HTK Tool Kit

HTK Tool Kit

What is HTK tool kit

The HTK language modeling tools are a group of programs designed for constructing and testing statistical *n-gram* language models

HTK Tool Kit

What to prepare

Training & Test Text

Dictionary

HTK Tool Kit

Training & Test Text

- Plain text sentences
- One sentence per line
- Sentence starts with <s>
- Sentence ends with </s>

HTK Tool Kit

Training Text Sample

<s> IT WAS ON A BITTERLY COLD NIGHT AND FROSTY MORNING TOWARDS THE END OF THE WINTER OF NINETY SEVEN THAT I WAS AWAKENED BY A TUGGING AT MY SHOULDER </s>

<s> IT WAS HOLMES </s>

HTK Tool Kit

Dictionary

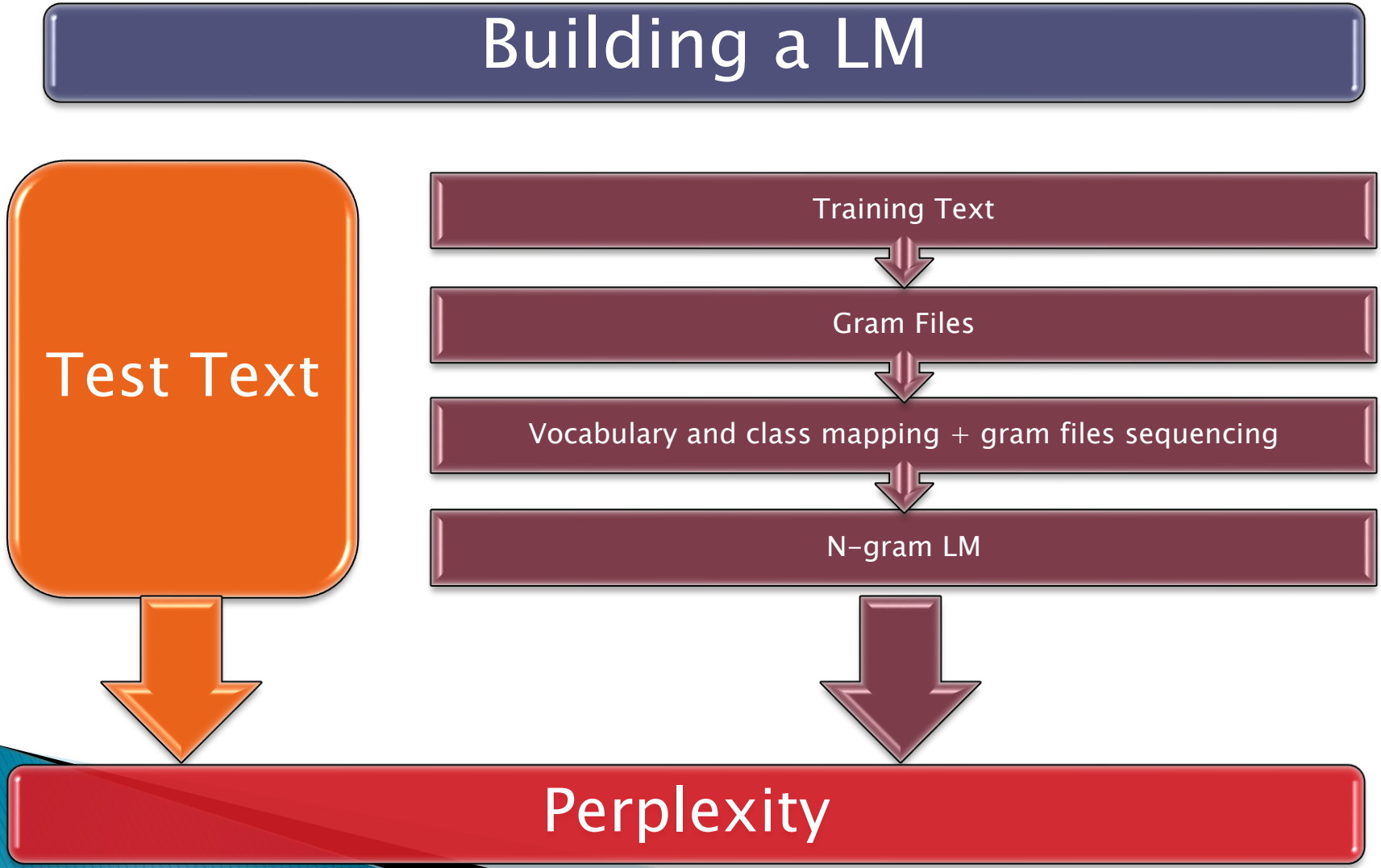
- Plain text wordlist
- One word per line
- Alphabetically ordered

