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

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

Identifying Confusable Contexts for Automatic Generation of Activities in Second Language Pronunciation Training

Oscar Saz, Maxine Eskenazi

Language Technologies Institute (LTI), Carnegie Mellon University (CMU), Pittsburgh, USA
{osaz,max}@cs.cmu.edu

Abstract

This paper presents a new technique in the development of tools for Computer-Aided Language Learning (CALL) in Second Language (L2) pronunciation training. Instead of the usual pinpointing and feedback strategies, which address all production errors in the L2 speaker equally, this technique aims at honing in on where a pronunciation error can lead to a confusion in the transmission of the message and providing these contexts for practice. To provide this new correction strategy, it is necessary to first identify the words and contexts which can lead to a confusion in a message when there is a mispronunciation in them. This paper describes a Basic Identification of Confusable Contexts (BICC) procedure for detecting these cases providing a measure of their relevance. It also includes a methodological validation of BICC with humans assessing the quality and comprehensibility of a dataset of automatically generated sentences which shows that the prediction of these contexts is possible with good precision.

Index Terms: second language, pronunciation training, natural language processing, confusable contexts

1. Introduction

There is currently a steady effort in the development of Computer-Aided Language Learning (CALL) tools for the pronunciation training of Second Language (L2) learning. These tools make use of various pinpointing techniques, Goodness of Pronunciation (GOP) [1] being the most popular one [2]. They have shown a solid ability to detect phonetic errors and inaccuracies in non-native speech in a variety of languages and tasks [3, 4]. When the pinpointing algorithm detects an error in the pronunciation of a non-native student, feedback is provided to the student showing this error and strategies for the student to improve pronunciation in the target language.

Nevertheless, all of these are purely focused on the production of speech; that is, they only check whether the student's speech is correct from the point of view of acoustic production and they do not consider the overall effect that a mispronunciation has on the perception and comprehension of a human listening to that speech. The result of this is that all the errors made by the L2 student are treated and presented equally when providing feedback to the student. However, when considering the influence of pronunciation errors, it can be argued how, in many cases, the error can be recovered by the listener (for instance, a native speaker of the language) due to the ability that humans have to predict meaning from context. This situation allows non-native speakers to be effective in communicating in the new language even if they still produce several errors in their speech.

On the other hand, there are some cases where a single pronunciation error can be critical in the efficient transmission of information from the non-native speaker to the other partner in the conversation. This is a case that might happen, for instance, with words that have minimal pairs (other words which only differ in one phoneme, like /*pin*/ with /*bin*/ or /*bin*/ with /*bean*/). When this happens, the interlocutor of the speaker will have to ask for a clarification or, in the worst case, will interpret a totally different meaning. Minimal pairs have been the subject in L2 pronunciation tutors before [5]; but, again, they only focused on evaluating the production of the phonemic contrast between the minimal pairs and did not consider the issues in perception and comprehension when there was a mistake between the pairs.

A tool which could assess and correct these cases that present comprehension issues would be very interesting for L2 speakers and a good complement to regular L2 classes and existing pinpointing tools, as it would provide a detection of pronunciation errors actually linked to the process of communication (which is the ultimate goal in learning a new language) in the same way experienced by non-natives when they start to get engaged in real dialogs with native speakers. The development of this type of tools requires a multi step process: First, it is necessary to understand how confusions happen and what leads to them in order to develop techniques which effectively identify contexts in which they are present; afterwards, a correct interface has to be designed that really engages and motivates the user; and, finally, evaluation with real students has to be carried out to assess how helpful these tools would really be.

This paper deals with the first step in this work. In order to create activities for L2 pronunciation training following this paradigm, it is necessary to develop a procedure which can automatically identify the cases of likely confusion from a given piece of text and select those which are most interesting to be practiced by the student. This would become the basis of a practice software for L2 students which, then, would perform pinpointing to detect if the student correctly pronounced the confusable word or made a critical error creating a confusion.

The organization of this paper is as follows: Section 2 provides the definition of confusable contexts and the basic methodology used to identify confusions from a given text. Then, Section 3 presents an evaluation carried out to determine how confusable the identified contexts are according to a human-based evaluation. Finally, Section 4 provides the conclusions and future work that this leads to.

