Homework #2: Statistical Machine Translation

Anoop Sarkar – anoop@cs.sfu.ca

(1) **Decoder for Phrase-based Statistical Machine Translation**

Given an input sentence $\mathbf{f}$ we wish to produce the best translation $\mathbf{e}$ according to a phrase-based machine translation model which uses a model $\Pr(\mathbf{e} \mid \mathbf{f})$ in order to find $\mathbf{e}$. The basic model consists of the following components:

1. the phrase translation probability $\phi(\bar{f} \mid \bar{e})$.
2. the re-ordering model $d(x) = \alpha^{\text{abs}(x)}$, and $\alpha \in (0, 1]$ is set by hand or tuning data.
3. the language model $p_{\text{LM}}(\mathbf{e})$.

The source sentence $\mathbf{f}$ is split up into I phrases $\bar{f}_i, 1 \leq i \leq I$. However, this split is not unique. There are many different values of I and we search over all of them. We look up the phrase table to get phrase pair probability $\phi(\bar{f}_i, \bar{e}_i)$ for all $\bar{e}_i$. To speed up decoding we will limit ourselves to the top ten $\bar{e}_i$ sorted by value of $\phi(\bar{f}_i, \bar{e}_i)$. The most likely $\mathbf{e}$ is computed using the following method:

$$
\begin{aligned}
\mathbf{e}_{\text{best}} &= \operatorname*{argmax}_{\mathbf{e}} \Pr(\mathbf{e} \mid \mathbf{f}) \\
&= \operatorname*{argmax}_{\mathbf{e}=\bar{e}_1,\ldots,\bar{e}_I} \prod_{i=1}^{I} \left(\phi(\bar{f}_i \mid \bar{e}_i) \times d(\text{start}_i - \text{end}_{i-1} - 1)\right) \times \prod_{i=1}^{I} p_{\text{LM}}(e_i \mid e_1, \ldots, e_{i-1})
\end{aligned}
$$

You do not need to add exponents $\lambda_\phi, \lambda_d, \lambda_{\text{LM}}$ for the three components of the model above for this homework.

Given a suitable phrase table (provided to you) implement the pseudo-code for stack decoding given in Figure 6.6 in the Koehn SMT book. The decoder will run faster if you assume a distortion limit of 5-10 but this is optional. A small phrase table for testing your decoder is provided to you.

Provide the decoder source code along with instructions to run on the sample test set (provided to you). If you use any parameters in your decoder (e.g. α) then include values for those parameters so that your decoder is ready to run on input sentences.

Previously trained language models are also provided to you along with an API that will let you query the language model from programming languages such as C, C++ or Python.

(2) **Competitive Grammar Writing**

This question involves writing or creating weighted context-free grammar files in order to parse English sentences and utterances. The vocabulary is fixed but the task is not trivial: you can choose to either write a weighted grammar by hand or you can use the various methods we have learned about in this course to automatically infer a weighted grammar from a suitably chosen set of data. One important thing to keep in mind is that you can solve this question without writing any code, but you probably will need to write some code to get a high ranking.

You will need to think about the following aspects of grammar development:

Linguistic Analysis You will need to consider different types of *linguistic phenomena* like *wh*-questions, agreement, adjunct clauses, embedded sentences, etc. The NLTK book contains a guide to some basic grammar development, but covering a large number of phenomena, even obscure ones like clefts, will help improve your rank in this question (check the detailed rules below to see why). Your previous homework contained some initial steps towards this kind of grammar development. The homework directory contains some sample grammars (explained in detail below) with default part of speech categories and weights. You can change or modify the provided grammars, especially the weights, as long as you do not change the vocabulary: *increase or decrease of the vocabulary is not allowed.* The grammar must be in **extended Chomsky Normal Form** (eCNF): rules are always of the form $A \rightarrow B\ C$, $A \rightarrow B$, or $A \rightarrow a$ where A, B, C are non-terminals and a is a terminal symbol (word in the vocabulary).

Parameter Tuning You can change the weights associated with rules. The weights can be anything you want, and should reflect the relative importance of the rules. This is especially true since in this question we will also be generating sentences from the grammars you write. So giving large weights to recursive rules can lead to the generation of very long sentences, or even a run-away generative process. You should also be aware that you may need to change the weights associated with the rules in the default grammar files provided in the homework directory.

Quantitative Evaluation This question will involve a very rigorous quantitative evaluation of your submitted grammar. Keep the details of the evaluation in mind as you develop your grammars. The goal is to get the highest evaluation score and to make it harder for others to reach your score. The evaluation is complicated because it tries to avoid some kinds of cheating to improve the ranking, so there is a whole section below devoted to explaining the evaluation procedure.

The Data All the data you need for grammar development is provided to you in the following files.

