The effect of non-tightness on Bayesian estimation of PCFGs

Shay B. Cohen and Mark Johnson (Proceedings of ACL, 2013)

This text describes step-by-step procedures to indicate the difference that the three approaches for handling tightness of PCFGs in Bayesian estimation have on the posterior
over syntactic structures with a uniform prior.
See paper for more discussion of this *Mathematica* procedure.

General problem setting

In this note, we describe a step-by-step procedure in *Mathematica* that follows from section 8 in the paper.

We assume we have the following grammar:

S -> S S S (rule probability p)
S -> S S (rule probability q)
S -> x (rule probability 1-p-q)

We are going to calculate the posterior over the three possible trees for the string w = “x x x”.

The three possible trees (t1, t2 and t3) are:

1. (S (S x) (S x) (S x))
2. (S (S x) (S (S x) (S x)))
3. (S (S (S x) (S x)) (S x))

Note that the second and third trees have identical probabilities.

We assume a uniform prior over all grammar rules, so that the prior is proportional to a constant.

The measure of t1, t2 and t3 are:

```
measureT1 := p (1 - p - q) ^3

In[13]:= measureT2 := q^2 (1 - p - q) ^3

measureT3 := q^2 (1 - p - q) ^3
```

We begin by calculating a function that computes the partition function for this grammar,following Chi (1999):

In[12]:=
```
sol = z /. Solve[z == p z^3 + q z^2 + 1 - p - q, z]
```

Out[12]=
$$\left\{1, \frac{-p-q-\sqrt{4\,p-3\,p^2-2\,p\,q+q^2}}{2\,p}, \frac{-p-q+\sqrt{4\,p-3\,p^2-2\,p\,q+q^2}}{2\,p}\right\}$$

In[2]:=
```
Z = Min[1, If [sol[[2]] < 0, 10000, sol[[2]]], If [sol[[3]] < 0, 10000, sol[[3]]]]
```

Out[2]=
$$\mathrm{Min}\left[1, \mathrm{If}\left[\frac{-p-q-\sqrt{4\,p-3\,p^2-2\,p\,q+q^2}}{2\,p} < 0, 10000, \mathrm{sol}[[2]]\right], \mathrm{If}\left[\frac{-p-q+\sqrt{4\,p-3\,p^2-2\,p\,q+q^2}}{2\,p} < 0, 10000, \mathrm{sol}[[3]]\right]\right]$$

The sink-element approach

For the sink-element approach, to compute p(t1 | w), p(t2 | w) and p(t3 | w), we need to integrate the probability of each over the probability simplex:

In[24]:=
```
sinkElementT1 = Integrate[measureT1, {p, 0, 1}, {q, 0, 1 - p}]
```

Out[24]=
$$\frac{1}{120}$$

In[16]:=
```
sinkElementT2 = Integrate[measureT2, {p, 0, 1}, {q , 0, 1 - p}]
```

Out[16]=
$$\frac{1}{420}$$

In[19]:=
```
sinkElementT3 = Integrate[measureT3, {p, 0, 1}, {q, 0, 1 - p}]
```

Out[19]=
$$\frac{1}{420}$$

In[21]:=
```
sinkElementProbT1 = sinkElementT1 / (sinkElementT1 + sinkElementT2 + sinkElementT3)
```

Out[21]=
$$\frac{7}{11}$$

In[22]:=
```
N[sinkElementProbT1]
```

Out[22]= 0.636364

The tight-only approach

For the tight-only approach, we need to create an indicator function that is 1 only if the grammar is tight.

In[26]:=
```
indTight = If [Z < 1, 0 , 1]
```

Out[26]=
$$\text{If}\left[\text{Min}\left[1, \text{If}\left[\frac{-p-q-\sqrt{4\,p-3\,p^2-2\,p\,q+q^2}}{2\,p} < 0, 10\,000, \text{sol}[[2]]\right], \text{If}\left[\frac{-p-q+\sqrt{4\,p-3\,p^2-2\,p\,q+q^2}}{2\,p} < 0, 10\,000, \text{sol}[[3]]\right]\right] < 1, 0, 1\right]$$

(By the way, it can be analytically shown that this grammar is tight when (p+1)/2 < q < 1-p.)

Repeat the same integration, only factoring in the tightness:

In[27]:=
```
tightOnlyT1 = Integrate[measureT1 indTight, {p, 0, 1}, {q, 0, 1 - p}]
```

Out[27]=
$$\frac{1597}{466\,560}$$

In[28]:=
```
tightOnlyT2 = Integrate[measureT2 indTight, {p, 0, 1}, {q, 0, 1 - p}]
```

Out[28]=
$$\frac{1007}{1\,088\,640}$$

In[29]:=
```
tightOnlyT3 = Integrate[measureT2 indTight, {p, 0, 1}, {q, 0, 1 - p}]
```

Out[29]=
$$\frac{1007}{1\,088\,640}$$

In[31]:=
```
tightOnlyT1Prob = tightOnlyT1 / (tightOnlyT1 + tightOnlyT2 + tightOnlyT3)
```

Out[31]=
$$\frac{11\,179}{17\,221}$$

```
N[tightOnlyT1Prob]
```

Out[32]= 0.649149

The renormalization approach

The last thing we need to compute is the probability of t1 with the renormalization approach.
For that, we repeat the integration of the measure of t1, t2 and t3, only this time factoring in the partition function:

In[33]:=

```
renormalizationT1 = Integrate[measureT1 / Z, {p, 0, 1}, {q, 0, 1 - p}]
```

Out[33]=

$$\frac{8109 + 160\sqrt{3}\,\pi + 1280\,\mathrm{ArcCoth}[3] - 640\,\mathrm{Log}[2]}{699\,840}$$

In[34]:=

```
renormalizationT2 = Integrate[measureT2 / Z, {p, 0, 1}, {q, 0, 1 - p}]
```

Out[34]=

$$\frac{13\,509 + 3984\sqrt{3}\,\pi + 73\,792\,\mathrm{ArcCoth}[3] - 13\,568\,\mathrm{Log}[2] - 11\,664\,\mathrm{Log}[3]}{9\,797\,760}$$

In[35]:=

```
renormalizationT3 = Integrate[measureT3 / Z, {p, 0, 1}, {q, 0, 1 - p}]
```

Out[35]=

$$\frac{13\,509 + 3984\sqrt{3}\,\pi + 73\,792\,\mathrm{ArcCoth}[3] - 13\,568\,\mathrm{Log}[2] - 11\,664\,\mathrm{Log}[3]}{9\,797\,760}$$

In[38]:=

```
renormalizationT1Prob =
 renormalizationT1 / (renormalizationT1 + renormalizationT2 + renormalizationT3)
```

Out[38]=

$$\frac{8109 + 160\sqrt{3}\,\pi + 1280\,\mathrm{ArcCoth}[3] - 640\,\mathrm{Log}[2]}{699\,840\left(\frac{8109+160\sqrt{3}\,\pi+1280\,\mathrm{ArcCoth}[3]-640\,\mathrm{Log}[2]}{699\,840} + \frac{13\,509+3984\sqrt{3}\,\pi+73\,792\,\mathrm{ArcCoth}[3]-13\,568\,\mathrm{Log}[2]-11\,664\,\mathrm{Log}[3]}{4\,898\,880}\right)}$$

In[39]:=

```
N[renormalizationT1Prob]
```

Out[39]= 0.619893