

# PHONETIC NAME MATCHING FOR CROSS-LINGUAL SPOKEN SENTENCE RETRIEVAL

Heng Ji\*, Ralph Grishman\*, Wen Wang<sup>+</sup>

\*Department of Computer Science, New York University    <sup>+</sup>SRI International

## ABSTRACT

Cross-lingual Spoken Sentence Retrieval (CLSSR) remains a challenge, especially for queries including OOV words such as person names. This paper proposes a simple method of fuzzy matching between query names and phones of candidate audio segments. This approach has the advantage of avoiding some word decoding errors in Automatic Speech Recognition (ASR). Experiments on Mandarin-English CLSSR show that phone-based searching and conventional translation-based searching are complementary. Adding phone matching achieved 26.29% improvement on F-measure over searching on state-of-the-art Machine Translation (MT) output and 8.83% over Entity Translation (ET) output.

*Index Terms*— Speech Recognition, Information Retrieval

## 1. INTRODUCTION

While there is growing interest in the ability to retrieve passages from foreign language audio, the performance of such systems is still quite limited by Automatic Speech Recognition (ASR) and Machine Translation (MT) errors, and in particular errors on OOV words. Given the fact that a large percentage of OOV words are person names [1], in this paper we shall study the task of retrieving sentences from Mandarin speech for person name queries in English.

Error analysis has shown that a major percentage of Mandarin ASR errors occur in word decoding, and these errors inhibit retrieval in conventional CLIR systems. Such errors can be circumvented by matching the query against the phone sequence. A number of recent studies have stressed the benefits of taking advantage of the audio source itself in this way [2, 3, 4].

Table 1 shows the wide range of cases that must be addressed in matching queries against phone sequences generated from ASR. In this paper we show how a mix of strategies can effectively address these cases. We investigate the limits of directly matching the query against phone sequences. We shall see how to improve this matching through query transliteration and translation; and use Japanese names as an example to show the effectiveness of country origin-specific query translation. We shall show that the two searching methods, phone-based and translation-based, complement each other. We shall also propose a confidence based combination approach and demonstrate its effectiveness.

Mandarin Phone Sequence for Doc.	English Query	Origin
hu jin taw	Hu Jintao	Mainland, China
ta la ba ni	Talabani	Iraq
pu jing	Putin	Russia
bu shi	Bush	US
sha bi er	Shabir	Afghanistan
bu le er	Blair	UK
ying show zhong tyan	Hidetoshi Nakata	Japan
lu wu xvan	Roh Moo-Hyun	Korea
rwan ming zhe	Nguyen Minh Triet	Vietnam
chen feng fu zhen	Margaret Chan	Hong Kong, China

Table 1. Person Name Examples

## 2. TASK AND DATA

The task we are addressing is Cross-lingual Sentence Retrieval: given a person name query in English the system should retrieve the sentences containing this name from Mandarin speech.

We used part of the GALE<sup>1</sup> Y2 audio MT development corpus as our candidate documents (19 broadcast conversation shows and 26 broadcast news shows, in total 668 sentences). 53 queries were constructed by selecting person names from the reference translations of reference transcripts for these shows. Then for each query, relevant sentences from the entire corpus were manually labeled as answer keys. The decisions were made against reference translations. We excluded ambiguous queries (e.g. “Li” can refer to “Li Huifen” in one show but to “Li Zhaoxing” in another)<sup>2</sup>. In total 164 sentences were labeled as relevant.

## 3. MOTIVATION

### 3.1. Document Processing Systems

We first consider the limitations of a more conventional retrieval approach, based on transcribing and translating the audio content. We used two baseline document processing systems in this paper. Mandarin shows are first recognized as Chinese texts by a state-of-the-art ASR system<sup>3</sup> [5], and then translated into English by two different translation models, one a statistical phrase-based MT

<sup>1</sup> Global Autonomous Language Exploitation

<sup>2</sup> Cross-document entity disambiguation is not in the research scope of this paper.

<sup>3</sup> For evaluation purposes ASR system retained reference sentence segmentations.

system [6] and the other a named entity translation (ET) system [7], which only translates names but in general does so more accurately.

### 3.2. Error Analysis

We begin our error analysis with an investigation of these two baseline pipelines, decomposing the errors into phone decoding errors, word decoding errors and translation errors. In Table 2 we report the detailed error distribution for processing one instance each for these 53 query names.

Table 2 shows that most Mandarin ASR errors occurred in word decoding. So if we search queries on the phone level we can achieve a higher potential recall. In fact even for partially correct phones (e.g. the “Blair” example), we could still obtain a successful match by fuzzy phone matching.

