利用现有软件构建统计机器翻译系统

刘洋

中国科学院计算技术研究所数字化技术研究室

yliu@ict.ac.cn

1 介绍

本文的主旨是介绍如何利用现有软件在很短的时间内构建一个统计机器翻译（Statistical Machine Translation）系统。重点将放在具体的操作上，而不是相关的理论。

2 准备工作

为了构建统计机器翻译系统，必须准备一下资源：

[1] **Linux 操作系统（附带 gcc）**

我用的是 Red Hat Linux 3.2.2-5，gcc 的版本是 3.2.2 20030222。如果您不知道自己的 Linux 系统的版本,可以输入以下命令查看:gcc -v。一般而言，只要版本不要太低的 Linux 都能胜任。

[2] **双语对齐语料库**

在本文演示的是建造将汉语翻译成英语的统计机器翻译系统，因此采用的汉英对齐语料库。我使用的是下载自http://www.nlp.org.cn/上的双语句对齐语料库，规模为 1500 句对。

[3] **目标语言语料库**

目标语言语料库是用做语言模型训练，在此为了简便，我采用了上述语料库中的 1500 句英语句子。

[4] **CMU-Cam Language Model Toolkit v2**

这是一个语言模型工具，用于生成语言模型，以便解码器调用。

[5] **GIZA++ v2 (2003-09-30)**

这是一个翻译模型工具，实现了 IBM Model 1-5，并且加入了一些新的特色。其前身是 GIZA。

[6] **mkcls (2003-09-30)**

这是生成 word class 的工具，是作为 GIZA++的辅助程序。

[7] **汉语的切分工具和英语的 tokenize 工具**

这是对语料进行预处理时需要使用的工具。汉语的切分工具可以使用 LDC Chinese Segmenter，英文的 tokenize 工具可以选择 EGYPT 软件包中内含的一个工具 tokenizeE.perl.tmpl。

3 总体流程

有了这些资源，我们就可以开始动手了。总体流程是这样的：

[1] 语料准备

[1] 构建语言模型

[2] 构建翻译模型

[3] 构建解码器

4 语料准备

首先您需要编程(用您偏爱并且能熟练使用的编程语言)将汉语句子和英语句子分别从1500 句对中抽取出来存在两个文本文件中，1500 个汉语句子存放在文件 chinese 中，1500 个英语句子存放在 english 中。每个句子一行，并且汉英对应句子的行号一一对应。然后，您需要对 chinese 中的汉语句子进行切分，也就是切成一个个的汉语词。对于 english 中的英语句子进行 tokenize。

之后 english 用做语言模型的训练语料，chinese 和 english 用做翻译模型的训练语料。

5 构建语言模型

5.1 语料预处理

构建语言模型要用到的语料是 english，但是需要对它进行一些改动。由于 ISI Rewrite Decoder 采用 XML 文件格式作为输入文件，有一些标记如<s>和</s>会用到。ISI Rewrite Decoder 要求语言模型必须能够识别<s>和</s>，把它当作一个句子的开始。为此，在构建语言模型时我们需要做两件事：

[1] 写一个 Context Cue File(.ccs)，让语言模型知道<s>和</s>是标记，而不是词汇。

[2] 在训练语料中包含<s>和</s>，这样在语言模型生成的词典中能包含<s>和</s>

这两个条件缺一不可。

5.2 编译源代码

下面开始才操作，我建立了一个文件夹/home/lonios/research/ICTSMTS，下载了 CMU-Cam_Toolkit_v2.tar.gz，将解压后的文件夹 CMU-Cam_Toolkit_v2 复制到 ICTSMTS 目录中。

目录/home/lonios/research/ICTSMTS/CMU-Cam_Toolkit_v2 下应当包含 5 个文件夹和两个文件：

文件夹 bin, doc, include, lib, src

文件 endian.sh, README

首先当然要看一下 README，里面讲述了编译源代码的方法。然后，进入 src 目录，找到 Makefile，用 vi 打开，将"#BYTESWAP_FLAG = -DSLM_SWAP_BYTES"中的"#"去掉即可。如

果使用的是 PC，就必须这样做。之后键入命令“make install”，这样就会编译源代码。编译成功后，去 bin 目录看看，就会发现已经生成了 12 个文件：

binlm2arpa，evallm，idngram2lm，idngram2stats，interpolate，mergeidngram，ngram2mngram，text2idngram，text2wfreq，text2wngram，wfreq2vocab，wngram2idngram

关于这 12 个文件的用法，请您参考相关文档。

5.3 生成语言模型

将 english 复制到 bin 目录中，将其重命名为“a.text”。然后在 a.text 中添加“<s>”和“</s>”，最好是分别置于一个句子的首尾。注意和句子中的其他词保持至少一个空格。这样才能保证“<s>”和“</s>”能出现在即将生成的词汇表中。

输入命令“./text2wfreq <a.text> a.wfreq”，这样就会生成 a.wfreq 文件。

输入命令“./wfreq2vocab <a.wfreq> a.vocab”，这样就会生成 a.vocab 文件。此时查看 a.vocab 文件，会发现“<s>”和“</s>”出现在词汇列表里面了。

输入命令“./text2idngram -vocab a.vocab -buffer 5 <a.text> a.idngram”，这样就生成 a.idngram 文件。

最后一步就是生成语言模型了，之前必须写一个 a.ccs 文件，来表明“<s>”和“</s>”是标记。a.ccs 文件的内容如下：

<s>

</s>

就这么简单，两个标记，一个一行。

然后,键入命令“./idngram2lm -idngram a.idngram -vocab a.vocab -context a.ccs -binary a.binlm”。OK，这样我们就得到了一个二进制文件 a.binlm，这就是语言模型！利用 evallm 程序，就可以计算任意英文句子的 P(e)了。

6 构建翻译模型

6.1 生成 word class

下载 mkcls.2003-09-30.tar.gz，解压后复制到 ICTSMTS 目录下，进入 mkcls-v2 目录。输入命令“make”，这样就会编译生成 mkcls。在 mkcls-v2 目录下建立一个子目录 temp，将 mkcls、chinese 和 english 拷贝到 temp 目录中.

