ISCA Archive
http://www.isca-speech.org/archive

Speech and Language Technology in Education (SLaTE2011)
Venice, Italy
August 24-26, 2011

Statistical Machine Translation Framework for Modeling Phonological Errors in Computer Assisted Pronunciation Training System

Theban Stanley, Kadri Hacioglu, Bryan Pellom

Rosetta Stone Labs, Boulder, Colorado, USA
{tstanley,khacioglu,bpellom}@rosettastone.com

Abstract

Computer Assisted Pronunciation Training (CAPT) is becoming more and more popular among language learners. Most effective CAPT systems take advantage of the learner's L1 and cater exercises and feedback specific to the language transfer effects. This paper presents a statistical machine translation (MT) based approach to model salient phonological errors present in an L1 population. The output of the MT system is coupled with a speech recognition system to detect non-native phonological errors. On a Korean learners of English corpus, the MT approach shows a 32.9% relative improvement in phone error detection and a 49% relative improvement in phone error identification compared to edit distance based modeling techniques. Similar performance improvements were observed on Japanese learners of English corpus.

Index Terms: phonological error modeling, machine translation, speech recognition

1. Introduction

The use of technology in classrooms has been steadily increasing in the past decade and the comfort level of students in using technology has never been higher. Computer Assisted Pronunciation Training (CAPT) has been quietly inching its way into many language learning curriculum. The high demand and shortage of language tutors especially in Asia has lead to CAPT systems playing a prominent and increasing role in language learning [1].

CAPT systems can be very effective among language learners who prefer to go through the curriculum at their own pace. Also, CAPT systems exhibit infinite patience while administering repeated practice drills which is a necessary evil in order to achieve automaticity. Most CAPT systems are L1 independent and cater to a wide audience of language learners from different language backgrounds. These systems take the learner through pre-designed prompts and provide limited feedback based on the closeness of the acoustics of the learners' pronunciation to that of native/canonical pronunciation [2]. In most of these systems, the corrective feedback, if any, is implicit in the form of pronunciation scores. The learner is forced to self-correct based on his/her own intuition about what went wrong. This method can be very ineffective especially when the learner suffers from the inability to perceive certain native sounds.

A recent trend in CAPT systems is to capture language transfer effects [4][5] between the learner's L1 and L2 languages. This makes the CAPT system better equipped to detect, identify and provide actionable feedback to the learner. These specialized systems have become more viable with enormous demand for English language learning products in Asian countries like China and India [1]. If the system is able to successfully pinpoint errors, it can not only help the learner identify and self-correct a problem, but can also be used as input for a host of other applications including content recommendation systems and individualized curriculum-based systems. For example, if the learner consistently mispronounces a phoneme contrast pair, he/she can be recommended remedial perception exercises before continuing the speech production activities. Also, language tutors can receive regular error reports on learners, which might be very useful in periodic tuning of customizable curriculum.

Linguistic experience and literature can be used to get a collection of error rules that represent negative transfer effects for a given L1-L2 pair. But this is not a foolproof process as most linguists are biased to certain errors based on their personal experience. Also, there are always inconsistencies among literature sources that list error rules for a given L1-L2 pair. Most of the relevant studies have been conducted on limited speaker population and most of them lack sufficient coverage of all phonological error phenomena [5]. It might be very convenient and cost effective to automatically derive error rules from L2 data.

In [5], Wai-Kit Lo et al. have tried automatically deriving context sensitive phonological rules by aligning the canonical pronunciations against phonetic transcriptions obtained from an annotator. Most alignment techniques used in similar automated approaches are variants of the basic edit distance (ED) algorithm. The algorithm is constrained to one to one mapping which is ineffective in discovering phonological error phenomena that occur over phone chunks. As edit distance based techniques poorly model dependencies between error rules, it's not straightforward to generate all possible non-native pronunciations given a set of error rules. Extensive rule selection and application criteria needs to be developed as it not modeled as part of the alignment process.

In order to remedy some of these short comings, in this paper, we formulate the phonological error modeling problem as a machine translation (MT) problem. The canonical pronunciation is considered to be in the source language and we attempt to generate the best non-native pronunciation (target language) that is a good representative translation of the canonical pronunciation (for a given L1 population). We can demonstrate that a statistical machine translation based framework can not only model phonological errors but also model dependencies between error rules. The framework also provides a more principled search paradigm that can generate N-best non-native pronunciations for a given canonical pronunciation.

