Repairing General-Purpose ASR Output to Improve Accuracy of Spoken Sentences in Specific Domains Using Artificial Development Approach

C. Anantaram, Sunil Kumar Kopparapu, Chirag Patel and Aditya Mittal
Innovation Lab, Tata Consultancy Services Limited, ASF Insignia, Gwal Pahari, Gurgaon, India
{c.anantaram, sunilkumar.kopparapu, patel.chiragkumar, mittal.aditya}@tcs.com

Abstract

General-purpose speech engines are trained on large corpus. However, studies and experiments have shown that when such engines are used to recognize spoken sentences in specific domains they may not produce accurate ASR output. Further, the accent and the environmental conditions in which the speaker speaks a sentence may induce the speech engine to recognize certain words/ sets of words inaccurately. Thus, the speech engine's output may need to be repaired for a domain before any further natural language processing is carried out. We present an artificial development (Art-Dev) based mechanism for such a repair. Our approach considers an erroneous ASR output sentence as a *biological cell* and repairs it through evolution and development of the inaccurate *genes* in the cell (sentence) with respect to the *genes* in the domain. Once the genotypes are identified, we 'grow' the genotypes into phenotypes to fill the missing gaps or erroneous words with appropriate domain concepts. We demonstrate our approach on the output of standard ASR engines such as Google Now and show how it improves the accuracy.

1 Introduction

Many working examples of automatic speech recognition (ASR) systems that convert human spoken speech into text can be seen in day to day use; popular examples are Google Now, IBM Watson and Siri. For freely spoken natural language sentences, the typical recognition accuracy achievable even for such state-of-the-art speech recognition systems have been observed to be about 60% to 90% in real-world environments [Fusayasu *et al*., 2015; Morbini *et al*., 2013]. However, such ASR engines may at times perform even worse when used in specific domains where specialized domain terms appear that are not available in general corpus [Twiefel *et al*., 2014]. Other factors such as variations in speaker's accent, background noise, poor ability to express on the part of the user also effect the accuracy of the ASR's output text. As a consequence if one takes the ASR output *as-is* and attempts natural language processing such as automated question-answering, of such erroneously and partially recognized text, then it becomes rather problematic. It is important, therefore, to repair the ASR output to improve the accuracy of the output-text vis-à-vis the specific domain.

In this demonstration, we present a mechanism to repair the ASR output of popular speech engines like Google Now and IBM Watson. Our mechanism is motivated by evolutionary development (Evo-Devo) inspired artificial development processes (Art-Dev) [Tufte, 2009; Harding and Banzhaf, 2008], and considers an erroneous sentence as a biological cell (*zygote*) and *grows* it through evolution and development of the *partial gene* present in the input sentence with respect to the genes in the domain. Once the *genotypes* are identified, we *grow* the genotypes into *phenotypes* to fill the missing gaps and replace erroneous words with appropriate domain concepts in the sentence. We present the results of such repair mechanism and demonstrate its usefulness.

2 Related Work

[Peng *et al*., 2013] have proposed a mechanism to improve ASR accuracy by re-ranking the N-best speech recognition hypothesis. Their method uses ASR output text as a set of queries for a web-search system and examine the search results to re-ranking the ASR hypothesis. [Fusayasu *et al*.,2015] use 'normalized relevance distance' as a measure for semantic similarity between words and use that to replace possibly erroneous words in the recognition. [Tur *et al*., 2013] use word confusion networks (WCNs) for more robust semantic parsing in a conditional random fields (CRF) framework to repair the errors and show significant semantic parsing performance improvements using WCNs. All the above methods, however, do not repair at the sentence level. In our work, we consider the ASR output as a biological cell that is repaired with respect to the environment through repair and development of its "*genes*" [Anantaram *et al*., 2015].

3 Repairing ASR output by Art-Dev method

In our approach the ASR output is treated as a *biological cell* that encodes 'species-specific' *genes* of the domain representing concepts in the sentences spoken by a speaker. However, in the process of ASR, wherein the spoken-audio is converted to text output, the genes are 'injured' during the recognition process. We repair all 'injured' genes through an artificial development approach, with evolution and development of the *partial gene* present in the ASR output with respect to the species-specific *genes* present in the domain (which encodes the concepts of the domain). In our context, the 'fittest' domain gene replaces the partial gene in the sentence. This is the first-level of repair. The set of all genes remaining in the sentence forms the *genotypes* of the sentence. Once the genotypes are identified, we grow them into *phenotypes* to remove the grammatical and linguistic errors in the sentence. The paritally repaired ASR sentence is parsed and the POS tags are evaluated to find any linguistic inconsistencies, which are then repaired. For example, a sentence may have a WP-tag (Wh-pronoun), but a WDT-tag may be missing (Wh-determiner). Using such clues we find appropriate parts and repair the sentence. Linguistic and semantic repairs form the genotype to phenotype repair – the second level of repair – and improves the sentence accuracy.