We can also see that translation errors are still dominant, and further analysis showed that about 1/4 of the errors were due to confusions between names and common words. For example, “Zhao Tiechui” was correctly recognized by ASR but “Tiechui” means “hammer”, and so MT mistakenly translated it into “Zhao hammer”. Developing a separate ET (name translation) system can certainly help, but for some difficult cases such as “Shi Guoqi (means ‘time national-flag’)” the ET system tends to miss the entire name during name tagging. About half of the translation errors were for names which never appeared in bitexts, in which case a name-unaware MT system tends to translate each character by meaning (e.g. “Hidetoshi Nakata” to “Chinese Descent”).

Error Type	No.	Examples	
		Reference	System
Phone Decoding Error	4	bu lai er (“Blair”)	bu le er
		wu yameng	yao me (“either”)
Word Decoding Error	8	姜瑜 (Jiang Yu)	将于 (“will be”)
		项飞 (Xiang Fei)	向飞 (“fly over”)
Baseline 1: MT Error	14	Hidetoshi Nakata	Chinese Descent
		Zhao Tiechui	Zhao Hammer
Baseline2: ET Error	6	Shi Guoqi	[Missing]
		Clark	Clack

Table 2. Document Name Processing Error Distribution

Furthermore, in such a sequential pipeline all ASR errors are naturally propagated to MT and ET. For example, “Xiang Fei” was translated into “fly over” because of the wrong recognition for the last name “Xiang”. To summarize, all these errors appeared because we treated translation of spoken documents in the same way as that for newswire texts. Therefore, for the remainder of this paper we shall focus on incorporating phone information to alleviate this shortcoming.

## 4. A MULTI-SOURCE CLSSR FRAMEWORK

Based on the above motivations we propose a multi-source CLSSR framework which tightly integrates phone information and where searching can be done on various levels. We will provide a general system overview in section 4.1, and then focus on query processing (section 4.2) and searching algorithms (4.3).

### 4.1. System Overview

The general system flow is depicted in Figure 1. We apply separate processing to the documents and queries, and then feed different combinations of query and document representation  $\langle Q_i, D_j \rangle$  into a sentence retrieval model.

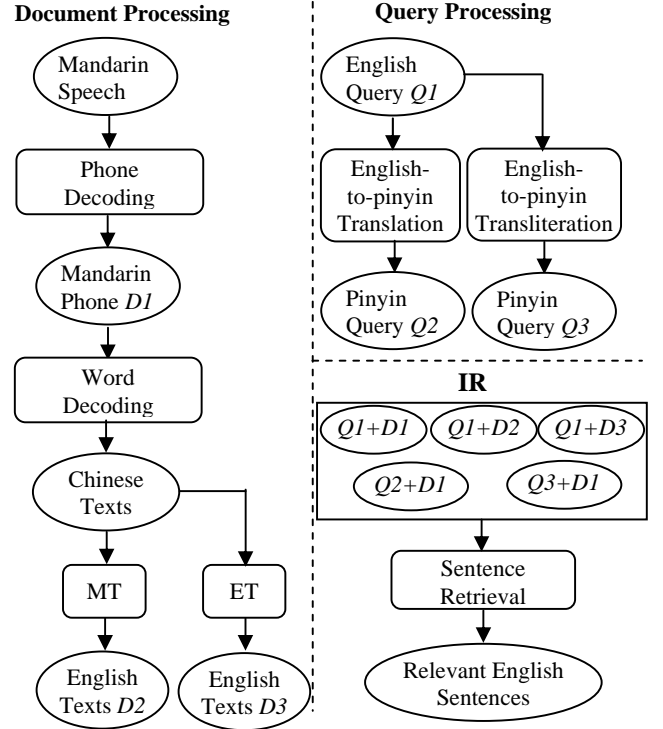


Figure 1. Cross-lingual Spoken Sentence Retrieval Framework

### 4.2. Query Translation and Transliteration

Directly matching English queries with Mandarin phones can help retrieve many Chinese names successfully, because they appear in the same form in query and document. However, for many non-Chinese names (e.g. “Schwarzenegger” has mandarin phone sequence “shi wa xin ge”) this direct matching produces very low confidence values. So we have developed a query processing pipeline to translate/transliterate English queries to Mandarin pinyin sequences to enhance matching confidence. We exploited a variety of resources as follows.