2. Statistical Machine Translation Framework

Machine translation is the problem of generating the best sequence of words in the target language that is a good representation of a sequence of words in the source language. The Bayesian formulation of the MT problem is as follows:

$$P(T \mid S) = \arg\max_{T} P(S \mid T) \cdot P(T) \qquad (1)$$

Where, T and S are word sequences in the target and source languages respectively. P(S|T) is the translation model that models word/phrase correspondences between the source and target languages. P(T) represents the language model of the target language. In this paper, we utilize the Moses [6] phrase-based machine translation system for phonological error modeling. In the following subsections, we will give a brief overview of Moses.

2.1. Translation model estimation

Estimation of the translation model requires a parallel corpus of sentences in the source and target languages. Word alignments between the source and target language are obtained using Giza++ tool kit [7]. Giza++ is the implementation of the original IBM machine translation models. It has some drawbacks including limitation to one-to-one mapping which is not necessarily true for most language pairs. In order to obtain more realistic alignments, Moses applies a series of transformations to the word alignments produced by Giza++ to grow the word alignments into phrasal alignments. The parallel corpus is aligned in both directions i.e., source language against the target language and vice versa. The two word alignments are reconciled by obtaining the intersection which gives the high precision alignment points (the points carrying high confidence). By taking the union of these two alignments one can obtain the high recall alignment points. In order to grow the alignments, the space between the high precision alignment points and high recall alignment points is explored. Moses starts with the intersection of the two word alignments and then adds new alignment points that exist in the union of the two word alignments. It uses various criteria and expansion heuristics for growing the phrases as explained by Och and Ney in [7]. This process generates phrase pairs of different word lengths with corresponding phrase translation probabilities based on their relative frequency of occurrence in the parallel corpus.

2.2. Language model estimation

The language model learns the most probable sequence of words that occur in the target language. It guides the search during the decoding phase by providing prior knowledge about the target language. Moses can read language models created from popular open source language modeling toolkits like SRI-LM, RandLM and IRST-LM. In this paper, we used the IRST-LM tool kit [8] to estimate the language models.

2.3. Decoder

The Moses decoder implements a beam search [9] to generate the best sequence of words in the target language that represents the word sequence in the source language. At each state, the current cost of the hypothesis is computed by combining the cost of previous state with the cost of the translating the current phrase and the language model cost of the phrase. The cost also includes a distortion metric that takes into account the difference in phrasal positions between the source and the target language. Competing hypotheses can potentially be of different lengths and a word can compete with a phrase as a potential translation. In order to solve this problem, a future cost is estimated for each competing path. As the search space is very large for an exhaustive search, competing paths are pruned away using a beam which is usually based on a combination of a cost threshold and histogram pruning. Additional details of the decoding algorithm are described in [9].

3. Modeling Phonological Errors of an L1-population

In this work, we implicitly model the phonological errors in L2 data by reformulating it as a machine translation problem. The native/canonical phone sequence is considered to be in the source language and we attempt to generate the best non-native phone sequence (target language) that represents a good translation of the canonical phone sequence. The corresponding Bayesian formulation looks like this:

$$P(NN \mid N) = \arg\max_{NN} P(N \mid NN) \cdot P(NN) \quad (2)$$

Where, N and NN are the corresponding native and non-native phone sequences. P(N|NN) is the translation model which models the phonological transformations between the native and non-native phone sequences. P(NN) is the language model for the non-native phone sequences, which models the likelihood of a certain non-native phone sequence occurring in L2 data.

3.1. Training the Phonological Error model and Non-native Phone Language model

A parallel phone corpus of canonical and annotated phone sequences is run through Giza++ [7] to obtain phone alignments. The phone alignments from Giza++ are loaded into Moses to grow the one to one alignments into phone-chunk based alignments. This process is analogous to growing word alignments into phrasal alignments in traditional machine translation. The resulting phonological error model has phone-chunk pairs with differing phone lengths and a translation probability associated with each one of them.

The non-native phone language model was trained using IRSTLM toolkit [8] by feeding in annotated phone sequences from the L2 data. It's a 3-gram phone model with Witten-Bell smoothing applied to its probabilities.

Given the phonological error model and a non-native phone language model, the Moses decoder can generate the N-best non-native phone sequences for a given canonical native phone sequence.

