every part can add a limitation in the selection of the whole NE. For example: "希/xi拉/la里/li · 克/ke林/lin顿/dun" is a compound name. "希/xi拉/la里/li" can be transliterate to "Hilary" or "Hilaly" and "克/ke林/lin顿/dun" can be transliterate to "Clinton" or "Klinton". But the combination of "Hilary·Clinton" will be selected for it is the most common combination. So the hit of combination query will be extracted as a feature in classifier.

Hint words around the NE: We can take some hint words around the NE into the query, in order to add some limitations to filter out noisy words. For example: "总统 (president)" can be used as hint word for "克林顿 (Clinton)". To find the hint words, we first search the Chinese name in Chinese web pages. The frequent words can be extracted as hint words and they will be translated to English using a bilingual dictionary. These hint words are combined with the revised candidates to search English web pages. So, the hit of the query will be extracted as feature.

The formula of AdaBoost is as follow.

$$H(x) = sign(\sum_{t=1}^{T} \alpha_t h_t(x)) \quad (7)$$

Where α_t is the weight for the *i*th weak classifier $h_t(x)$. α_t can be calculated based on the precision of its corresponding classifier.

5 Experiments

We carry out experiments to investigate how much the revision process and the re-ranking process can improve the performance compared with the baseline of statistical transliteration model. We will also evaluate to which extents we can solve the two problems mentioned in section 1 with the assistance of Web resources.

5.1 Experimental data

The training corpus for statistical transliteration model comes from the corpus of Chinese <-> English Name Entity Lists v 1.0 (LDC2005T34). It contains 565,935 transliteration pairs. Ruling out those pairs which are not suitable for the research on Chinese-English backward transliteration, such as Chinese-Japanese, we select a training set which contains 14,443 pairs of Chinese-European & American person names. In the training set, 1,344 pairs are selected randomly as the close test data. 1,294 pairs out of training set are selected as the open test data. To set up the word list, a 2GB-sized collection of web pages is used. Since 7.42% of the names in the test data don't appear in the list, we use *Google* to get the web page containing the absent names and add these pages into the collection. The word list contains 672,533 words.

5.2 Revision phase vs. statistical approach

Using the results generated from statistical model as baseline, we evaluate the revision module in recall first. The statistical transliteration model works in the following 4 steps: 1) Chinese name are transformed into pinyin representation and the English names are split into syllables. 2) The *GIZA++*[1] tool is invoked to align pinyin to syllables, and the alignment probabilities $P(py \mid es)$ are obtained. 3) Those frequent sequences of syllables are combined as phrases. For example, "be/r/g"→"berg", "s/ky"→"sky". 4) *Camel*[2] decoder is executed to generate 100-best candidates for every name.

We compare the statistical transliteration results with the revised results in Table 2. From Table 2 we can find that the recall of top-100 after revision is improved by 13.26% in close test set and 17.55% in open test set. It proves that the revision module is effective for correcting the mistakes made in statistical transliteration model.

	Transliteration results		Revised results	
	close	open	close	open
Top1	33.64%	9.41%	27.13%	11.04%
Top5	40.37%	13.38%	42.83%	19.69%
Top10	47.79%	17.56%	56.98%	26.32%
Top20	61.38%	25.44%	71.05%	37.81%
Top50	66.49%	36.19%	82.16%	46.22%
Top100	**72.32%**	**41.73%**	**85.78%**	**59.28%**

Table 2. Statistical model vs. Revision module

To show the effects of the revision on the two above-mentioned problems in which the statistical model does not solve well: the losing of silent syllables and the selection bias problem, we make a statistics of the improvements with a measurement of "correction time".

For a Chinese word whose correct transliteration appears in top-100 candidates only if it has been

[1] http://www.fjoch.com/GIZA++.html
[2] http://www.nlp.org.cn