The ET system [7] includes about 80,000 Chinese-English name pairs mined from cross-lingual Wikipedia pages, parallel corpora, web data and LDC bi-lingual name dictionary. By converting Chinese names into Mandarin pinyin sequences we obtained an accurate database for English-to-pinyin translation.

In addition, Japanese names were handled specially. Japanese names in the English query (preserving Japanese pronunciation) and Mandarin phones (preserving Mandarin pronunciation) hardly match at all. Therefore we exploited a Japanese-English person name dictionary [8] (20126 entries) and converted Chinese characters to pinyin. For example, this database can provide 10 possible translations for “Hidetoshi Nakata” (“Yingshou Zhongtian”, “Yingjun Zhongtian”, “Xiushou Zhongtian”, etc.), one

of which matches the query name exactly and all the others match with high confidence.

However such a name pair database cannot cover all names that will possibly appear in queries. So we took a further step by using an English-to-pinyin transliteration model developed by Fair Isaac Corp. (a reverse version of the averaged perceptron based pinyin-to-English transliteration model described in [9]) to transliterate those names not in our database.

### 4.3. Searching Algorithms

The above query and document processing steps result in various possible combinations of  $\langle Qi, Dj \rangle$  as shown in Figure 1. This section shall present a uniform searching algorithm and confidence based combination scheme.

#### 4.3.1. Cognates between Phone and Pinyin

There are some slight differences of representation between Mandarin phone and pinyin, so we automatically learned 37 frequent cognate pairs from a separate development set (35 broadcast conversation shows and 22 broadcast news shows) from GALE Y2 MT dev set. For example, “ey” and “ei”, “van” and “uan” are considered as cognates in matching.

#### 4.3.2. Fuzzy Name Matching and Confidence Estimation

For each combination of query and document representation  $\langle Qi, Dj \rangle$ , we define the match *cost* between the query and a document substring as the Damerau–Levenshtein edit distance [10, 11], with each operation assigned unit cost. We used the spelling for the phone sequence and computed the edit cost in terms of letter sequences, with blanks removed. Based on this cost, we computed a matching confidence value:

$$\text{conf}(query, \text{doc-string}) = \frac{\max(\text{length}(query), \text{length}(\text{doc-string})) - \text{cost}}{\max(\text{length}(query), \text{length}(\text{doc-string}))}$$

We then define the sentence-level match confidence, *sent-conf*, as the maximum value of *conf* between the query and any sentence substring starting and ending on a syllable (for phone sequences) or token boundary.

#### 4.3.3. Confidence Based Sentence Ranking and Filtering

We apply the above fuzzy searching algorithm for all combinations  $\langle Qi, Dj \rangle$  in Figure 1 and rank them by matching confidence. Then in the final output we produce a set of most confident sentences. If the highest confidence score is 1 (exact match), then we only keep the results produced by one combination in the following priority order:  $Q1+D2 > Q1+D3 > Q1+D1 > Q2+D1 > Q3+D1$ . This can help filter some noise produced by using phone matching alone.

## 5. EXPERIMENTAL RESULTS AND ANALYSIS

In this section we present the experimental results of adding phone matching into CLSSR.

### 5.1. Overall Performance

Table 3 and Figure 2 show the overall Precision (P), Recall (R) and F-Measure (F) scores for different searching methods, including two baselines *MTS* and *ETS*, the phone matching based searching *PHSTT*, and the final combined system *COMBINE*. Each

curve in Figure 2 shows the effect on precision and recall of varying the threshold for the fuzzy searching confidence. The labeled point on each curve shows the best F-measure that can be obtained by adjusting the threshold at an optimal value (t) for that searching method. The best performance for each approach is summarized in Table 3.

Approach	t	P(%)	R(%)	F(%)
<i>MTS: (Q1+D2) MT Search</i>	0.80	96.25	46.95	63.11
<i>ETS: (Q1+D3) ET Search</i>	0.80	95.80	69.51	80.57
<i>PHS: (Q1+D1) Phone Search</i>	0.75	66.43	57.93	61.89
<i>PHST1: PHS + (Q2+D1) Query Translation</i>	0.77	75.95	73.17	74.53
<i>PHST2: PHS + (Q3+D1) Query Transliteration</i>	0.80	66.86	68.90	67.87
<i>PHSTT: PHST1 + PHST2</i>	0.80	78.43	73.17	75.71
<b><i>COMBINE: MTS+ETS+PHSTT</i></b>	0.80	<b>97.82</b>	<b>82.32</b>	<b>89.40</b>

