Named Entity Transliteration with Comparable Corpora

Richard Sproat, Tao Tao, ChengXiang Zhai
University of Illinois at Urbana-Champaign, Urbana, IL, 61801
rws@uiuc.edu, {taotao,czhai}@cs.uiuc.edu

Abstract

In this paper we investigate Chinese-English name transliteration using *comparable corpora*, corpora where texts in the two languages deal in some of the same topics — and therefore share references to named entities — but are not translations of each other. We present two distinct methods for transliteration, one approach using phonetic transliteration, and the second using the temporal distribution of candidate pairs. Each of these approaches works quite well, but by combining the approaches one can achieve even better results. We then propose a novel score propagation method that utilizes the co-occurrence of transliteration pairs within document pairs. This propagation method achieves further improvement over the best results from the previous step.

1 Introduction

As part of a more general project on multilingual named entity identification, we are interested in the problem of name transliteration across languages that use different scripts. One particular issue is the discovery of named entities in "comparable" texts in multiple languages, where by comparable we mean texts that are about the same topic, but are *not* in general translations of each other. For example, if one were to go through an English, Chinese and Arabic newspaper on the same day, it is likely that the more important international events in various topics such as politics, business, science and sports, would each be covered in each of the newspapers. Names of the same persons, locations and so forth — which are often *transliterated* rather than translated — would be found in comparable stories across the three papers.[1] We wish to use this expectation to leverage transliteration, and thus the identification of named entities across languages. Our idea is that the occurrence of a cluster of names in, say, an English text, should be useful if we find a cluster of what looks like the same names in a Chinese or Arabic text.

An example of what we are referring to can be found in Figure 1. These are fragments of two stories from the June 8, 2001 Xinhua English and Chinese newswires, each covering an international women's badminton championship. Though these two stories are from the same newswire source, and cover the same event, they are *not* translations of each other. Still, not surprisingly, a lot of the names that occur in one, also occur in the other. Thus *(Camilla) Martin* shows up in the Chinese version as 马尔廷 *ma-er-ting*; *Judith Meulendijks* is 于·默伦迪克斯 *yu mo-lun-di-ke-si*; and *Mette Sorensen* is 迈·索伦森*mai su-lun-sen*. Several other correspondences also occur. While some of the transliterations are "standard" — thus Martin is conventionally transliterated as 马尔廷 *ma-er-ting* — many of them were clearly more novel, though all of them follow the standard Chinese conventions for transliterating foreign names.

These sample documents illustrate an important point: if a document in language L_1 has a set of names, and one finds a document in L_2 containing a set of names that look as if they could be transliterations of the names in the L_1 document, then this should boost one's confidence that the two sets of names are indeed transliterations of each other. We will demonstrate that this intuition is correct.

[1] Many names, particularly of organizations, may be translated rather than transliterated; the transliteration method we discuss here obviously will not account for such cases, though the time correlation and propagation methods we discuss will still be useful.

Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pages 73–80,
Sydney, July 2006. ©2006 Association for Computational Linguistics

Dai Yun Nips World No. 1 Martin to Shake off Olympic Shadow . . . In the day's other matches, second seed Zhou Mi overwhelmed Ling Wan Ting of Hong Kong, China 11-4, 11-4, Zhang Ning defeat Judith Meulendijks of Netherlands 11-2, 11-9 and third seed Gong Ruina took 21 minutes to eliminate Tine Rasmussen of Denmark 11-1, 11-1, enabling China to claim five quarterfinal places in the women's singles.

羽毛球世锦赛中国女单选手全部跻身八强 . . . 马尔廷还认为,她不可能连续战胜4个中国人,即使 . . . 三号种子龚睿那今晚以两个11:1轻取丹麦选手蒂·拉斯姆森,张宁在上午以11:2和11:9淘汰了荷兰的于·默伦迪克斯,周蜜在下午以11:4和11:1战胜了中国香港选手凌婉婷

Figure 1: Sample from two stories about an international women's badminton championship.

2 Previous Work

In previous work on Chinese named-entity transliteration — e.g. (Meng et al., 2001; Gao et al., 2004), the problem has been cast as the problem of producing, for a given Chinese name, an English equivalent such as one might need in a machine translation system. For example, for the name 维·威廉姆斯 *wei wei-lian-mu-si*, one would like to arrive at the English name *V(enus) Williams*. Common approaches include source-channel methods, following (Knight and Graehl, 1998) or maximum-entropy models.