HTK Tool Kit

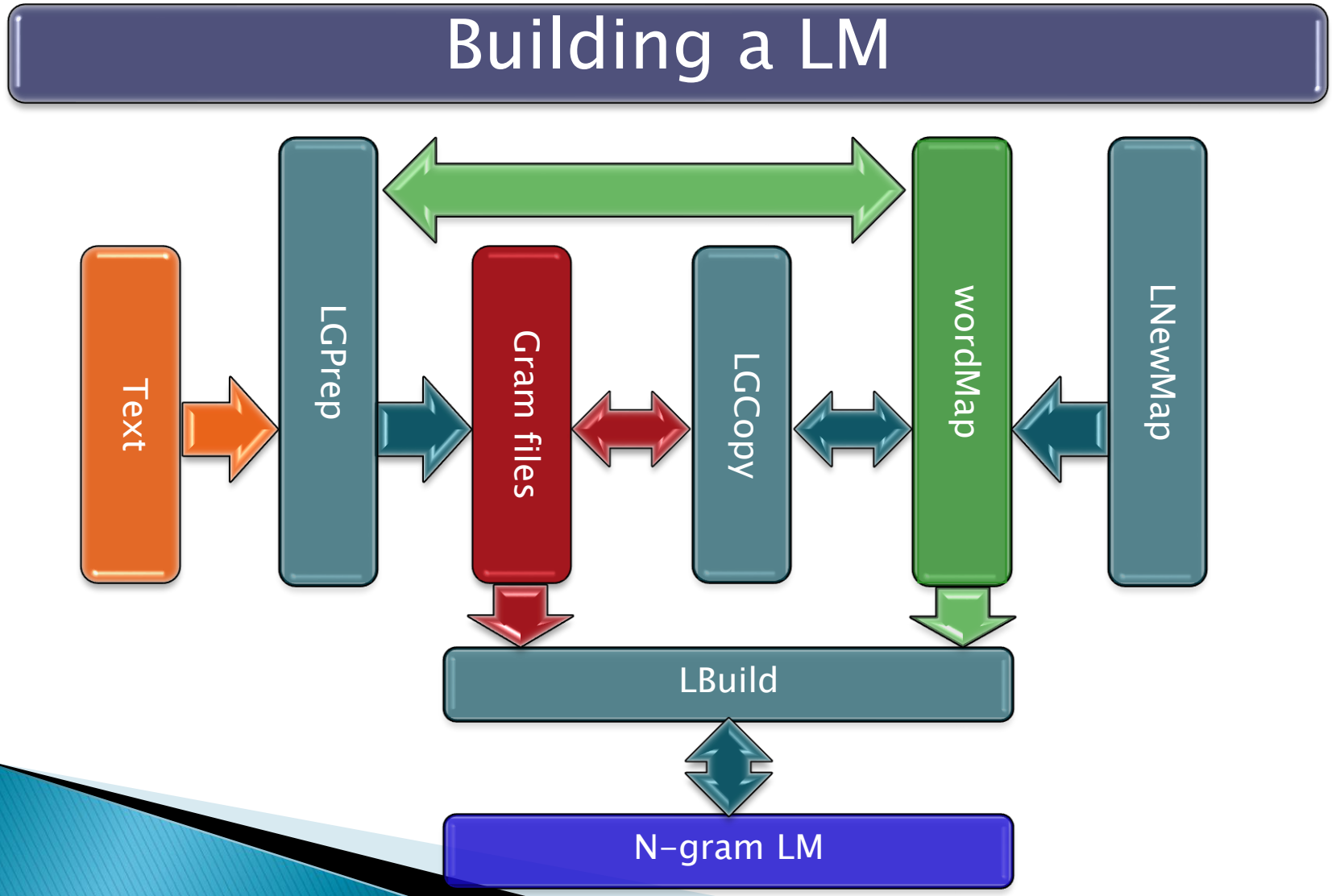
Dictionary Sample

</s>
<s>
A
A.
ABANDON
ABANDONED
ABBEY
ABDULLAH
ABE

HTK Tool Kit


Building a LM
Test Text
Training Text
Gram Files
Vocabulary and class mapping + gram files sequencing
N-gram LM
Perplexity

HTK Tool Kit


Building a LM
Text
LGPrep
Gram files
LGCopy
wordMap
LNewMap
LBuild
N-gram LM

HTK Tool Kit

LNewMap

LNewMap [options] name mapfn

-e esc Change the contents of the EscMode header to esc.
Default is RAW.