Table 3. Best Performance

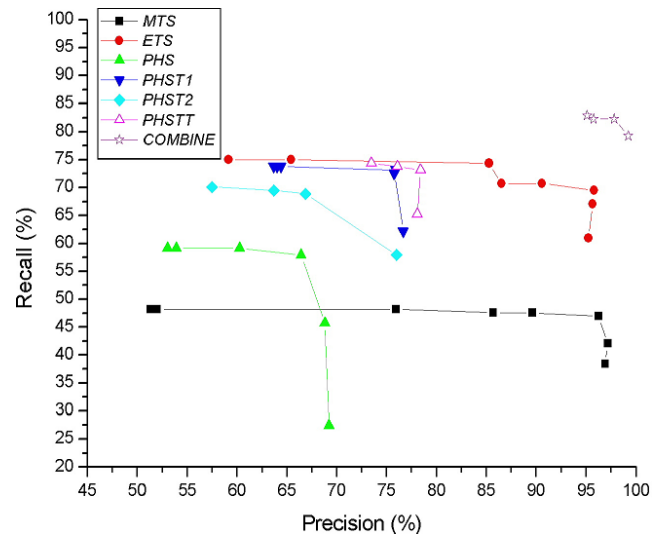


Figure 2. Overall Performance

We can see that searching for the English Query directly in the Mandarin phones can obtain comparable results (61.89%) to searching the MT output (63.11%), but it performed worse than another higher baseline of searching ET output (80.57%). However, adding query translation and transliteration significantly enhanced its performance to 75.71%. Adding query transliteration provided small but consistent gains for each threshold setting (1.2% gain for best f-measure).

The main observation is that combining all various approaches together achieved much better Precision and Recall (about 89.4% F-measure) than each alone. In the next subsection we shall give more detailed analysis about how phone-based searching and conventional translation-based approach are complementary.

From Table 3 we can also see that the best thresholds for all approaches were in the range of [0.7, 0.8], which indicated that fuzzy searching performs better than exact searching (represented

as the rightmost points on the curves in Figure 2 with threshold  $t=1$ ) for CLSSR, regardless of the setup of query and candidate document. Compare to exact matching, fuzzy matching has the advantage of matching different spellings for the same entity (“Al-Maliki” = “Maliki”), sequences with different spellings (“pei luo xi” = “Pelosi”), as well as being robust to recognition errors (“bu le er” = “Blair”). [12] used the same idea of fuzzy matching and ranked documents based on the matching confidence in a monolingual French task in CLEF 2005. [13] applied fuzzy matching in cross-lingual track of CLEF 2001 to recover some query terms and also found some relevant query terms during dictionary look-up.

### 5.3. Comparison on Each Query

Setting the threshold  $t=0.8$ , we found that only for about 31% of the queries did these three approaches obtain the same performance. For 27% of the queries *PHSTT* performed better than both *MTS* and *ETS*; while for 21% of the queries either *ETS* or *MTS* was superior to *PHSTT*. This provides strong evidence that these two general searching approaches – phone-based and translation-based – complement each other.

Most cases in which the translation approach performed worse involved the document processing errors as presented in section 3.2; about 65% of these queries involve names of Chinese persons. Interestingly, we also found that since the MT system is not sensitive to phone information, even when ASR successfully decoded some names such as “Song Dahan”, MT mistakenly translated it into traditional pinyin sequence “Song Dae-Keun”.

On the other hand, unlike the translation-based approach, phone-based searching does not incorporate any context or frequency information, so it tends to produce more confusion with common words even for very famous names. For example, “bu shi (Bush)” in Mandarin phones can also represent the word “is not”; “wu yi (Wu Yi)” can also represent the word “no doubt”. Therefore relying on phone matching alone will produce some more false alarms. Also if the query and the document involve two name variants referring to the same famous entity (e.g. “Chen Feng Fu Zhen” = “Margaret Chan”), statistical MT using language modeling can in fact produce high-quality translations to match the query name. In addition, phone matching had difficulty in matching those foreign names represented based on Chinese pronunciations in documents (e.g. Vietnamese name “rwan ming zhe” for “Nguyen Minh Triet”). However MT system may translate them correctly if they appear in bitexts.

## 6. CONCLUSIONS AND FUTURE WORK

Information Retrieval techniques have been quite successful for text genres such as newswire, but when adapting the same techniques to spoken documents the system performance suffers from ASR and MT errors. In this paper we followed the idea of taking advantage of audio source itself, and investigated how to effectively incorporate phone matching into CLSSR. Experimental results on Mandarin-English CLSSR showed that this new approach and conventional pipelines are complementary. The method should be applicable for other language pairs as long as person names are mostly represented based on pronunciations.