Comparable corpora have been studied extensively in the literature (e.g.,(Fung, 1995; Rapp, 1995; Tanaka and Iwasaki, 1996; Franz et al., 1998; Ballesteros and Croft, 1998; Masuichi et al., 2000; Sadat et al., 2003)), but transliteration in the context of comparable corpora has not been well addressed.

The general idea of exploiting frequency correlations to acquire word translations from comparable corpora has been explored in several previous studies (e.g., (Fung, 1995; Rapp, 1995; Tanaka and Iwasaki, 1996)).Recently, a method based on Pearson correlation was proposed to mine word pairs from comparable corpora (Tao and Zhai, 2005), an idea similar to the method used in (Kay and Roscheisen, 1993) for sentence alignment. In our work, we adopt the method proposed in (Tao and Zhai, 2005) and apply it to the problem of transliteration. We also study several variations of the similarity measures.

Mining transliterations from multilingual web pages was studied in (Zhang and Vines, 2004); Our work differs from this work in that we use comparable corpora (in particular, news data) and leverage the time correlation information naturally available in comparable corpora.

3 Chinese Transliteration with Comparable Corpora

We assume that we have comparable corpora, consisting of newspaper articles in English and Chinese from the same day, or almost the same day. In our experiments we use data from the English and Chinese stories from the Xinhua News agency for about 6 months of 2001.[2] We assume that we have identified names for persons and locations—two types that have a strong tendency to be transliterated wholly or mostly phonetically—in the English text; in this work we use the named-entity recognizer described in (Li et al., 2004), which is based on the SNoW machine learning toolkit (Carlson et al., 1999).

To perform the transliteration task, we propose the following general three-step approach:

1. Given an English name, identify candidate Chinese character n-grams as possible transliterations.

2. Score each candidate based on how likely the candidate is to be a transliteration of the English name. We propose two different scoring methods. The first involves phonetic scoring, and the second uses the frequency profile of the candidate pair over time. We will show that each of these approaches works quite well, but by combining the approaches one can achieve even better results.

3. Propagate scores of all the candidate transliteration pairs globally based on their co-occurrences in document pairs in the comparable corpora.

The intuition behind the third step is the following. Suppose several high-confidence name transliteration pairs occur in a pair of English and Chinese documents. Intuitively, this would increase our confidence in the other plausible transliteration pairs in the same document pair. We thus propose a score propagation method to allow these high-confidence pairs to propagate some of their

[2] Available from the LDC via the English Gigaword (LDC2003T05) and Chinese Gigaword (LDC2003T09) corpora.

scores to other co-occurring transliteration pairs. As we will show later, such a propagation strategy can generally further improve the transliteration accuracy; in particular, it can further improve the already high performance from combining the two scoring methods.

3.1 Candidate Selection

The English named entity candidate selection process was already described above. Candidate Chinese transliterations are generated by consulting a list of characters that are frequently used for transliterating foreign names. As discussed elsewhere (Sproat et al., 1996), a subset of a few hundred characters (out of several thousand) tends to be used overwhelmingly for transliterating foreign names into Chinese. We use a list of 495 such characters, derived from various online dictionaries. A sequence of three or more characters from the list is taken as a possible name. If the character "·" occurs, which is frequently used to represent the space between parts of an English name, then at least one character to the left and right of this character will be collected, even if the character in question is not in the list of "foreign" characters.

Armed with the English and Chinese candidate lists, we then consider the pairing of every English candidate with every Chinese candidate. Obviously it would be impractical to do this for all of the candidates generated for, say, an entire year: we consider as plausible pairings those candidates that occur within a day of each other in the two corpora.

