Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 34 (2015) 154–185

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Employing distance-based semantics to interpret spoken referring expressions☆,☆☆

Ingrid Zukerman*, Su Nam Kim, Thomas Kleinbauer, Masud Moshtaghi

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

Received 26 May 2014; received in revised form 29 October 2014; accepted 12 January 2015
Available online 23 January 2015

Abstract

In this paper, we present *Scusi?*, an anytime numerical mechanism for the interpretation of spoken referring expressions. Our contributions are: (1) an anytime interpretation process that considers multiple alternatives at different interpretation stages (speech, syntax, semantics and pragmatics), which enables *Scusi?* to defer decisions to the end of the interpretation process; (2) a mechanism that combines scores associated with the output of the different interpretation stages, taking into account the uncertainty arising from a variety of sources, such as ambiguity or inaccuracy in a description, speech recognition errors and out-of-vocabulary terms; and (3) distance-based functions with probabilistic semantics that represent lexical similarity between objects' names and similarity between stated requirements and physical properties of objects (viz colour, size and positional relations). We considered two approaches for combining these descriptive attributes, viz multiplicative and additive, and determined whether prioritizing certain interpretation stages and descriptive attributes affects interpretation performance. We conducted two experiments to evaluate different aspects of *Scusi?*'s performance: Interpretive, where we compared *Scusi?*'s understanding of descriptions that are mainly ambiguous or inaccurate with people's understanding of these descriptions, and Generative, where we assessed *Scusi?*'s understanding of naturally occurring spoken descriptions. Our results show that *Scusi?*'s understanding of the descriptions in the Interpretive trial is comparable to that of people; and that its performance is encouraging when given arbitrary spoken descriptions in diverse scenarios, and excellent for the corresponding written descriptions. In both experiments, *Scusi?* significantly outperformed a baseline system that maintains only top same-score interpretations.

Keywords: Spoken language understanding; Numerical approach; Semantic interpretation; Distance-based semantics; Performance evaluation

1. Introduction

People often express themselves ambiguously or inaccurately (Trafton et al., 2005; Moratz and Tenbrink, 2006; Funakoshi et al., 2012). An ambiguous reference to an object matches several objects well, while an inaccurate reference

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

☆☆ This paper describes research conducted on the *Scusi?* system, amending, extending and adding detail to the work of Zukerman et al. (2008), Makalic et al. (2008) and Kleinbauer et al. (2013).

* Corresponding author. Tel.: +61 399055202.
E-mail addresses: Ingrid.Zukerman@monash.edu (I. Zukerman), Su.Kim@monash.edu (S.N. Kim), Thomas.Kleinbauer@monash.edu (T. Kleinbauer), Masud.Moshtaghi@monash.edu (M. Moshtaghi).

http://dx.doi.org/10.1016/j.csl.2015.01.002
0885-2308/

matches one or more objects partially. For instance, in a household domain, a reference to a "big blue mug" is ambiguous if there is more than one big blue mug in the room, and inaccurate if there are two mugs in the room, one big and red, and one small and blue. In addition, ambiguous or inaccurate references may result from different parse trees (e.g., due to variants in prepositional attachments), or from misheard utterances in spoken interactions. Computer systems that interact with people in natural language must be able to cope with such issues. This does not mean that a system must always obtain an intended interpretation of an utterance (although that would be nice), but when it misunderstands an utterance, the misunderstanding should be plausible.

Like Funakoshi et al. (2012), Lison and Kruijff (2009) and Ross (2010), we posit that simply considering the best interpretation of an utterance would not address these problems. In fact, our approach is part of a growing trend which harnesses numerical formalisms to consider multiple interpretations in order to handle the uncertainty inherent in real-world problems, e.g., (Funakoshi et al., 2012; Lison and Kruijff, 2009; Ross, 2010). However, we defer decisions to the end of the interpretation process, which like Punyakanok et al.'s (2008) approach, allows us to make global, rather than local, decisions. The score associated with an interpretation typically represents its goodness, which in turn enables the ranking of candidate interpretations. Additionally, scores may be used, in combination with utilities, to determine a course of action, e.g., ask a clarification question, or perform a requested action. Considering multiple interpretations, coupled with numerical scores, enables a system to recover from situations where the highest-scoring interpretation is not the intended one (e.g., due to speech recognition errors), or detect situations where several interpretations are similarly plausible.

In this paper, we present a numerical mechanism for the interpretation of spoken referring expressions which considers multiple interpretations at different levels (i.e., speech, syntax, semantics and pragmatics), and incorporates the physical reality of the context at the pragmatics level of the interpretation process. The disparate processes carried out at each level of interpretation require our mechanism to combine scores and probabilities obtained from several sources. Specifically, our mechanism combines scores returned by an Automatic Speech Recognizer (ASR), probabilities estimated by a probabilistic parser, scores that estimate the complexity of an interpretation, and scores that represent how well the properties of candidate objects (viz lexical item, colour, size and location) match a user's requirements (Section 4). The scores may be viewed as *subjective probabilities*, which represent one's state of certainty regarding the truth of a proposition (Pearl, 1988).

Our mechanism produces a ranked list of interpretations at each of the above levels, culminating in instantiated objects in a specific context. For example, consider the description "the large blue mug" uttered in a room which contains a large aqua mug, a small blue mug and a large red mug, denoted `mug0`, `mug1` and `mug2` respectively, among other objects. Our mechanism yields alternative interpretations where in principle the referent may be each of the objects in the room, and each interpretation is associated with a score. Interpretations comprising `mug0`, `mug1` or `mug2` as referents have a higher score than interpretations comprising other objects (which have a low score); and interpretations where the referent is a blue object have a higher score than interpretations where the referent has another colour (this helps in situations where the name of a referent has been misheard, but its colour was heard correctly). Since none of the mugs in our example match the description precisely, their ranking depends on how well the given requirements match the actual colour and size of the mugs, and on the relative importance of colour and size. These rankings and their scores support the determination of a course of action by a robotic agent. For instance, if the score of, say, the `mug0` interpretation is significantly higher than the scores of the other mugs, the agent can simply retrieve `mug0`. If the scores of the interpretations involving the three mugs are similar, the agent should ask a clarification question. However, if the description was uttered in a room with no mugs, the scores of all the candidate objects would be low, which may prompt the agent to look for a mug in a different room. The implementation of such a decision process is the next step in this project.

Our mechanism was evaluated in two experimental settings: an *Interpretive* web-based setting to determine how well our system's understanding of given written referring expressions matches people's understanding; and the more common *Generative* setting, e.g., (Gandrabur et al., 2006; Thomson et al., 2008; DeVault et al., 2009), to assess our system's performance in various scenarios where people give spoken descriptions of designated referents.

The contributions of this paper are:

- An anytime interpretation process that considers multiple alternatives at different interpretation stages (speech, syntax, semantics and pragmatics), which enables our system to defer decisions to the end of the interpretation process.