In the short-term we are planning the following improvements.

We have only exploited 1-Best phones from ASR so far; in the future we are interested in using graphemes and phone lattice as described in [3, 4] to get a better match. This scheme should be

able to fix some phone decoding errors which we didn’t address in this paper. For example, “Medici” was mistakenly decoded to the Mandarin phones “wai di qi” but a better result “mai di qi” does exist in the phone lattice. Also using graphemes can improve matching more OOV words.

In our current fuzzy matching algorithm, we assign equal weights to different edit operations. [14, 15] showed that automatically learned weights can achieve better matching accuracy. We plan to explore similar approach for CLSSR.

We are also interested in training an ET system (name tagging and name translation) on Mandarin phones instead of Chinese words, so that when searching queries we can assign more weight to those phone sequences which are tagged as names.

Ultimately we want to use our system output to correct ASR and MT output so it can benefit downstream processing tasks such as Information Extraction and Question Answering.

## ACKNOWLEDGMENTS

This material is based upon work supported by the Defense Advanced Research Projects Agency under Contract No. HR0011-06-C-0023, and the National Science Foundation under Grant IIS-00325657. Any opinions, findings and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the U. S. Government.

## 7. REFERENCES

- [1] David D. Palmer and Mari Ostendorf. Improving out-of-vocabulary name resolution. 2005. *Computer Speech & Language*. Volume 19, Issue 1, pp. 107-128.
- [2] Helen M. Meng, Wai-Kit Lo, Berlin Chen and Karen Tang. 2001. Generating Phonetic Cognates to Handle Named Entities in English-Chinese Cross-Language Spoken Document Retrieval. *ASRU 2001*.
- [3] Tee Kiah Chia, Khe Chai Sim, Haizhou Li and Hwee Tou Ng. 2008. A Lattice Based Approach to Query-by-Example Spoken Document Retrieval. *SIGIR 2008*.
- [4] Murat Akbacak, Dimitra Vergyri and Andreas Stolcke. 2008. Open-Vocabulary Spoken Term Detection Using Grapheme-Based Hybrid Recognition Systems. *ICASSP 2008*.
- [5] Mei-Yuh Hwang, Gang Peng, Wen Wang, Ario Faria, Aaron Heidel and Mari Ostendorf. 2007. Building a Highly Accurate Mandarin Speech Recognizer. *ASRU 2007*.
- [6] Richard Zens, Oliver Bender, Sasa Hasan, Shahram Khadivi, Evgeny Matusov, Jia Xu, Yuqi Zhang and Hermann Ney. 2005. The RWTH Phrase-based Statistical Machine Translation System. *IWSLT 2005*.
- [7] Heng Ji, Matthias Blume, Dayne Freitag, Ralph Grishman, Shahram Khadivi and Richard Zens. 2007. NYU-Fair Isaac-RWTH Chinese to English Entity Translation 07 System. *NIST ET 2007 PI/Evaluation Workshop*.
- [8] Sadao Kurohashi, Toshihisa Nakamura, Yuji Matsumoto and Makoto Nagao: Improvements of Japanese Morphological Analyzer JUMAN. 1994. *The International Workshop on Sharable Natural Language Resources*, pp.22-28.
- [9] Dayne Freitag and Shahram Khadivi. 2007. A Sequence Alignment Model Based on the Averaged Perceptron. *EMNLP-CONLL 2007*.
- [10] F.J. Damerau. 1964. A technique for computer detection and correction of spelling errors, *Communications of the ACM*.

- [11] V.I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*.
- [12] Annabelle Mercier, Amélie Imafouo and Michel Beigbeder. 2005. ENSM-SE at CLEF 2005 : Using a Fuzzy Proximity Matching Function. 6th Workshop of the cross-language evaluation forum (CLEF 2005)..
- [13] Fredric C. Gey, Hailing Jiang, Vivien Petras and Aitao Chen. 2001. Cross-Language Retrieval for the CLEF Collections - Comparing Multiple Methods of Retrieval. *CLEF 2001*.
- [14] Daniel Gillick, Dilek Hakkani-Tur and Michael Levit. 2008. Unsupervised Learning of Edit Parameters for Matching Name Variants. *ICASSP 2008*.
- [15] Inderjeet Mani, Alex Yeh and Sherri Condon. 2008. Learning to Match Names Across Languages. *COLING 2008 Workshop on Multi-source, Multilingual Information Extraction and Summarization*.