2. Definition of Confusable Contexts

A problem in oral communication might appear when the perception and understanding of a spoken message by a person is different from the meaning that the speaker tried to convey. This can happen as a misunderstanding, when the listener gets an incorrect message but accepts it as the correct message; or as

a confusion when the listener has problems to understand the message and, thus, will require further information. Both situations slow down communication and make the transmission of information troublesome. There are many reasons why this happens, related to errors in the production of the message by the speaker due to poor grammar and syntax or to pronunciation errors; difficulties in the perception and comprehension by the listener or environment conditions like excessive noise.

This work is focused on identifying these effects due to pronunciation errors by the speaker. Pronunciation errors may change a relevant word in the message (noun, verb, adjective, adverb) and, hence, create misunderstanding or confusion in the listener. Other effects in understanding like errors in suprasegmental features and prosody by the non-native speaker are not covered in this work, but they could required for further work.

Confusions due to mispronunciations appear when an N-word sentence, $S = \{W_1, ..., W_n, ..., W_N\}$, is converted into an alternative sentence S'_{i_j} with the change of the i-th word W_i into the minimal pair W'_{i_j} out of the M minimal pairs, $\{W'_{i_1}, ..., W'_{i_m}, ..., W'_{i_M}\}$, existing for W_i. Also, the sentence S'_{i_j} has to be grammatically and syntactically correct and provide a different meaning than the original sentence S.

2.1. Basic Identification of Confusable Contexts (BICC)

The Basic Identification of Confusable Contexts (BICC) procedure is a two-step algorithm which makes use of the Carnegie Mellon University Pronunciation Dictionary (CMUDict) [6] and the Stanford Part-of-Speech (PoS) tagger [7] for automatically providing potential confusable contexts.

Initially, every word W_i in the sentence is fed to CMUDict to find the corresponding pronunciations of the word. All combinations of phoneme substitution, insertion and deletion are made over the original pronunciation and those that are the pronunciation of a real word in CMUDict are retained as minimal pairs of the word ($\{W'_{i_1}, ..., W'_{i_m}, ..., W'_{i_M}\}$). Then, the set of sentences S'_{i_j} is created where a word in the sentence (W_i) is replaced by one of its minimal pairs (W'_{i_j}).

The Stanford PoS tagger is used to prune those sentences which do not match the original PoS tags of the original sentence. The assumption for this is that a sentence can more likely create confusion if it has grammatical and syntax coherence, especially in comparison to the original sentence. The substitution, for example, of a verb in the sentence for a noun will less likely create confusion, even if they are minimal pairs, because a native listener would predict the verb for syntactic correctness.

With the proposed method a sentence like "The cat is white" produces 42 alternative sentences including "The cat is wide", "The can is white", and the "The cut is white". The minimal pairs involved in these sentences are $/white/$ ($/waIt/$ in SAMPA) with $/wide/$ ($/waId/$), $/cat/$ ($/k\{t/$) with $/can/$ ($/k\{n/$) or $/cut/$ ($/kVt/$).

2.2. Predictors of Confusability

A set of 3 predictors were initially believed to relate to how relevant the identified confusions were going to be. These measures were designed to predict how confusable and intelligible the generated sentences were, and to favor the most relevant ones in the practice of the students.

The first predictor works at the phonological level. It is based on how certain phonological contexts are more frequent than others and listeners prioritize words that contain them in a sentence. The phonological predictor of a word is the sum of the phonological frequency of all the triphones that form it normalized by the total number of triphones in it. The frequency of apparition of triphones was learned by counting its presence in all the words in CMUDict, where a total of 19,421 triphones are present. The most frequent triphone was $AH - N - SIL$ ($\{-n - SIL$ in SAMPA) with 9,180 appearances and the least frequent were 3,605 triphones like $AA - AA - K$ ($A - A - k$), or $Z - Y - OW$ ($z - I - oU$) with only one appearance.

The second predictor is the frequency of apparition of the word inserted in the sentence, which was also considered relevant when looking for the intelligibility of a sentence. Very frequent words are expected to be more likely accepted by a listener than the infrequent ones, as humans are more accustomed to them than to words that they have rarely seen before. The word frequencies can be easily learned from large text corpora which model a given language or a domain of the language, by simply counting the repetitions of each word.

The final predictor relates to the co-occurrences of the word inserted in the sentence with the surrounding words. Co-occurrence is a measure of how habitually certain words appear together in sentences and phrases in a domain or context. Highly co-occurring words create clusters of words (Bag-of-Words) where the apparition of some of the words is a predictor of the future presence of the rest of the words in the group. Two measures were used to define co-occurrence: On one hand, the frequency of co-occurrence is defined as the number of times a certain word appears in the same sentence with another one; and, on the other hand, the distance of co-occurrence measures the distance (in words) between those two words when they appear in the same sentence. Co-occurrences are also learned from text corpora in a language and/or domain.