3.2. Advantages of the MT approach

The MT approach offers several distinct advantages over previous phonological modeling approaches. In contrast to knowledge-based approaches in which linguistic rules are applied to model an L1 population, the MT approach automatically learns all phenomena that consistently occur in the annotated data. This includes language transfer effects and other phenomena like mispronunciations caused by interference due to word orthography. For example, non-native speakers of English often pronounce the "b" in the word "debt".

At an algorithmic level, the MT approach employs a more complex algorithm than an edit distance based extraction of phonological errors. It's capable of one-to-many, many-to-one and many-to-many alignments which allow for better modeling of phonological error phenomena that span across phone chunks. This can be especially effective in modeling severe insertion or deletion problems spanning across phoneme subsequences. For example, Korean learners of English often mispronounce the word refrigerator (pronounced "ɹ ə f ɹ ɪ ʤ ə ɚ eɪ ɾ ə ɚ") as "l i p ɹ ɪ ʤ i ɾ eɪ ɾ ə ɚ". The MT system has the ability to learn the correspondence between the native phone chunk "ɹ ə f" and the non-native chunk "l i p". Also, due to one-to-many alignment capability, it can learn the rule "ə ɚ"=>"i ɾ". In the case of an edit distance based approach,

this rule needs to be modeled as separate substitution and insertion rules. In traditional MT, word sequences in a language pair are not expected to be in the same order. In order to accommodate this behavior, a distortion metric is employed to allow alignments between words that are at different positions in the language pair. For example, the first word in a sentence in one language can be aligned to the last word in the corresponding sentence pair in another language. This feature can be useful in modeling after effects of certain mispronunciations that can manifest itself in the 2nd or 3rd syllable from the actual location of the mispronunciation.

Current approaches to phonological error modeling lack a strong mathematical framework that would facilitate auto-generation of non-native phone sequences. Most of these systems employ heuristic based rule selection and application techniques (as rule interdependencies are not explicitly modeled). The decoding paradigm that the MT approach offers is a much more principled way of combining error rule probabilities and interdependencies between error rules to generate the most probable non-native phone sequences.

4. Corpora

For experimentation, we collected and phonetically annotated a corpus consisting of Japanese learners of English (RS-JLE) and Korean learners of English (RS-KLE). The corpora consist of prompted speech data from an assortment of different types of content. It includes minimal pairs (e.g. right/light), stress minimal pairs (e.g. CONtent/conTENT), short paragraphs of text, sentence prompts, isolated loan words and words with particularly difficult consonant clusters (e.g. refrigerator). Phone level annotation was conducted on each corpus by three human annotators. The RS-JLE corpus used in this paper consists of 105 speakers who are residents of Japan. In total, data from 6 speakers were inter-annotated and another 15 speakers were intra-annotated. The RS-KLE corpus consists of 111 speakers of which data from 15 speakers have been inter-annotated and another 15 speakers are intra-annotated. We used 80% of the speakers for model training and the rest 20% for evaluation. Each speaker recorded only one session in the corpus and there is no speaker overlap between the test and train data. As each speaker recorded roughly the same amount of data, a speaker level split equates to a rough 80-20 split at the corpus level. The train split was used for modeling the phonological errors using MT. The test split consisted of inter and intra graded directories and was used for evaluation of the MT approach against an edit distance based approach.

5. System Architecture

Figure 1 summarizes the phonological error modeling as described in section 3 and its coupling with Rosetta Stone's speech recognition engine (SRE). We utilize the phonological error model and non-native phone language model to automatically generate non-native alternatives for every native pronunciation. The Moses decoder has the ability to generate N-best lists and based on empirical observations, we use a 4-best list to strike a good balance between under generation and over generation of pronunciation alternatives. The generated non-native lexicon (includes canonical pronunciations) along with an American English acoustic model is used to recognize the spoken utterance. For this paper, we assume the expected utterance to be produced is known and perform utterance verification followed by a Viterbi alignment of the audio and the expected text. The search space is constrained to the native and non-native variants of the expected utterance. The phone sequence which maximizes the Viterbi path probability is then aligned against the native/canonical phone sequence to extract the phonological errors produced by the learner.