3.2 Candidate scoring based on pronunciation

We adopt a source-channel model for scoring English-Chinese transliteration pairs. In general, we seek to estimate $P(e|c)$, where e is a word in Roman script, and c is a word in Chinese script. Since Chinese transliteration is mostly based on pronunciation, we estimate $P(e'|c')$, where e' is the pronunciation of e and c' is the pronunciation of c. Again following standard practice, we decompose the estimate of $P(e'|c')$ as $P(e'|c') = \prod_i P(e'_i|c'_i)$. Here, e'_i is the ith subsequence of the English phone string, and c'_i is the ith subsequence of the Chinese phone string. Since Chinese transliteration attempts to match the syllable-sized characters to equivalent sounding spans of the English language, we fix the c'_i to be syllables, and let the e'_i range over all possible subsequences of the English phone string. For training data we have a small list of 721 names in Roman script and their Chinese equivalent.[3] Pronunciations for English words are obtained using the Festival text-to-speech system (Taylor et al., 1998); for Chinese, we use the standard pinyin transliteration of the characters. English-Chinese pairs in our training dictionary were aligned using the alignment algorithm from (Kruskal, 1999), and a hand-derived set of 21 rules-of-thumb: for example, we have rules that encode the fact that Chinese /l/ can correspond to English /r/, /n/ or /er/; and that Chinese /w/ may be used to represent /v/. Given that there are over 400 syllables in Mandarin (not counting tone) and each of these syllables can match a large number of potential English phone spans, this is clearly not enough training data to cover all the parameters, and so we use Good-Turing estimation to estimate probabilities for unseen correspondences. Since we would like to filter implausible transliteration pairs we are less lenient than standard estimation techniques in that we are willing to assign zero probability to some correspondences. Thus we set a hard rule that for an English phone span to correspond to a Chinese syllable, the initial phone of the English span must have been seen in the training data as corresponding to the initial of the Chinese syllable some minimum number of times. For consonant-initial syllables we set the minimum to 4. We omit further details of our estimation technique for lack of space. This phonetic correspondence model can then be used to score putative transliteration pairs.

3.3 Candidate Scoring based on Frequency Correlation

Names of the same entity that occur in different languages often have correlated frequency patterns due to common triggers such as a major event. Thus if we have comparable news articles over a sufficiently long time period, it is possible to exploit such correlations to learn the associations of names in different languages. The idea of exploiting frequency correlation has been well studied. (See the previous work section.) We adopt the method proposed in (Tao and Zhai, 2005), which

[3] The LDC provides a much larger list of transliterated Chinese-English names, but we did not use this here for two reasons. First, we have found it it be quite noisy. Secondly, we were interested in seeing how well one could do with a limited resource of just a few hundred names, which is a more realistic scenario for languages that have fewer resources than English and Chinese.

works as follows: We pool all documents in a single day to form a large pseudo-document. Then, for each transliteration candidate (both Chinese and English), we compute its frequency in each of those pseudo-documents and obtain a raw frequency vector. We further normalize the raw frequency vector so that it becomes a frequency distribution over all the time points (days). In order to compute the similarity between two distribution vectors, The Pearson correlation coefficient was used in (Tao and Zhai, 2005); here we also considered two other commonly used measures – **cosine** (Salton and McGill, 1983), and **Jensen-Shannon divergence** (Lin, 1991), though our results show that Pearson correlation coefficient performs better than these two other methods.

3.4 Score Propagation

In both scoring methods described above, scoring of each candidate transliteration pair is *independent* of the other. As we have noted, document pairs that contain lots of plausible transliteration pairs should be viewed as more plausible document pairs; at the same time, in such a situation we should also trust the putative transliteration pairs more. Thus these document pairs and transliteration pairs mutually "reinforce" each other, and this can be exploited to further optimize our transliteration scores by allowing transliteration pairs to propagate their scores to each other according to their co-occurrence strengths.

Formally, suppose the current generation of transliteration scores are (e_i, c_i, w_i) $i = 1, ..., n$, where (e_i, c_i) is a distinct pair of English and Chinese names. Note that although for any $i \neq j$, we have $(e_i, c_i) \neq (e_j, c_j)$, it is possible that $e_i = e_j$ or $c_i = c_j$ for some $i \neq j$. w_i is the transliteration score of (e_i, c_i).

These pairs along with their co-occurrence relation computed based on our comparable corpora can be formally represented by a graph as shown in Figure 2. In such a graph, a node represents (e_i, c_i, w_i). An edge between (e_i, c_i, w_i) and (e_j, c_j, w_j) is constructed iff (e_i, c_i) and (e_j, c_j) co-occur in a certain document pair (E_t, C_t), i.e. there exists a document pair (E_t, C_t), such that $e_i, e_j \in E_t$ and $c_i, c_j \in C_t$. Given a node (e_i, c_i, w_i), we refer to all its directly-connected nodes as its "neighbors". The documents do not appear explicitly in the graph, but they implicitly affect the graph's topology and the weight of each edge. Our idea of score propagation can now be formulated as the following recursive equation for updating the scores of all the transliteration pairs.