3. BICC Assessment

The validation of the BICC procedure was designed to assess if the automatically generated sentences containing minimal pairs could be understood as real sentences by humans and, hence, would be potential sources of confusion. Although the ultimate interest of this work is the effect of confusions in spoken language, it was considered necessary to assess the quality of the BICC in text before starting the resource-consuming process of acquiring and labeling a speech corpus suitable for this validation.

3.1. Methodological Framework

To find a sufficiently large number of examples to process, the 10k Corpus, a large corpus of text from the business and financial domains from 1996 to 2006 [8], was used. Although this type of domain was far from ideal for L2 learning and their possible sentences had little interest for L2 learners, it provided a large amount of data to analyze and the results achieved could be generalized to any other corpus.

The subsets of this corpus that were chosen for training and testing are summarized in Table 1. The training subset was used to learn the frequency of the words in the corpus and the Bag-of-Words for each of the words as explained in Section 2.2. The number of different words in the training subset was about 196,000 words (including proper names and numbers).

Table 1: *Training and testing subsets from the 10k corpus*

	Years	Files	Sentences	Words
Training	1996-99	8,651	15,671,044	376,892,191
Testing	2000	1,613	484,383	11,353,653

Testing was carried out on sentences from the year 2000 subset of the corpus. The number of sentences in the test sub-

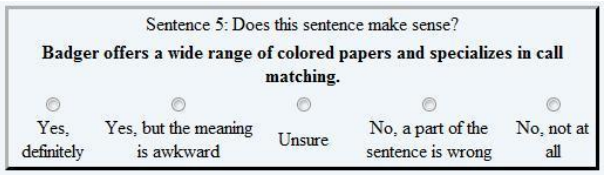

Figure 1: *Presentation of a sentence in the AMT task.*

corpus was large, nearly half a million as seen in Table 1, so a set of about 1,300 sentences with a length in words between 10 and 20 words from this subset were randomly selected and were processed with the BICC procedure. The confusable sentences obtained from these sentences were the final group of sentences for evaluation.

3.2. Validation through Human Perception

The only way to validate the level of comprehensibility of the automatically generated sentences was through human validation. However, to validate a large amount of sentences and achieve significant results it was not feasible to rely on a reduced set of local human experts, as this would have required a large amount of time and resources. For this reason, crowdsourcing through Amazon Mechanical Turk (AMT) was chosen to carry out the validation. Crowdsourcing has been shown to be a reliable way to assess the quality of the work in a variety of NLP tasks like machine translation [9] and speech labeling and transcription [10]. For these tasks, the knowledge of a group of naïve humans has been proven to equal the knowledge of experts [11]. A similar task was previously performed in [12], but it evaluated a much smaller number of sentences proposing different questions to the workers.

The task presented to the workers was entitled "Tell us if these are sentences that make sense" and they were asked to rate sentences according to the question "Does this sentence make sense?" with the following five possible answers:

- "Yes, definitely": A normal correct sentence.
- "Yes, but the meaning is awkward": A correct sentence with a word(s) out of context.
- "No, a part of the sentence is wrong": An incorrect sentence that is still readable in part.
- "No, not at all": An incorrect and unreadable sentence.
- "Unsure": When a decision is not clear.

A total of 11,790 sentences were evaluated this way in 1,310 Human Intelligence Tasks (HITs). Each HIT contained 9 automatically-generated sentences to evaluate and 1 of the original sentences as gold standard control. The expected answer to that sentence should be "Yes, definitely". Five different workers were requested to do each task, so in total 6,550 HITs were uploaded into AMT.

Figure 1 shows a screenshot of the presentation of the task to the AMT workers. Each sentence was displayed with the 5 choices provided for the worker to choose. The HIT included a set of examples, to help workers understand the procedure.

