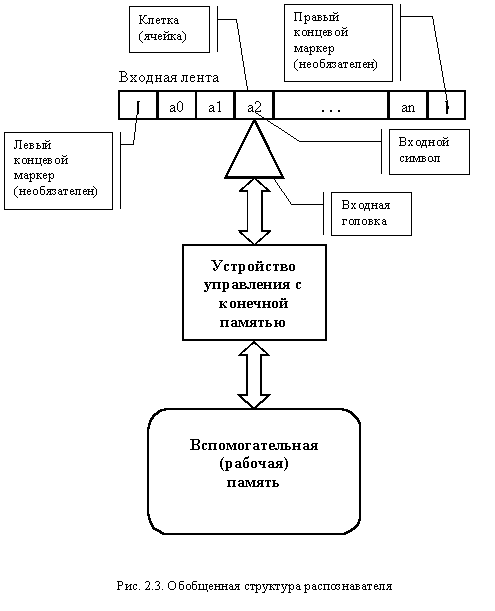
КЗ – грамматики



В общем случае любой распознаватель выглядит так.

<https://studopedia.ru/8_59983_mashina-tyuringa-kak-universalniy-raspoznavatel.html>

Это принцип работы машины тьюринга. +/- так же работает ЛОА (только там есть маркеры начала и конца и невозможно уйти за них)

Тут можно сделать вывод, что

Способа разобрать текст по контекстно-зависимой грамматике в один проход нет. Точно так же не справится и разбор с возвратами и стеком. Существующие алгоритмы имеют чрезмерно большую вычислительную сложность и не подходят для быстрой компиляции исходного кода. Поэтому все формальные языки используют простой трюк:

* грамматика языка составляется как контекстно-свободная, и по этой грамматике пишется парсер
* вводятся дополнительные правила (семантика языка), которая накладывает ограничение на уже разобранные парсером деревья разбора (parse tree), или на абстрактные синтаксические деревья (abstract syntax tree)

Применение – переводчики

Система, проводящая полный анализ текста на естественном языке должна состоять из нескольких модулей, осуществляющих поэтапный анализ исходного текста:

− графематический анализ;

− морфологический анализ;

− предсинтаксический анализ;

− синтаксический анализ;

− постсинтаксический анализ;

− семантический анализ.

Графематический анализ – первичный этап в процессе автоматической обработки текстов на естественном языке. Основной задачей графематического анализа является выделение структурных единиц из входного текста, а именно предложений абзацев, слов, знаков препинания и т. д. Входной текст может иметь как линейную структуру, содержащую единый фрагмент текста, так и 7 нелинейную. В этом случае текст содержит различные структурные единицы: основной текст, заголовки, примечания, комментарии и т. д. Выходными данными этапа графематического анализа является графематическая таблица, которая в свою очередь служит входными данными для следующего этапа автоматической обработки текста – морфологического анализа.

Задача морфологического анализа – определение нормальной формы, от которой была образована данная словоформа, а так же получение набора её морфологических характеристик. Это делается для того, чтобы в последующих этапах анализа использовать только нормальную форму слова, а не все его словоформы и использовать морфологические характеристики для проверки согласованности слов. Нормальной формой слова называется форма, принятая для обозначения понятия, связанного с данным словом. Нормальной формой считается слово в именительном падеже и единственном числе. Морфологические характеристики – это набор пар «ключ, значение». В роле ключа выступает, например, род, число, падеж, склонение, время и другие признаки слов, используемые в русском языке. Значением является какое-либо конкретное значение, которое может принимать данный признак (ключ). Например, ключ «число» может принимать значения «единственное», «множественное» или «двойственное», ключ «род» может принимать значения «мужской», «женский», «средний» и т. д.

Предсинтаксический анализ выполняет следующие функции: объединение отдельных лексических единиц в одну синтаксическую и, наоборот, разделение одной лексической единицы на несколько синтаксических. В одну синтаксическую единицу объединяются неразрывные словосочетания. Например, «золотые руки», «тянуть за язык» и т. д. Еще до проведения синтаксического анализа есть возможность провести поверхностный анализ структуры разбираемого предложения.

Синтаксический анализ в свою очередь сможет, опираясь на результаты такого анализа, отбросить часть вариантов синтаксического разбора предложения. Анализ 8 структуры разбираемого предложения называется синтаксической сегментацией. Первая задача синтаксической сегментации – используя определенные правила языка, уменьшить количество омонимов, соответствующих каждой словоформе. Второй задачей является выделение различных синтаксических конструкций. Например, выделение простых предложений в составе сложных для проведения в дальнейшем над ними независимого синтаксического анализа, выделение синтаксических групп (именных, глагольных, и т. д.) [2]. Задачей синтаксического анализа является определение роли слов и их связи между собой. Результатом данного этапа является набор синтаксических деревьев, отражающих такие связи. Синтаксический анализ является самым сложным этапом анализа текста на естественном языке. В связи с неоднозначностью как входных данных (одна словоформа может быть получена от нескольких различных нормальных форм), так и самих правил разбора, как правило, существует большое количество различных вариантов разбора [2].

Постсинтаксический анализ решает две следующие задачи. Первая заключается в необходимости уточнить смысл слов, который выражается с помощью различных средств языка: предлогов, префиксов и т. д. Проблематика второй задачи заключается в том, что одну и ту же мысль можно выразить разными конструкциями языка. Например, в многоязычно диалоговой системе одна и та же мысль может быть выражена различными синтаксическими конструкциями. В связи с этим появляется необходимость в нормализации дерева, чтобы свести конструкции, которые выражают одно действие различным образом для разных ситуаций, к одному нормализованному дереву. Так же на этапе постсинтаксического анализа может проводиться поиск изменяемых словосочетаний, составные части (слова) которых могут быть разделены другими словами [2]. 9

Семантический анализ – завершающий этап автоматической обработки текста. На данном этапе производится смысловой анализ текста – уточняются связи, которые не смог уточнить постсинтаксический анализ, т. к. большое количество ролей выражается не только при помощи средств языка, но и с помощью значения слова. Вторая задача семантического анализа – исключить некоторые значения слов, а иногда даже целые варианты разбора как семантически несвязные [2].

<https://www.hse.ru/data/2013/06/03/1285521221/report_v2.pdf>

Статья из вшэ

<https://www.keldysh.ru/papers/2007/prep89/prep2007_89.html#_Toc185855267>