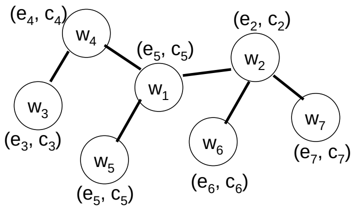

Figure 2: Graph representing transliteration pairs and cooccurence relations.

$$w_i^{(k)} = \alpha \times w_i^{(k-1)} + (1-\alpha) \times \sum_{j \neq i, j=1}^{n} (w_j^{(k-1)} \times P(j|i)),$$

where $w_i^{(k)}$ is the new score of the pair (e_i, c_i) after an iteration, while $w_i^{(k-1)}$ is its old score before updating; $\alpha \in [0, 1]$ is a parameter to control the overall amount of propagation (when $\alpha = 1$, no propagation occurs); $P(j|i)$ is the conditional probability of propagating a score from node (e_j, c_j, w_j) to node (e_i, c_i, w_i).

We estimate $P(j|i)$ in two different ways: 1) The number of cooccurrences in the whole collection (Denote as CO). $P(j|i) = \frac{C(i,j)}{\sum_{j'} C(i,j')}$, where $C(i, j)$ is the cooccurrence count of (e_i, c_i) and (e_j, c_j); 2) A mutual information-based method (Denote as MI). $P(j|i) = \frac{MI(i,j)}{\sum_{j'} MI(i,j')}$, where $MI(i, j)$ is the mutual information of (e_i, c_i) and (e_j, c_j). As we will show, the CO method works better. Note that the transition probabilities between indirect neighbors are always 0. Thus propagation only happens between direct neighbors.

This formulation is very similar to PageRank, a link-based ranking algorithm for Web retrieval (Brin and Page, 1998). However, our motivation is propagating scores to exploit cooccurrences, so we do not necessarily want the equation to converge. Indeed, our results show that although the initial iterations always help improve accuracy, too many iterations actually would decrease the performance.

4 Evaluation

We use a comparable English-Chinese corpus to evaluate our methods for Chinese transliteration. We take one day's worth of comparable news articles (234 Chinese stories and 322 English stories), generate about 600 English names with the entity recognizer (Li et al., 2004) as described above, and find potential Chinese transliterations also as previously described. We generated 627 Chinese candidates. In principle, all these 600×627 pairs are potential transliterations. We then apply the phonetic and time correlation methods to score and rank all the candidate Chinese-English correspondences.

To evaluate the proposed transliteration methods quantitatively, we measure the accuracy of the ranked list by Mean Reciprocal Rank (MRR), a measure commonly used in information retrieval when there is precisely one correct answer (Kantor and Voorhees, 2000). The reciprocal rank is the reciprocal of the rank of the correct answer. For example, if the correct answer is ranked as the first, the reciprocal rank would be 1.0, whereas if it is ranked the second, it would be 0.5, and so forth. To evaluate the results for a set of English names, we take the mean of the reciprocal rank of each English name.

We attempted to create a complete set of answers for all the English names in our test set, but a small number of English names do not seem to have any standard transliteration according to the resources that we consulted. We ended up with a list of about 490 out of the 600 English names judged. We further notice that some answers (about 20%) are not in our Chinese candidate set. This could be due to two reasons: (1) The answer does not occur in the Chinese news articles we look at. (2) The answer is there, but our candidate generation method has missed it. In order to see more clearly how accurate each method is for ranking the candidates, we also compute the MRR for the subset of English names whose transliteration answers are in our candidate list. We distinguish the MRRs computed on these two sets of English names as “AllMRR” and “CoreMRR”.

Below we first discuss the results of each of the two methods. We then compare the two methods and discuss results from combining the two methods.

4.1 Phonetic Correspondence

We show sample results for the phonetic scoring method in Table 1. This table shows the 10 highest scoring transliterations for each Chinese character sequence based on all texts in the Chinese and English Xinhua newswire for the 13th of August, 2001. 8 out of these 10 are correct. For all the English names the MRR is 0.3, and for the core names it is 0.89. Thus on average, the correct answer, if it is included in our candidate list, is ranked mostly as the first one.