3.3. Results

The AMT task was completed in 4 days. 508 workers participated in the work, spending 86 seconds on average to complete each HIT. The distribution of all the single votes received by the 5 possible options can be seen in Figure 2. It shows how a very small percentage of votes (3.40%) were cast by the workers to sentences that did not make sense at all for them. This result

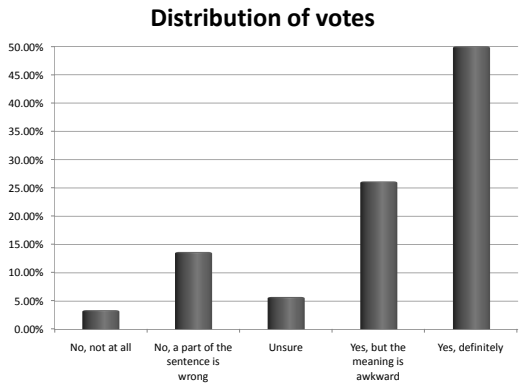

Figure 2: *Distribution of all the votes cast in the AMT task.*

was, hence, very positive as workers very rarely found the sentences to be totally unintelligible. The low number of votes for the "Unsure" option (5.75%) also showed that the workers were able to finish the work properly with the guidelines provided, without experiencing major difficulties with it.

A majority of the votes were cast towards the most positive option ("Yes, definitely"; 50.98%), indicating that workers gave most cases the highest understandability rating. The rest of the votes were cast for the two intermediate options (26.16% and 13.69%) providing a measure of cases where the sentences created comprehension issues for the workers.

The agreement among workers evaluating the same sentence was studied in terms of the weighted Kappa coefficient, which measures agreement in cases when the possible choices are ordered as in this case. The Kappa value measured was 0.4 that indicated a minimum agreement but it did not provide much evidence of real agreement among workers. The subjectivity of the task showed up as workers did not seem to agree to one precise category, rather presenting a distribution of votes around a certain value.

To give relevance to each individual vote, a scoring system was designed to translate the votes cast by the workers to a numeric score in the following way: "Yes, definitely" provided a score of 1, "Yes, but the meaning is awkward" provided a score of 0.75, "Unsure" provided a score of 0.5, "No, a part of the sentence is wrong" provided a score of 0.25 and "No, not at all" provided a score of 0. The final score awarded to each sentence was the average of the scores assigned by each worker evaluating the sentence. Thus, all sentences obtained a final score between 0 and 1.

To provide a descriptive view to the score of each sentence, three cases were defined depending on the final score obtained by them:

- Cases of misunderstanding, where the automatically created sentence had a high score (above 0.9) and was mostly understood by the workers as a real sentence
- Cases of confusion, where the sentence had a good score (between 0.5 and 0.9) and were mostly considered correct by the workers but with some errors.
- Erroneously generated sentences obtaining very low scores (below 0.5) and that were errors in the BICC procedure because workers did not see them as plausible sentences.

To avoid the noise introduced by the sentences with low agreement, only those where the standard deviation among votes was below 0.2 were studied. 77.74% of the sentences

Table 2: *Samples of automatically generated sentences and their score assigned by human voting*

Sentence (minimal pair is highlighted in bold face, with the original word in parenthesis)	Score
We could be adversely **effected** (affected) by continuing or sustained price discounting in the fast food industry.	1.0
All of the Company's stores are subject to inspection and regulation by public **help** (health) authorities.	0.95
The FDA **lessor** (letter) provided the conditions that must be satisfied before final approval.	0.85
The Option may be exercised in **sole** (whole) or in part at any time on or prior to June 15 2002.	0.75
Carter has been experiencing **rain** (pain) and discomfort and has been unable to return to work.	0.55
Upward trends in prices could have a **port** (short) term impact on margins.	0.35
This switch features a non blocking architecture to avoid the loss of **dates** (data) packets.	0.25

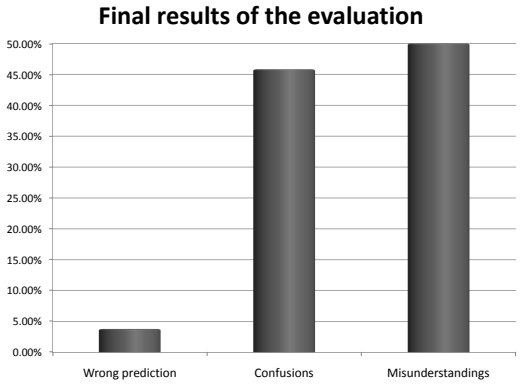

Figure 3: *Final results of the AMT task.*

fulfilled this condition and their distribution in the 3 final categories can be seen in Figure 3. Only 3.8% of the sentences were considered as wrongly identified confusions, while confusions and misunderstandings accounted for 45.8% and 50.4% respectively. With these result, the precision of the proposed results is 96.2%, although 23% of the sentences were not studied due to the agreement issues. A small sample of the sentences generated and the different final scores provided by the workers in the AMT task can be seen in Table 2. Sentences are presented belonging rated in three final categories created (misunderstandings, confusions and BICC errors).