-f fld Add the field fld to the Fields header.

HTK Tool Kit

LNewMap

Example:

LNewMap -f WFC Holmes empty.wmap

```
Name = Holmes
SeqNo = 0
Entries = 0
EscMode = RAW
Fields = ID,WFC
\Words\
```

HTK Tool Kit

LGPrep

LGPrep [options] wordmap [textfile ...]

-a n Allow upto n new words in input texts (default 100000).

-b n Set the internal gram buffer size to n (default 2000000). LGPrep stores incoming n-grams in this buffer. When the buffer is full, the contents are sorted and written to an output gram file. Thus, the buffer size determines the amount of process memory that LGPrep will use and the size of the individual output gram files.

HTK Tool Kit

LGPrep cont'd

LGPrep [options] wordmap [textfile ...]

-d Directory in which to store the output gram files (default current directory).

-i n Set the index of the first gram file output to be n (default 0).

-n n Set the output n-gram size to n (default 3).

-r s Set the root name of the output gram files to s (default "gram").

HTK Tool Kit

LGPrep cont'd

LGPrep [options] wordmap [textfile ...]

-s s Write the string s into the source field of the output gram files. This string should be a comment describing the text source.

-z Suppress gram file output. This option allows LGPrep to be used just to compute a word frequency map. It is also normally applied when applying edit rules to the input.

HTK Tool Kit

LGPrep cont'd

Example:

LGPrep -T 1 -a 100000 -b 2000000 -d holmes.0 -n 4 -s "Sherlock Holmes" empty.wmap D:\train\abbey_grange.txt, D:\train\beryl_coronet.txt,...

HTK Tool Kit

LGPrep cont'd

WMAP file

```
Name = Holmes
SeqNo = 1
Entries = 18080
EscMode = RAW
Fields = ID,WFC
\Words\
<s>     65536  33669
IT      65537  8106
WAS     65538  7595
...
```

HTK Tool Kit

LGCopy

LGCopy [options] wordmap [mult] gramfiles

-b n Set the internal gram buffer size to n (default 2000000). LGPrep stores incoming n-grams in this buffer. When the buffer is full, the contents are sorted and written to an output gram file. Thus, the buffer size determines the amount of process memory that LGPrep will use and the size of the individual output gram files.

-d Directory in which to store the output gram files (default current directory).

HTK Tool Kit

LGCopy cont'd

LGCopy [options] wordmap [mult] gramfiles

-o n Output class mappings only. Normally all input *n*-grams are copied to the output, however, if a class map is specified, this options forces the tool to output only *n*-grams containing at least one class symbol.

HTK Tool Kit

LGCopy cont'd

Example:

LGCopy -T 1 -b 2000000 -d D:\holmes.1
D:\ holmes.0\wmap D:\ holmes.0\gram.1 D:\
holmes.0\gram.2.....

HTK Tool Kit

LBuild

LBuild [options] wordmap outfile [mult] gramfile ..

-c n c Set cutoff for n-gram to c.

-n n Set final model order to n.

HTK Tool Kit

LBuild cont'd

Example:

LBuild -T 1 -c 2 1 -c 3 1 -n 3 D:\lm_5k\5k.wmap
D:\lm_5k\tg2-1_1 D:\holmes.1\data.1
D:\holmes.1\data.2... D:\lm_5k\data.1 D:\lm_5k\data.12

HTK Tool Kit

LPlex

LPlex [options] langmodel labelFiles

-n n Perform a perplexity test using the n-gram component of the model. Multiple tests can be specified. By default the tool will use the maximum value of n available.

-t Text stream mode. If this option is set, the specified test files will be assumed to contain plain text.

HTK Tool Kit

LPlex cont'd

Example:

Lplex -n 3 -t D:\lm_5k\tg1_1 D:\test\red-headed_league.txt