∗paris	佩雷斯	pei-lei-si	3.51
iraq	伊拉克	yi-la-ke	3.74
staub	斯塔伯	si-ta-bo	4.45
canada	加拿大	jia-na-da	4.85
belfast	贝尔法斯特	bei-er-fa-si-te	4.90
fischer	菲舍尔	fei-she-er	4.91
philippine	菲律宾	fei-lü-bin	4.97
lesotho	莱索托	lai-suo-two	5.12
∗tirana	铁路内	tye-lu-na	5.15
freeman	弗里曼	fu-li-man	5.26

Table 1: Ten highest-scoring matches for the Xinhua corpus for 8/13/01. The final column is the $-log\ P$ estimate for the transliteration. Starred entries are incorrect.

4.2 Frequency correlation

Similarity	AllMRR	CoreMRR
Pearson	0.1360	0.3643
Cosine	0.1141	0.3015
JS-div	0.0785	0.2016

Table 2: MRRs of the frequency correlation methods.

We proposed three similarity measures for the frequency correlation method, i.e., the Cosine, Pearson coefficient, and Jensen-Shannon divergence. In Table 2, we show their MRRs. Given that the only resource the method needs is comparable text documents over a sufficiently long period, these results are quite encouraging. For example, with Pearson correlation, when the Chinese transliteration of an English name is included in our candidate list, the correct answer is, on average, ranked at the 3rd place or better. The results thus show that the idea of exploiting frequency correlation does work. We also see that among the three similarity measures, Pearson correlation performs the best; it performs better than Cosine, which is better than JS-divergence.

Compared with the phonetic correspondence method, the performance of the frequency correlation method is in general much worse, which is not surprising, given the fact that terms may be correlated merely because they are topically related.

4.3 Combination of phonetic correspondence and frequency correlation

Method	AllMRR	CoreMRR
Phonetic	0.2999	0.8895
Freq	0.1360	0.3643
Freq+PhoneticFilter	0.3062	0.9083
Freq+PhoneticScore	0.3194	0.9474

Table 3: Effectiveness of combining the two scoring methods.

Since the two methods exploit complementary resources, it is natural to see if we can improve performance by combining the two methods. Indeed, intuitively the best candidate is the one that has a good pronunciation alignment as well as a correlated frequency distribution with the English name. We evaluated two strategies for combining the two methods. The first strategy is to use the phonetic model to filter out (clearly impossible) candidates and then use the frequency correlation method to rank the candidates. The second is to combine the scores of these two methods. Since the correlation coefficient has a maximum value of 1, we normalize the phonetic correspondence score by dividing all scores by the maximum score so that the maximum normalized value is also 1. We then take the average of the two scores and rank the candidates based on their average scores. Note that the second strategy implies the application of the first strategy.

The results of these two combination strategies are shown in Table 3 along with the results of the two individual methods. We see that both combination strategies are effective and the MRRs of the combined results are all better than those of the two individual methods. It is interesting to see that the benefit of applying the phonetic correspondence model as a filter is quite significant. Indeed, although the performance of the frequency correlation method alone is much worse than that of the phonetic correspondence method, when working on the subset of candidates passing the phonetic filter (i.e., those candidates that have a reasonable phonetic alignment with the English name), it can outperform the phonetic correspondence method. This once again indicates that exploiting the frequency correlation can be effective. When combining the scores of these two methods, we not only (implicitly) apply the phonetic filter, but also exploit the discriminative power provided by the phonetic correspondence scores and this is shown to bring in additional benefit, giving the best performance among all the methods.

4.4 Error Analysis

From the results above, we see that the MRRs for the core English names are substantially higher than those for all the English names. This means that our methods perform very well whenever we have the answer in our candidate list, but we have also missed the answers for many English names. The missing of an answer in the candidate list is thus a major source of errors. To further understand the upper bound of our method, we manually add the missing correct answers to our candidate set and apply all the methods to rank this augmented set of candidates. The performance is reported in Table 4 with the corresponding performance on the original candidate set. We see that, as expected, the performance on the augmented candidate list, which can be interpreted as an upper bound of our method, is indeed much better, suggesting that if we can somehow improve the candidate generation method to include the answers in the list, we can expect to significantly improve the performance for all the methods. This is clearly an interesting topic for further research. The relative performance of different methods on this augmented candidate list is roughly the same as on the original candidate list, except that the "Freq+PhoneticFilter" is slightly worse than that of the phonetic method alone, though it is still much better than the performance of the frequency correlation alone. One possible explanation may be that since these names do not necessarily occur in our comparable corpora, we may not have sufficient frequency observations for some of the names.