When correlating the scores provided by the human raters to the predictors proposed in Section 2.2, however, there was no strong correlation. Word frequency and the Bag-of-Words frequency and distance were seen as weak predictors since sentences considered as possible misunderstandings and confusions had higher values of the predictors in mean that those considered as incorrectly identified confusable contexts. Several factors could influence this: first, the way workers consider sentences as intelligible depends on many factors not all of them reflected by the predictors; furthermore, the predictors measured the impact of the new word in the sentence and not the comprehensibility of the sentence as a whole.

4. Conclusions

This paper has presented a procedure for identifying confusable contexts, BICC, based on NLP tools, to process real sentences and automatically generate confusable sentences. These confusable sentences contain minimal pairs which can be mistaken in the pronunciation of a L2 learner and create misunderstandings or confusions in communicating with another person.

The results have showed that BICC can achieve a good performance in this task, as reported by humans through a crowd sourcing task. A precision of 95% was measured, although 22% of the sentences had to be removed from the analysis due to the lack of agreement among workers to provide a value. Further work oriented to design stronger predictors would help differentiate cases where misunderstandings from confusions, as their influence in human comprehension is different.

Once seen that that it is possible to identify these contexts in a reliable way, the future work consists in the inclusion of these confusable sentences as the basis of a new type of pronunciation tutor which provides feedback to the student when mispronunciations in confusable contexts are made. Future work will also address other causes of misunderstandings in L2 learners like prosody and events such as dropped schwas.

5. Acknowledgements

Oscar Saz is supported by a Fulbright/MEC scholarship. This work was partially supported by projects PSLC (NSF-SBE-0836012) and EDECAN (CICYT-TIN2008-06856-C05-04). The opinions expressed in this paper do not necessarily reflect those of NSF.

6. References

[1] S.-M. Witt and S.-J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Communication*, vol. 30, no. 2–3, pp. 95–108, 2000.

[2] H. Strik, K. Truong, F. deWet, and C. Cucchiarini, "Comparing different approaches for automatic pronunciation error detection," *Speech Communication*, vol. 51, no. 10, pp. 845–852, 2009.

[3] C. Cucchiarini, A. Neri, and H. Strik, "Oral proficiency training in Dutch L2: The contribution of ASR-based corrective feedback," *Speech Communication*, vol. 51, no. 10, pp. 853–863, 2009.

[4] M. Black, J. Tepperman, A. Kazemzadeh, S. Lee, and S. Narayanan, "Pronunciation verification of english letter-sounds in preliterate children," in *Proceedings of Interspeech*, Brisbane, Australia, 2008, pp. 2783–2786.

[5] Z. Handley, M. Sharples, and D. Moore, "Training novel phonemic contrasts: A comparison of identification and oddity discrimination training," in *Proc. of SLaTE*, Birmingham, UK, 2009.

[6] R.-L. Weide, "The CMU pronunciation dictionary," 2005. [Online]. Available: http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[7] K. Toutanova, D. Klein, C. Manning, and Y. Singer, "Feature-rich Part-of-Speech tagging with a cyclic dependency network," in *Proceedings of HLT-NAACL*, Berkeley, US, 2003, pp. 252–259.

[8] S. Kogan, D. Levin, B.-R. Routledge, J.-S. Sagi, and N.-A. Smith, "Predicting risks from financial reports with regression," in *Proceedings of NAACL-HLT*, Boulder, USA, 2009.

[9] V. Ambati, S. Vogel, and J. Carbonell, "Active learning and crowd-sourcing for machine translation," in *Proceedings of LREC2010*, Valletta, Malta, 2010.

[10] G. Parent and M. Eskenazi, "Toward better crowdsourced transcription: Transcription of a year of the Let's Go bus information system data." in *Proceesings of SLT2010*, Berkeley, USA, 2010.

[11] R. Snow, B. O'Connor, D. Jurafsky, and A.-Y. Ng, "Cheap and fast, but is it good? Evaluating non-experts annotation for natural language tasks," in *Proceedings of EMNLP*, Waikiki, USA, 2008.

[12] B. Lambert, R. Singh, and B. Raj, "Creating a linguistic plausibility dataset with non-expert annotators," in *Proceedings of Interspeech 2010*, Tokyo, Japan, 2010, pp. 1906–1909.