`S1.gr` The default main grammar file which contains a weighted context-free grammar in extended Chomsky Normal Form (eCNF).

$$
\begin{array}{rrcr}
1 & \mathit{S1} & \rightarrow & \mathit{NP\ VP} \\
1 & \mathit{S1} & \rightarrow & \mathit{NP\ \_VP} \\
1 & \mathit{\_VP} & \rightarrow & \mathit{VP\ Punc} \\
20 & \mathit{NP} & \rightarrow & \mathit{Det\ Nbar} \\
1 & \mathit{NP} & \rightarrow & \mathit{Proper}
\end{array}
$$

Notice the non-terminal $\_VP$ which is used to keep the grammar in eCNF. The probability of a particular rule is obtained by normalizing the weights for each left-hand side non-terminal in the grammar. For example, for rule $NP \rightarrow Det\ Nbar$ the probability is $20/(20+1)$.

`S2.gr` The default backoff grammar file. This grammar file ensures that any input constructed using the allowed vocabulary will be accepted by the parser (although it may accept it with

low probability depending on the weights). The files `S1.gr` and `S2.gr` are connected via the following rules in `S1.gr`:

$$\begin{array}{rrcr} 99 & \textit{TOP} & \rightarrow & \textit{S1} \\ 1 & \textit{TOP} & \rightarrow & \textit{S2} \\ 1 & \textit{S2} & \rightarrow & \textit{Misc} \end{array}$$

Notice how the grammar in `S1.gr` is highly weighted (99 times out of 100) so that `S2.gr` is used as a backoff grammar. You may want to consider optimizing this weight to improve your evaluation score. *Misc* expands to vocabulary items defined in `Vocab.gr`.

One way to improve the weights for the rules in `S2.gr` is to use Hidden Markov Model learning, but there are many other ways to obtain a good backoff grammar.

`Vocab.gr` The vocabulary file containing rules of the type $A \rightarrow a$ where A is a part of speech tag and a is a word that will appear in the input sentences. Many words are assigned to the default part of speech *Misc*, so you will probably want to re-assign these words to more useful part of speech tags in order to aid your grammar development.

`allowed_words.txt` This file contains all the words allowed in the evaluation. You should make sure that your grammar does not generate any words not in this file. It does not specify the part of speech for each word, so you can choose to model the ambiguity of words in terms of part of speech in your `Vocab.gr` file.

`cgw-devset.txt` This file contains example sentences that you can use as a starting point for your grammar development. Only the first two sentences of this file can be parsed using the default `S1.gr` grammar. The rest are parsed with the backoff `S2.gr` grammar. You can augment this data with any other data you can find on the web (but make sure you do not expand the vocabulary).

`unseen.tags` Used to deal with unknown words. You should not have to use this file during parsing, but the parser provided to you can optionally use this file in order to deal with unknown words in the input. Since the vocabulary is fixed for this question you do not need to use this file, but it is provided in case you want to try the parser on your own data.

The Parser and Generator You are given a parser that takes sentences as input and produces parse trees and also a generator which generates a random sample of sentences from the weighted grammar. Parsing and generating will be useful steps in your grammar development strategy. You can learn the various options for running the parser and generator using the following command. The parser has several options to speed up parsing, such as beam size and pruning. Most likely you will not need to use those options (unless your grammars are huge).

```
$ python2.6 pcfg_parse_gen.py -h
```

Parsing input: The parser provided to you reads in the grammar files and a set of input sentences. It prints out the single most probable parse tree for each sentence (using the weights assigned to each rule in the input context-free grammar). The parser also reports the negative cross-entropy score for the whole set of sentences. Assume the parser gets a text of n sentences to parse: $s_1, s_2, \ldots, s_n$ and we write $|s_i|$ to denote the length of each sentence s_i. The probability assigned to each sentence by the parser is $P(s_1), P(s_2), \ldots, P(s_n)$. The negative cross entropy is the average log probability score (bits per word) and is defined as follows:

$$\text{score}(s_1, \ldots, s_n) = \frac{\log P(s_1) + \log P(s_2) + \ldots + \log P(s_n)}{|s_1| + |s_2| + \ldots + |s_n|}$$

We keep the value as negative cross entropy so that higher scores are better. For example, running the parser with the default grammar files on the development set of sentences `cgw-devset.txt` gives the following output (ignoring the output parse trees for each input sentence):

```
$ python2.6 pcfg_parse_gen.py -i -g "*.gr" < cgw-devset.txt
#loading grammar files: S1.gr, S2.gr, Vocab.gr
#reading grammar file: S1.gr
#reading grammar file: S2.gr
#reading grammar file: Vocab.gr
#skipping comment line in input: # this is the devset for hw5 q9
#-cross entropy (bits/word): -10.0557
```