Method	ALLMRR	
	Original	Augmented
Phonetic	0.2999	0.7157
Freq	0.1360	0.3455
Freq+PhoneticFilter	0.3062	0.6232
Freq+PhoneticScore	0.3194	0.7338

Table 4: MRRs on the augmented candidate list.

Method	AllMRR			CoreMRR		
	init.	CO	MI	init.	CO	MI
Freq+PhoneticFilter	0.3171	0.3255	0.3255	0.9058	0.9372	0.9372
Freq+PhoneticScore	0.3290	0.3373	0.3392	0.9422	0.9659	0.9573

Table 5: Effectiveness of score propagation.

4.5 Experiments on score propagation

To demonstrate that score propagation can further help transliteration, we use the combination scores in Table 3 as the initial scores, and apply our propagation algorithm to iteratively update them. We remove the entries when they do not co-occur with others. There are 25 such English name candidates. Thus, the initial scores are actually slightly different from the values in Table 3. We show the new scores and the best propagation scores in Table 5. In the table, "init." refers to the initial scores. and "CO" and "MI" stand for best scores obtained using either the co-occurrence or mutual information method. While both methods result in gains, CO very slightly outperforms the MI approach. In the score propagation process, we introduce two additional parameters: the interpolation parameter α and the number of iterations k. Figure 3 and Figure 4 show the effects of these parameters. Intuitively, we want to preserve the initial score of a pair, but add a slight boost from its neighbors. Thus, we set α very close to 1 (0.9 and 0.95), and allow the system to perform 20 iterations. In both figures, the first few iterations certainly leverage the transliteration, demonstrating that the propagation method works. However, we observe that the performance drops when more iterations are used, presumably due to noise introduced from more distantly connected nodes. Thus, a relatively conservative approach is to choose a high α value, and run only a few iterations. Note, finally, that the CO method seems to be more stable than the MI method.

5 Conclusions and Future Work

In this paper we have discussed the problem of Chinese-English name transliteration as one component of a system to find matching names in comparable corpora. We have proposed two methods for transliteration, one that is more traditional and based on phonetic correspondences, and one that is based on word distributions and adopts methods from information retrieval. We have shown that both methods yield good results, and that even better results can be achieved by combining the methods. We have further showed that one can improve upon the combined model by using reinforcement via score propagation when transliteration pairs cluster together in document pairs.

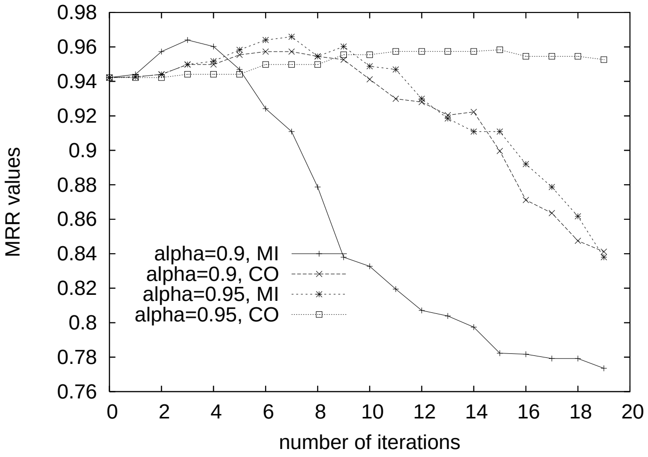

Figure 3: Propagation: Core items

The work we report is ongoing. We are investigating transliterations among several language pairs, and are extending these methods to Korean, Arabic, Russian and Hindi — see (Tao et al., 2006).

6 Acknowledgments

This work was funded by Dept. of the Interior contract NBCHC040176 (REFLEX). We also thank three anonymous reviewers for ACL06.

References

Lisa Ballesteros and W. Bruce Croft. 1998. Resolving ambiguity for cross-language retrieval. In *Research and Development in Information Retrieval*, pages 64–71.

Sergey Brin and Lawrence Page. 1998. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30:107–117.