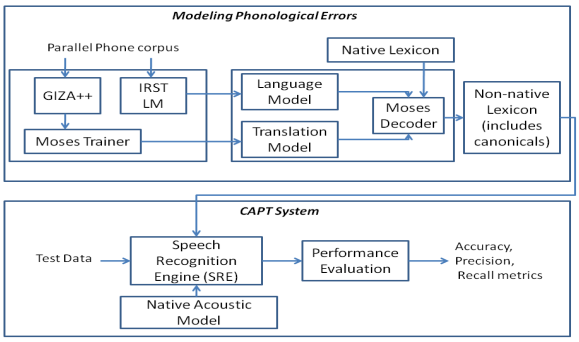

Figure 1: *Block Diagram summarizing Phonological Error modeling and its coupling with the SRE.*

6. System Evaluation

The MT approach was evaluated against an edit distance based approach similar to the one described in [5]. For the MT approach, the system described in Figure 1 was used to detect phonological errors in the test set. In order to build the edit distance based baseline system, we initially extracted phonological errors using ED from the training set. The phonological errors were ranked by occurrence probability. From empirical observations, we set the cutoff probability threshold at 0.001. This gave us approximately 1500 frequent error patterns. The frequent error rules were loaded into the Lingua Phonology Perl module [10] to generate non-native phone sequences. The tool was constrained to apply rules only once for a given triphone context as the edit distance approach does not model interdependencies between error rules. The N-best list obtained from the Lingua module was ranked by the occurrence probability of the rules that were applied to obtain that particular alternative. The non-native lexicon was created with an N-best cutoff of 4 so that it's comparable to the non-native lexicon produced by the MT approach. The approaches were evaluated using the following metrics: (i) Overall accuracy of the system (ii) Diagnostic performance as measured by precision and recall (iii) F-1 score, which is the harmonic mean of precision and recall. It gives us one number to track changes in operating point of the system. These metrics were calculated for the phone detection and phone identification tasks along with their corresponding human annotator upper bounds.

Table 1. *Performance of MT and ED approaches normalized to Human performance (set at 100%) in phone error detection*

RS-KLE	Accuracy	Precision	Recall	F-1
ED	92.7	58.2	41.1	48.7
MT	91.9	64.6	64.9	64.7

RS-JLE	Accuracy	Precision	Recall	F-1
ED	90.0	66.3	48.1	56.2
MT	90.2	72.4	71.3	71.9

6.1. Phone error detection

Phone error detection is defined as the task of flagging a phoneme as containing a mispronunciation. The accuracy metric measures overall classification accuracy of the system on the phone error detection task, while precision and recall measure the diagnostic performance of the system. Precision measures the number of correct mispronunciations over all the

mispronunciations flagged by the system. Recall measures the number of correct mispronunciations over the total number of mispronunciations found in the test set (as flagged by the annotator).

As shown in Table 1, across the corpora, the MT system achieves between 65 to 72% of the performance achieved by humans on F-1 score. The more holistic modeling approach employed by the MT system is evidenced by higher normalized performance (NP) in recall in comparison to precision. The MT system achieves a 28-33% relative improvement in F-1 in comparison to the ED approach. Figure 2 shows NP on F-1 for varying number of pronunciation alternatives. There is a significant increase in performance for lexicons with 3-4 best alternatives beyond which the performance asymptotes.

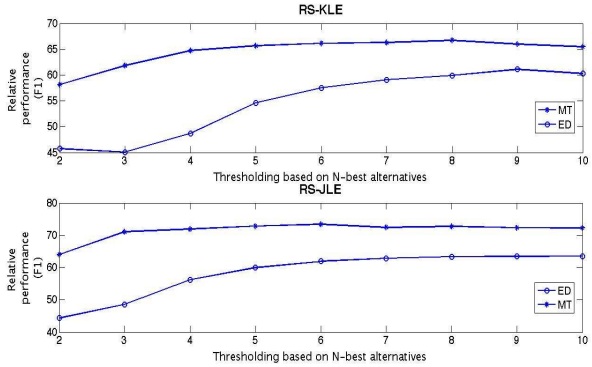

Figure 2: *The normalized performance of F-1 score in phone error detection for varying number of pronunciation alternatives*

6.2. Phone identification

Phone identification is defined as the task of identifying the phone label spoken by the learner. The identification accuracy metric measures the overall performance on the identification task. Precision measures the number of correctly identified error rules over the total number of error rules discovered by the system. Recall measures the number of correctly identified error rules over the number of error rules in the test set (as annotated by the human annotator).

