Dear Mark,
Sorry for the late reply. Since the ASR system is slow when decoding with the weak LMs (and without LM), it took me awhile to run the experiment.

Before showing the results, I want to state the parameters I was tuning throughout the experiment. Two parameters of the HVite tool were considered in my experiment:

1. **-s f** grammar scale factor, which is the fudge factor mentioned in the comments
2. **-p f** inter model trans penalty (log), which is the insertion penalty at the phone level (if using phone models)

	Cross-entropy	Perplexity	Word Error Rate
3-gram Language Model	5.87	58	5.1%
2-gram Language Model	6.78	110	7.4%
1-gram Language Model	9.53	742	32.8%
No Language Model	12.28	4987	69.2%

In the table above, the WER values for 1,2, and 3-gram LM are derived by setting the insertion penalty to **-4.0**, and the LM scale to **15.0**. In the *No Language Model* case, I simply set the LM scale to **0.0** (in other words, telling HVite to not using LM at all) and let the insertion penalty being **0.0**.

In the LM scale tuning experiment for the 1-gram LM system, I leave the insertion penalty **-4.0**. The following table shows the result

LM scale	WER		LM scale	WER		LM scale	WER
15	32.82		10	27.35		5	29.82
14	31.05		9	27.14		4	31.57
13	30.08		8	26.92		3	37.28
12	28.75		7	27.46		2	41.94
11	**27.65**		6	28.47		1	49.97

I also did the same experiment for the 0-gram LM system. The insertion penalty is set to **-4.0** too (notice that the *No Language Model* system in the first table use insertion penalty 0.0.)

LM scale	WER		LM scale	WER		LM scale	WER
10	35.23		6	35.42		2	48.09

9	34.63		5	36.20		1	56.43
8	**33.96**		4	38.76		0	68.86
7	34.15		3	42.39			

And if we further change the insertion penalty from -4.0 to **0.0**, the WER of *LM scale=0* system becomes 78.14%.

Two things can be observed here:

1. The insertion penalty (-p) is important -- especially for weak LM. It carries some information about the distribution of the occurrence of words with different length.
2. 0-gram LM (uniform distributed) doesn't mean No Language Model. As one of the reader said, it can be seen as the word-level insertion penalty (comparing to the model-level insertion penalty (-p) throughout my experiment.) In fact, we can see the use of N-gram language model as a way to introduce word-level insertion penalty to the ASR system -- the less frequent word-level transitions are penalized more.