Available online at www.sciencedirect.com
ScienceDirect
Computer Speech & Language xxx (2017) xxx-xxx
www.elsevier.com/locate/csl

Improving the understanding of spoken referring expressions through syntactic-semantic and contextual-phonetic error-correction☆

Q1 Ingrid Zukerman*, Andisheh Partovi

Faculty of Information Technology, Monash University, Clayton, Victoria 3800, Australia

Received 15 April 2016; received in revised form 28 February 2017; accepted 12 May 2017

Abstract

Despite recent advances in automatic speech recognition, one of the main stumbling blocks to the widespread adoption of Spoken Dialogue Systems is the lack of reliability of automatic speech recognizers. In this paper, we offer a two-tier error-correction process that harnesses syntactic, semantic and pragmatic information to improve the understanding of spoken referring expressions, specifically descriptions of objects in physical spaces. A syntactic-semantic tier offers generic corrections to perceived ASR errors on the basis of syntactic expectations of a semantic model, and passes the corrected texts to a language understanding system. The output of this system, which consists of pragmatic interpretations, is then refined by a contextual-phonetic tier, which prefers interpretations that are phonetically similar to the mis-heard words. Our results, obtained on a corpus of 341 referring expressions, show that syntactic-semantic error correction significantly improves interpretation performance, and contextual-phonetic refinements yield further improvements.

Keywords: Spoken language understanding; Referring expressions; Error correction; Pragmatic interpretation; Physical spaces; Syntactic-semantic model; Contextual-phonetic model

1. Introduction

In recent times, there have been significant improvements in Automatic Speech Recognition (ASR) (Pellegrini and Trancoso, 2010; Chorowski et al., 2015). Nonetheless, ASR errors and inconsistent performance across domains still impede the widespread adoption of Spoken Dialogue Systems (SDSs). For example, a research prototype of a spoken slot-filling dialogue system reported a Word Error Rate (WER) of 13.8% when using "a generic dictation ASR system" (Mesnil et al., 2015), and Google reported an 8% WER for its ASR API,[1] but this API had a WER of 54.6% when applied to the Let's Go corpus (Lange and Suendermann-Oeft, 2014). The accuracy of the commercial ASR employed in this research (Microsoft Speech SDK 6.1) falls between these numbers, with a WER of 30% for

☆ This paper has been recommended for acceptance by Prof. R. K. Moore.

* Corresponding author.
E-mail address: Ingrid.Zukerman@monash.edu (I. Zukerman), andi.partovi@monash.edu (A. Partovi).

[1] venturebeat.com/2015/05/28/google-says-its-speech-recognition-technology-now-has-only-an-8-word-error-rate.

http://dx.doi.org/10.1016/j.csl.2017.05.005
0885-2308/

descriptions of household objects. On one hand, these descriptions are more open-ended than the utterances employed in slot-filling applications, but on the other hand, they do not contain proper nouns, such as names of cities and foreign entities, which are rather error prone (Bulyko et al., 2005).

ASR errors not only produce mis-heard (wrongly recognized) entities or actions, but may also yield ungrammatical utterances that cannot be processed by subsequent interpretation modules of a Spoken Language Understanding (SLU) system (e.g., "the plate inside the microwave" being mis-heard as "*of plating sight* the microwave"), or yield incorrect results when processed by these modules (e.g., hesitations, often accompanied by fillers such as "hmm" or "ah", being mis-heard as "and" or "on") – all of which happened in our trials. Thus, further improvements are required in order to enable the widespread adoption of SDSs.

Two approaches for achieving such improvements are (1) enhancing ASR performance, e.g., through the use of deep neural networks (Hinton et al., 2012; Bahdanau et al., 2016) or by reducing the noise in the input signal (Maas et al., 2012); and (2) performing post-ASR error correction, e.g., (Alumäe and Kurimo, 2010; Béchet et al., 2014; Fusayasu et al., 2015). Clearly, perfect ASR would obviate the need for the latter approach, but we are not there yet, even with recent advances in ASRs based on deep neural networks (Tam et al., 2014). Further, Ringger and Allen (1996) argue that "even if the SR engine's language model can be updated with new domain-specific data, the post-processor trained on the *same* new data can provide additional improvements in accuracy". Hence, at present, post-ASR error correction is an active area of research (Section 9), which offers a practical way of de-coupling innovation cycles within target applications from ASR innovation cycles (Feld et al., 2012), and increases system portability (Ringger and Allen, 1996).

In this paper, we offer a mechanism for improving the understanding of spoken referring expressions, specifically descriptions of objects in physical spaces, by harnessing syntactic, semantic and pragmatic information after obtaining output from the ASR. Our mechanism consists of two main stages: syntactic-semantic error correction (Kim et al., 2013) and contextual-phonetic error-correction (Zukerman et al., 2015b).

The ***syntactic-semantic*** tier receives as input textual alternatives returned by the ASR. It first invokes a classifier that postulates wrong words in the ASR output (Zavareh et al., 2013) (Section 4) and a shallow semantic parser that breaks up these texts into semantically labelled segments (Kim et al., 2013) (Section 5). It then activates a syntactic-semantic error-correction model that proposes modifications for selected words in each text on the basis of the information provided by the word-error classifier and syntactic expectations of the semantic segments. The following modifications are considered during this stage: removal of noise (which is often due to filled pauses), insertion of missing prepositions, and replacement of mis-heard words. We consider two main types of replacements: (1) words or phrases expected to be closed class are replaced with phonetically similar closed-class words or phrases, e.g., "*for there a wave*" in prepositional phrase position is replaced with "further away"; and (2) words or phrases expected to be open class incur *generic* replacements, e.g., given a text such as "the *played* inside the microwave" (mis-heard by our ASR from the description "the plate inside the microwave" in the context of Fig. 1b[2]), "*played*" is replaced with the generic noun "thing", yielding "the thing inside the microwave". The rationale for this replacement is that "*played*" would lead a parser astray, while the proposed replacement allows an SLU system to proceed.

The resultant, possibly modified, texts are then given as input to the *Scusi?* SLU system (Zukerman et al., 2015a), which is a component of an SDS for human-robot interactions. *Scusi?* generates a ranked list of pragmatic interpretations for these texts, where an interpretation comprises candidate things in a physical space and the spatial relations between them, e.g., `plateD`-*location_in*-`microwave1`, and the ranking of an interpretation corresponds to the extent to which it matches the description (Section 2). In our example, the interpretation comprising Plate D is ranked first, as it is the only object inside the microwave. When the ASR made the same error for the description "the plate on the table" (which ambiguously refers to Plate E in Fig. 1b), yielding "the *played* on the table", the syntactic-semantic tier generated "the thing on the table". At this stage, this change offers no benefit in the context of Fig. 1b, as all the items are on the table. However, in the context of Fig. 1a, the three objects on the table are ranked equal first, ahead of the other objects in the scene.

The ***contextual-phonetic*** tier receives as input the top-*N* pragmatic interpretations produced by *Scusi?*, and re-ranks them according to the phonetic similarity between the mis-heard head nouns in the texts that led to each

[2] The ASR mis-heard "plate" as "*played*" or "*play*" in 404 out of the 1697 texts produced for referring expressions describing plates: items D and E in Fig. 1b, and items G, H and I in Fig. 1c.

Please cite this article as: I. Zukerman, A. Partovi, Improving the understanding of spoken referring expressions through syntactic-semantic and contextual-phonetic error-correction, Computer Speech & Language (2017), http://dx.doi.org/10.1016/j.csl.2017.05.005