As an example, let the sentence spoken by a Human speaker (H1) on retail sales domain be "*In two thousand fourteen which industry had the peak sales*". In our experiment Google Now recognised this sentence as (G1): "in two thousand fourteen which industry had the pixels". Through our partial gene matching method, we find that there is no relationship in the domain ontology between industry and pixels, whereas there is a relationship between industry and peak sales. Further the phonetic match between pixels and peak sales gives us the fitness match to replace the inaccurately recognized word pixels with the domain word peak sales. Thus we repair (G1) to (E1): "in two thousand fourteen which industry had the peak sales" which is more accurate with respect to the domain.

4 Experiments

We show results of our experiments on a financial domain.

Human spoken sentence (H*x*), Google ASR output (G*x*), Sentence repaired by our method (E*x*)	**WER Accuracy % of G, E**
H2: The return on plan assets include interest earned, dividends earned, realized and unrealized gains or losses less taxes payable by the plan less administrative costs of the plan.	
G2: the return on plan acids include interest on dividends and realized and unrealized games or losses less taxes payable by the plan let administrative process of the plan	**G2:** 64.2
E2: the return on plan assets include interest on dividends and realized and unrealized gains or losses less taxes payable by the plan let administrative process of the plan	**E2:** 75.0
H3: Actuarial gain or loss refers to an increase or decrease to a company's estimate of the Fair Value of Plan Assets as a result of either change in assumption or experience adjustments.	
G3: actuarial gain or loss refers to an increase or decrease tour companies estimate of the fair value of planets as a result of either changed and assumption or experience at judgements	**G3:** 62.5
E3: actuarial gain or loss refers to an increase or decrease tour companies estimate of the fair value of plan assets as a result of either changed and assumption or experience at judgements	**E3:** 71.8
H4: Expenditure in foreign currency includes research development expenses and intangible assets related charges	
G4: Expenditure in foreign currency includes resource development expenses and intangible assets let it charge	**G4:**61.53
E4: Expenditure in foreign currency includes research development expenses and intangible assets charges	**E4:**92.3

Table 1: Evo-devo experiments with Google ASR

5 Conclusions

Thus our evo-devo inspired artificial development process helps repair ASR output of spoken sentences in a domain.

References

[Anantaram *et al.*, 2015] C. Anantaram, R. Gupta, N. Kini, and S. K. Kopparapu. Adapting general-purpose speech recognition engine output for domain-specific natural language question answering. *Workshop on Replicability and Reproducibility in Natural Language Processing: adaptive methods, resources and software,* IJCAI 2015.

[Fusayasu *et al.*,2015] Yohei Fusayasu, Katsuyuki Tanaka, Tetsuya Takiguchi, Yasuo Ariki. Word-Error Correction of Continuous Speech Recognition Based on Normalized Relevance Distance. *IJCAI 2015*, pages 1257-1262. 2015.

[Harding and Banzhaf, 2008] Simon Harding and Wolfgang Banzhaf. 'Artificial development', Organic Computing. Springer Berlin Heidelberg, 2008. 201-219.

[Morbini *et al.*, 2013] Fabrizio Morbini, Kartik Audhkhasi, Kenji Sagae, Ron Artstein, Doğan Can, Panayiotis Georgiou, Shri Narayanan, Anton Leuski, David Traum. Which ASR should I choose for my dialogue system? In *Proceedings of the SIGDIAL 2013 Conference*, 2013.

[Peng *et al.*, 2013] Fuchun Peng, Scott Roy, Ben Shahshahani, Francoise Beaufays. Search Results Based N-Best Hypothesis Rescoring with Maximum entropy classification. In *Proceedings of ASRU*, 2013.

[Tufte, 2009] Gunnar Tufte, "From Evo to EvoDevo: Mapping and Adaptation in Artificial Development," in *Evolutionary Computation*, edited by Wellington Pinheiro dos Santos, Chapter 12, October, 2009.

[Tur *et al.*, 2013] Gokhan Tur, Anoop Deoras, Dilek Hakkani Tur. Semantic Parsing Using Word Confusion Networks With Conditional Random Fields. *Interspeech* 2013.

[Twiefel *et al.*, 2014] Johannes Twiefel, Timo Baumann, Stefan Heinrich, StefanWermter. 'Improving Domain-Independent Cloud-Based Speech Recognition with Domain-Dependent Phonetic Post-Processing', Twenty-Eighth AAAI Conference, 2014.