输入命令“./mkcls -c80 -n10 -pchinese -Vchinese.vcb.classes opt”，生成两个文件：chinese.vcb.classes 和 chinese.vcb.classes.cats。输入命令“./mkcls -c80 -n10 -penglish -Venglish.vcb.classes opt”， 生成两个文件：english.vcb.class 和 english.vcb.classes.cats。

6.2 编译并运行 GIZA++

下载 GIZA++.2003-09-30.tar.gz，解压后复制到 ICTMTES 目录下，进入 GIZA++-v2 目录，输入命令“make”，就可以编译生成 GIZA++。同时还生成一个 plain2snt.out。

输入命令“./plain2snt.out chinese english”，生成四个文件：chinese.vcb，english.vcb，chinese_english.snt 和 english_chinese.snt。

在 GIZA++-v2 目录建立一个子目录 test，将 GIZA++，chinese.vcb，english.vcb，chinese_english.snt，chinese.vcb.classes，chinese.vcb.classes.cats，english.vcb.class 和 english.vcb.classes.cats 复制到 test 目录里。

输入命令“./GIZA++ -S english.vcb -T chinese.vcb -C english_chinese.snt”。随后就开始 IBM Model 1-5 的训练，生成许多文件，不再详述这些文件。

7 构建解码器

7.1 设定环境变量

下载 isi-rewrite-decoder-r1.0.0a.tar.gz，解压缩后复制到 ICTSMTS 目录，进入 isi-rewrite-decoder-r1.0.0a 目录。进入 linux 目录。输入命令

```
export LD_LIBRARY_PATH="/home/lonios/research/ICTSMTS/isi-rewrite-decoder-r1.0.0a/linux
```

7.2 编写配置文件

编写配置文件 decoder.cfg，内容如下:

```
LanguageModelFile = /home/lonios/research/ICTSMTS/CMU-Cam_Toolkit_v2/bin/a.binlm
TranslationModelConfigFile =
/home/lonios/research/ICTSMTS/GIZA++-v2/test/104-06-11.161715.lonios.Decoder.config
PrintAlignment = true
PrintProbabilities = true
```

7.3 编写输入文件

编写输入文件 input.xml，内容如下：

```
<?xml version="1.0" encoding="gbk"?>
<doc>
<s id="1">中国 孩子 都 十分 活泼 .</s>
<s id="2">政府 应该 大力 促进 经济 发展 .</s>
<s id="3">谁 是 我们 足球队 里 最 强壮 的 人 ?</s>
</doc>
```

7.4 复制相关文件

将/home/lonios/research/ICTSMTS/GIZA++-v2/test/中的所有文件复制到/home/lonios/research/ICTMTS/isi-rewrite-decoder-r1.0.0a/linux/中

7.5 生成 FZeroWords

查看 104-06-11.161715.lonios.Decoder.config 文件，在最后一行会发现需要一个 104-06-11.161715.lonios.fe0_3.final 文件，可是 GIZA++并不生成这个文件！

怎么办呢？将 104-06-11.161715.lonios.n3.final 和 104-06-11.161715.lonios.trn.src.vcb 这两个文件复制到 /home/lonios/research/ICTSMTS/isi-rewrite-decoder-r1.0.0a/util 下，这里面有个工具叫 rewrite.mkZeroFert.perl，我们要用到它。输入命令：

./rewrite.mkZeroFert.perl 104-06-11.161715.lonios.trn.src.vcb 104-06-11.161715.lonios.n3.final

这时候终端上会打印以下内容：

the 0.978169
of 0.977899
to 0.962034
, 0.879314
in 0.839314
and 0.809735
a 0.783997

这说明这些词是 FZeroWords，在 /home/lonios/research/ICTSMTS/isi-rewrite-decoder-r1.0.0a/linux 下编写 104-06-11.161715.lonios.fe0_3.final 文件，内容如下：

the
of
to
,
in
and
a

7.6 终于可以运行了!!!

经过许多复杂的配置，现在终于可以运行了，回到 /home/lonios/research/ICTSMTS/isi-rewrite-decoder-r1.0.0a/linux 目录，输入以下命令：

./decoder.linux.public --config decoder.cfg input.xml

然后，您就可以在屏幕上看到那三个句子的译文了！

8 后记

本文主要描述了构建一个汉译英统计机器翻译系统的主要步骤。关于具体每个指令的众多可选项并没有展开讨论，这需要您自己去探索体会。另外，每个软件包都有文档和说明，您可以通过阅读这些资料获得更深刻的认识。

希望通过这篇短文，您也能构建一个自己的 SMTS！

历史

2004 年 6 月

初稿。

2004 年 9 月

主要是用 GIZA++自带的 plain2snt.out 代替 EGYPT 里的 whittle，同时修正了一些错误。