Table 2. *Performance of MT and ED approaches normalized to Human performance (set at 100%) in phone error identification*

RS-KLE	Accuracy	Precision	Recall	F-1
ED	94.1	47.5	33.5	39.7
MT	92.5	59.1	59.4	59.2

RS-JLE	Accuracy	Precision	Recall	F-1
ED	91.8	62.2	45.1	52.8
MT	90.9	72.0	70.9	71.4

Table 2 shows that the MT approach achieves a 59-71% NP on F1-score across the corpora. This constitutes a 35-49% relative improvement compared to the ED approach. Given the difficulty of error identification task, it should be noted that the performances are relatively lower in comparison to phone error detection. Similar to the behavior in phone error detection, Figure 3 shows that the highest NPs are achieved with 3-4 best alternatives.

7. Conclusions

We present an automatic, data driven approach to modeling phonological errors in L2 data. The MT approach not only offers conceptual advantages (as elaborated in section 3.2) to existing techniques but also produces significant improvements on phone error detection and phone error identification tasks. On the RS-KLE corpus, the MT approach achieves a 32.9% relative improvement (F-1 score) in phone error detection and a 49% relative improvement in phone error identification compared to the edit distance based approach. On RS-JLE corpus, a 27.8% relative improvement was achieved in phone error detection and a 35.3% improvement in phone error identification. Tighter coupling between the MT system and the SRE is a good candidate for future exploration. Likelihood scores from the phonological error model can be imported into the SRE to bias the recognition results. This concept is analogous to language model weights used in traditional large vocabulary speech recognition. This could lead to precise tuning of CAPT systems to the proficiency level of the language learner.

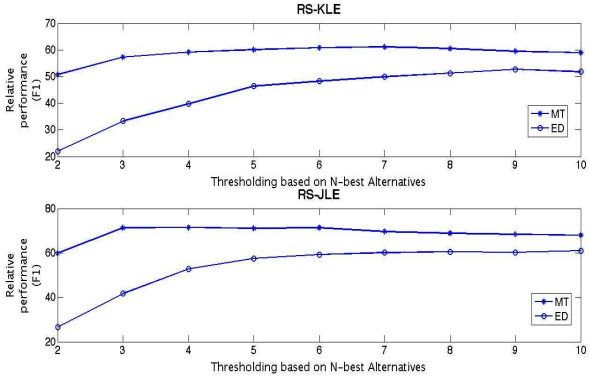

Figure 3: *The normalized performance of F-1 score in phone error identification for varying number of pronunciation alternatives*

8. References

[1] Kachru, B. B., Asian Englishes: Beyond the Canon, Hong Kong University Press, 2005.

[2] Franco, H. et al., "The SRI EduSpeak System: Recognition and Pronunciation Scoring for Language Learning", Proc. of InSTIL, Scotland, 123-128, 2000.

[3] Meng, H., Lo, Y. Y., Wang, L., and Lau, W. Y., "Deriving Salient Learners' Mispronunciations from Cross-language Phonological Comparisons," in Proc. of ASRU2007.

[4] Harrison, A. M., Lau, W. Y., Meng, H., and Wang, L., "Improving Mispronunciation Detection and Diagnosis of Learners' Speech with Context-sensitive Phonological Rules based on Language Transfer," in Proc. of INTERSPEECH2008.

[5] Lo W. K., Zhang, S., and Meng, H., "Automatic Derivation of Phonological Rules for Mispronunciation Detection in a Computer-Assisted Pronunciation Training System", Interspeech, Makuhari, Japan, 2010.

[6] Koehn, P., et al., "Moses: Open Source Toolkit for Statistical Machine Translation", Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic, June 2007.

[7] Och, F. J., Ney, H., "A Systematic Comparison of Various Statistical Alignment Models", Computational Linguistics, volume 29, number 1, pp. 19-51, March 2003.

[8] Federico, M., Bertoldi, N., Cettolo, M., "IRSTLM: an Open Source Toolkit for Handling Large Scale Language Models", Proceedings of Interspeech, Brisbane, Australia, 2008.

[9] Koehn, P., "Pharaoh: A Beam Search Decoder for Phrase-Based Statistical Machine Translation Models", AMTA, 2004.

[10] Lingua Phonology Perl module, http://search.cpan.org/~jaspax/Lingua-Phonology-0.33