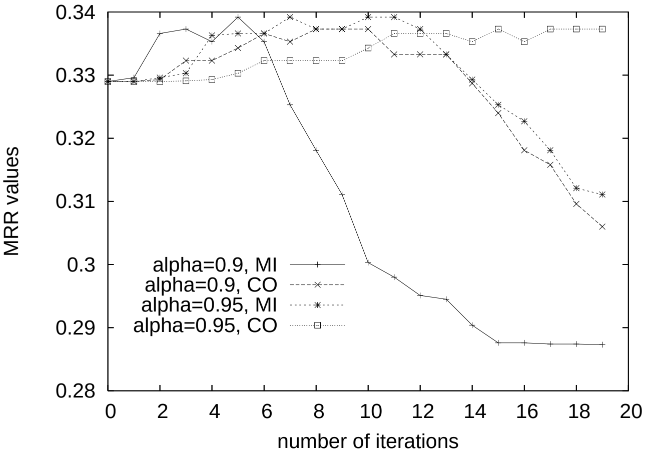

Figure 4: Propagation: All items

A. Carlson, C. Cumby, J. Rosen, and D. Roth. 1999. The SNoW learning architecture. Technical Report UIUCDCS-R-99-2101, UIUC CS Dept.

Martin Franz, J. Scott McCarley, and Salim Roukos. 1998. Ad hoc and multilingual information retrieval at IBM. In *Text REtrieval Conference*, pages 104–115.

Pascale Fung. 1995. A pattern matching method for finding noun and proper noun translations from noisy parallel corpora. In *Proceedings of ACL 1995*, pages 236–243.

W. Gao, K.-F. Wong, and W. Lam. 2004. Phoneme-based transliteration of foreign names for OOV problem. In *IJCNLP*, pages 374–381, Sanya, Hainan.

P. Kantor and E. Voorhees. 2000. The TREC-5 confusion track: Comparing retrieval methods for scanned text. *Information Retrieval*, 2:165–176.

M. Kay and M. Roscheisen. 1993. Text translation alignment. *Computational Linguistics*, 19(1):75–102.

K. Knight and J. Graehl. 1998. Machine transliteration. *CL*, 24(4).

J. Kruskal. 1999. An overview of sequence comparison. In D. Sankoff and J. Kruskal, editors, *Time Warps, String Edits, and Macromolecules*, chapter 1, pages 1–44. CSLI, 2nd edition.

X. Li, P. Morie, and D. Roth. 2004. Robust reading: Identification and tracing of ambiguous names. In *NAACL-2004*.

J. Lin. 1991. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151.

H. Masuichi, R. Flournoy, S. Kaufmann, and S. Peters. 2000. A bootstrapping method for extracting bilingual text pairs.

H.M. Meng, W.K Lo, B. Chen, and K. Tang. 2001. Generating phonetic cognates to handle named entities in English-Chinese cross-languge spoken document retrieval. In *Proceedings of the Automatic Speech Recognition and Understanding Workshop*.

R. Rapp. 1995. Identifying word translations in non-parallel texts. In *Proceedings of ACL 1995*, pages 320–322.

Fatiha Sadat, Masatoshi Yoshikawa, and Shunsuke Uemura. 2003. Bilingual terminology acquisition from comparable corpora and phrasal translation to cross-language information retrieval. In *ACL '03*, pages 141–144.

G. Salton and M. McGill. 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill.

R. Sproat, C. Shih, W. Gale, and N. Chang. 1996. A stochastic finite-state word-segmentation algorithm for Chinese. *CL*, 22(3).

K. Tanaka and H. Iwasaki. 1996. Extraction of lexical translation from non-aligned corpora. In *Proceedings of COLING 1996*.

Tao Tao and ChengXiang Zhai. 2005. Mining comparable bilingual text corpora for cross-language information integration. In *KDD'05*, pages 691–696.

Tao Tao, Su-Youn Yoon, Andrew Fister, Richard Sproat, and ChengXiang Zhai. 2006. Unsupervised named entity transliteration using temporal and phonetic correlation. In *EMNLP 2006*, Sydney, July.

P. Taylor, A. Black, and R. Caley. 1998. The architecture of the Festival speech synthesis system. In *Proceedings of the Third ESCA Workshop on Speech Synthesis*, pages 147–151, Jenolan Caves, Australia.

Ying Zhang and Phil Vines. 2004. Using the web for automated translation extraction in cross-language information retrieval. In *SIGIR '04*, pages 162–169.