Generating output: In order to aid your grammar development you can also generate sentences from the weighted grammar to test if your grammar is producing grammatical sentences with high probability. The following command samples 20 sentences from the `S1.gr,Vocab.gr` grammar files. Always sample only from these two files and ignore the backoff grammar (as you will see from the evaluation procedure).

```
$ python2.6 pcfg_parse_gen.py -o 20 -g S1.gr,Vocab.gr
#loading grammar files: S1.gr, Vocab.gr
#reading grammar file: S1.gr
#reading grammar file: Vocab.gr
```

```
every pound covers this swallow
no quest covers a weight
Uther Pendragon rides any quest
the chalice carries no corner .
any castle rides no weight
Sir Lancelot carries the land .
a castle is each land
every quest has any fruit .
no king carries the weight
that corner has every coconut
the castle is the sovereign
the king has this sun
that swallow has a king
another story rides no story
this defeater carries that sovereign
each quest on no winter carries the sovereign .
another king has no coconut through another husk .
a king rides another winter
that castle carries no castle
every horse covers the husk .
```

During the time that the homework is assigned to you, you are allowed to share samples of your output with others in the class (the more students share samples the better it is for everyone). To avoid flooding the mailing list with samples from your grammars, please use an online pastebin, e.g. the site *pastebin.mozilla.org*, to create a pastebin with your samples and mail the link to the mailing list. You can edit the pastebin with additional samples as you develop better grammars but please do not email the mailing list each time you update the pastebin. For example, the above set of 20 sentences sampled from the default grammar is available at *http://pastebin.mozilla.org/1162241*.

You should modify your grammar to improve the cross entropy score assigned to samples from the grammars written by others. **Important:** You cannot share your grammars, only the samples from your grammars.

The Evaluation For this question you should submit your grammar files `"*.gr"` which must contain at least one rule with the start symbol `TOP` and your submission must have at least one file called `S1.gr`. *Please do not put the grammar files into subdirectories. We want to run an automatic script on your grammar files, so please keep them at the top directory level in your submission.* Based on the `"*.gr"` you will submit as your solution to this question, we will use the following steps in order to provide a ranking to each submission.

Step 1: Sample sentences We will sample 20 sentences from the `S1.gr` and `Vocab.gr` file from each submission using the following command (this example uses the default grammar files):

```
$ python2.6 pcfg_parse_gen.py -o 20 -g S1.gr,Vocab.gr
#loading grammar files: S1.gr, Vocab.gr
#reading grammar file: S1.gr
#reading grammar file: Vocab.gr
```

If your submitted grammar files cause an error in the sentence generator then you will receive zero marks for this question. We might have to delete some lines from your output for misbehaving incomplete submissions.

In order to score a high rank, it is important to note that the `S2.gr` file is **not** used to sample sentences.

Step 2: Parse open test set We will concatenate the 20 sentence samples obtained from all the submissions into a single file called `cgw-testset.txt`. We will parse this file with the parser using your grammar files (including `S2.gr`). The score obtained on `cgw-testset.txt` will be part of the rank given to your grammar submission.

Step 3: Parse with test set During testing your submission we will parse a file called `cgw-holygrail.txt` using your `S1.gr` and `Vocab.gr` grammar files. The file `cgw-holygrail.txt` contains previously unseen sentences from the same domain as the file `cgw-devset.txt` (the sentences use the words given in `allowed_words.txt`). The score obtained on `cgw-holygrail.txt` will be part of the rank given to your grammar submission, but you will not have access to this file during your grammar development process. The reason to include this test is to penalize submissions that use the following trick to get a high ranking: put a very small weight on `S1.gr` and put a high weight on `S2.gr` and then your grammar will get a score for the sentences in `cgw-testset.txt` mainly using your `S2.gr`. With this weighting the trick is to use ungrammatical sentences in `S1.gr` which is used for sampling sentences for others. This trick will not work since we use `S1.gr` to parse the unseen text file `cgw-holygrail.txt` and the ranking depends on this score as well.

Step 4: Rank each submission Each submission will be ranked using the following formula:

$$\text{rank} = 0.5 \times \text{score}(\texttt{cgw-testset.txt}) + 0.5 \times \text{score}(\texttt{cgw-holygrail.txt})$$

We will look at your submission to assign a grade for this question – especially the effort made in `S1.gr` and in `S2.gr`. You can submit a text file called `cgw-readme.txt` that includes your strategy for obtaining a high rank and the process you used to write or automatically produce `S1.gr` and `S2.gr`.

Further Reading This question is an adaptation of the idea presented in the following paper:

> Jason Eisner and Noah A. Smith. Competitive Grammar Writing. In *Proceedings of the ACL Workshop on Issues in Teaching Computational Linguistics*, pages 97-105, Columbus, OH, June 2008. *http://aclweb.org/anthology/W/W08/W08-0212.pdf*

You can read this paper for a discussion on how to get the best score on this question, but be aware that some of the details have been changed, so follow the instructions provided here.