

Описание

AGFL 2.0 parser generator system for Windows

Введение

AGFL (Affix Grammar over a Finite Lattice) является формальной аффиксальной грамматикой, специально разработанной для описания синтаксиса естественных языков. Как и другие аффиксальные грамматики, в AGFL есть первый (или нижний) уровень, состоящий из контекстно-свободных синтаксических правил, переписывающих нетерминальные символы в терминальные или другие нетерминальные символы, и второй уровень, который добавляет к правилам первого уровня параметры (аффиксы).

Формализм AGFL был реализован в программном **продукте AGFL 2.0 parser generator system for Windows**¹ (Nijmegen University, Subfaculty of Computer Science), который описывается ниже.

Файлы дистрибутива

Дистрибутив системы может быть загружен с AGFL сайта <http://www.cs.kun.nl/agfl/> . Он включает в себя четыре архивных файла:

- **epforir.zip**
- **agfl-dev_v2-0_win32_sources.zip**
- **agfl-rt_v2-0_win32.zip**
- **cdl3-rt_v2.01_win32.zip**

epforir.zip

Включает в себя все грамматические файлы (*.gra), файлы, задающие набор известных системе терминальных символов с их грамматической характеристикой (*.dat), а так же файл **epforir.lex**, объединяющий .dat файлы для разных частей речи.

agfl-dev_v2-0_win32_sources.zip

Содержит файлы с исходным текстом программ (определяющих работу всех модулей системы), написанных на языке программирования C++ (*.cc/*.c), соответствующие *.h файлы, необходимые dll-библиотеки (*.dll), а так же *.k3 файлы, определяющие последовательность действий при парсинге.

agfl-rt_v2-0_win32.zip

Содержит скомпилированные C++ файлы, то есть исполнимые *.exe файлы, объектные файлы (*.o), *.a файлы и некоторые dll-библиотеки.

cdl3-rt_v2.01_win32.zip

Программы ввода/вывода и работы с библиотеками.

Основные принцип работы системы

В первую очередь, после запуска **setagfl.bat** система “прописывает” себя в файле autoexec.bat, определяя пути к подкаталогам в папке **c:\agfl**. Следующим этапом является уже само создание на основе пользовательской корневой грамматики (напр. **mygram.gra**) исполнимого .exe файла, при помощи которого можно осуществлять парсинг входного текста. После ввода в командной строке

```
c:\agfl\bin gen mygram
```

Последовательно запускаются .exe файлы из каталога **bin** которые:

- Считывают корневую грамматику **mygram**;
- Считывают все грамматики, на которые существуют ссылки в корневой грамматике;

¹ Более ранние версии существуют для платформы UNIX

- Считывают корневой лексический файл **.lex**;
- Считывают все **.dat** файлы, на которые существуют ссылки в **.lex** файле;
- Создается **.lif** файл (**lexicon input file**), задающий соответствие между лексиконом и грамматикой;
- Осуществляется токенизация **.lex** и ***.gra** файлов;
- Создается **.blf** файл (**Binary Lexicon File**), который в дальнейшем используется скомпилированным парсером для парсинга произвольного текста на данном языке как машинный файл с лексической информацией;
- Создается объектный файл **.o**, который связывается со скомпилированными заранее С-библиотеками;
- Создается исполнимый **.exe** файл, который и является программой парсинга текста на данном языке.

Назначение компонентов системы и ее структура

Все файлы системы можно условно разделить на четыре группы, согласно их назначению. Это файлы с исходным текстом программ; скомпилированные файлы; файлы для создания системы парсинга английского текста; файлы созданные для парсинга на основе конкретной грамматики.

Файлы с исходным текстом программ

Этими файлами являются все файлы, включенные в архив **agfl-dev_v2-0_win32_sources.zip**. Поскольку **AGFL 2.0** является открытой некоммерческой системой, то данные файлы включены в пакет на тот случай, если пользователю будет нужно изменить ход работы программы или воспользоваться лишь ее отдельными компонентами. Кроме того, эти файлы предоставляют возможность разобраться в основных принципах работы системы и назначении ее отдельных модулей.

Скомпилированные файлы

Это результат компиляции исходных файлов первой группы и некоторые библиотечные файлы. Можно сказать, что эти файлы и составляют систему **AGFL 2.0 parser generator system**. Их главное назначение состоит в создании на основе пользовательской грамматики и лексикона программы, позволяющей производить парсинг текста на данном языке. Пользователь должен составить **AGFL** описание интересующего его языка, которое должно храниться в различных ***.gra** файлах; корневую грамматику, объединяющую остальные ***.gra** файлы; ***.dat** файлы, составляющие, по сути, лексикон данного языка с грамматической характеристикой терминальных символов, а так же **.lex** файл, объединяющий ***.dat** файлы.

После компиляции этих файлов создаются файлы четвертой группы.

Файлы для создания системы парсинга английского текста

Разработчики формализма **AGFL** предоставляют загрузить со своего сайта не только саму систему **AGFL 2.0 parser generator**, но и все файлы, необходимые для генерации парсера английского текста (архив **epforir.zip**).

Файлы созданные для парсинга на основе конкретной грамматики

Эти файлы являются результатом компиляции пользовательских грамматических и лексических файлов. Если корневая грамматика называется, например, **epforir.gra**, то такими файлами будут **epforir.o**, **epforir.blf** и **epforir.exe**. Эти файлы являются необходимыми для дальнейшего парсинга произвольных предложений на основе пользовательской грамматики.