Anonymous ACL submission

Abstract

1 Introduction

* ASR is getting better all the time, but just being able to accurately recognize speech doesn't mean intent is understood. We have a way of showing intent understanding. * PAs come in a continuum of predictability, from very little "you say everything in one go" to fully "the system knows what you want before you even speak". We explore to what degree users want predictability in their systems. * PAs either fully act out a command or they fully misunderstand and start over. We offer something that works incrementally and signals understanding by verifying more fine-grained aspects of intent as it unfolds, allowing the user to observe the "internal state" of the intent understanding so corrections can be made quickly and easily * Who really knows what a PA can do? The only way to find out is to attempt to give it a command and see if it works (if not, then Google, for example, runs a web search). We show the user what the PA could potentially do * Dialogue systems are notoriously bad at signalling affordances

2 Related Work

3 System Description: DiaTree

This section introduces and describes our SDS. Our SDS is modularised into four main components: automatic speech recognition (ASR), natural langauge understanding (NLU), dialogue management (DM), and the user interface (UI) which, as explained below, is visualised as a right-branching tree. For the remainder of this section, each module is exlpained in turn. First, however, we explain what is meant by incremental processing and the role that plays in the work presented here.

//todo: figure of system modules//

3.1 Incremental Dialogue

Of prime importance in our SDS–the aspect of our SDS that sets it apart from others–is the requirement that it processes *incrementally*. An often-cited concern with incremental processing is regarding informativeness: why act so soon when waiting (even just for a moment) would allow additional information, resutling in more-informed decisions? The trade-off here is all-important: *naturalness* as perceived by the end user who is interacting with the SDS. Indeed, it has been shown that humans perceive incremental systems as being more natural than traditional, turn-based systems (Aist et al.(2006)Aist, Allen, Campana, Galescu, Gallo, Stoness, Swift, and Tanenhaus; Skantze and Schlangen(2009); Skantze and Hjalmarsson(1991); Asri et al.(2014)Asri, Laroche, Pietquin, and Khouzaimi), offer a more human-like experience for the human users (Edlund et al.(2008)Edlund, Gustafson, Heldner, and Hjalmarsson) and are more satisfying to interact with than non-incremental systems (Aist et al.(2007)Aist, Allen, Campana, Gallo, Stoness, and Swift). Psycholinguistic research has also shown that humans process (i.e., comprehend) utterances as they unfold and do not wait until the end of an utterance to begin the comprehension process (Tanenhaus(1995); Spivey et al.(2002)Spivey, Tanenhaus, Eberhard, and Sedivy).

The trade-off between informativeness and naturalness can be reconciled when mechanisms are in place where earlier desicions can be repaired. Such mechanisms were introduced in the incremental unit (IU) framework for SDS (Schlangen and Skantze(2009); Schlangen and Skantze(2011)). Following (Kennington et al.(2014)Kennington, Kousidis, and

Schlangen), SDSs based on the IU-network approach consist of a network of processing *modules*. A typical module takes input from its *left buffer*, performs some kind of processing on that data, and places the processed result onto its *right buffer*. The data are packaged as the payload of *incremental units* (IUs) which are passed between modules. The IUs themselves are also interconnected via *same level links* (SLL) and *grounded-in links* (GRIN), the former allowing the linking of IUs as a growing sequence, the latter allowing that sequence to convey what IUs directly affect it. A complication particular to incremental processing is that modules can "change their mind" about what the best hypothesis is, in light of later information, thus IUs can be *added*, *revoked*, or *committed* to a network of IUs.

The modules exlpained in the remained of this section are implemented as IU-modules and process incrementally. Each will now be explained.

3.2 Speech Recognition

Incremental processing begins with modules that take in input; in the case of our SDS, that is the ASR component. Incremental ASR must transcribe uttered speech into words and words must be forthcoming from the ASR as early as possible (i.e., the ASR must not wait for endponiting in order to act). Each module that follows must also process incrementally, acting in lock-step upon input as it is received. Incremental ASR is not new (Baumann et al.(2009)Baumann, Atterer, and Schlangen) and many of the current freely-accessible ASR systems can produce output (semi-) incrementally.

In our SDS, we opt for Google ASR because of its wide vocbaulary coverage of the language we are interested in (German). We are able to package ASR output from the Google service into IUs as explained above. Those word IUs are passed to the NLU module, which will now be explained.

3.3 Language Understanding

We approach the task of NLU as a slot-filling task (a very common approach; see (Tur et al.(2012)Tur, Deng, Hakkani-Tür, and He)) where the system can fill the task when all slots of a frame are filled. The main driver of the NLU in our SDS is the SIUM model of NLU introduced in (Kennington et al.(2013)Kennington, Kousidis, and Schlangen). Though originally a model of reference resolution, the authors hinted that it could be used for general NLU, which we do here. The model is formalised as follows:

$$P(I|U) = \frac{1}{P(U)} P(I) \sum_{r \in R} P(U|R) P(R|I) \quad (1)$$

That is, $P(I|U)$ is the probability of the intent I (i.e., a frame slot) behind the speaker's (ongoing) utterance U. This is recovered using the mediating variable R, a set of *properties* which map between aspects of U and aspects of I. These properties could be visual properties of visible objects, or they could be more abstract properties that intents might have, which we opt for here (e.g., the intent of a `destination` might be filled by `new york` which has (among others) properties like `new york`, `biggest-us-city`, `has-statue-of-libery`, etc.). Properties are pre-defined by a system designer and can match words that might be uttered to describe the intent in question. The mapping betwen properties and aspects of U can be learned from data. During application, $P(U|R)$ can produce a distribution over words (or properties, see below) which are summed over and the probability mass for each property is accumulated for each intent, resulting in a distribution over possible intents. This occurs at each word increment, where the distribution from the previous increment is combined via $P(I)$, keeping track of the distribution over time.

In our SDS, we apply an instantion of SIUM for each slot (explained in Section 4), all of which update at each word increment. At each word increment, the updated slots (and their corresponding) distributions are given to the DM, which will now be explained.

3.4 Dialogue Manager

//todo: explain opendial and how it handles four states and CRs//

3.5 User Display

//todo: show examples of the tree and how it portrays the current state//

4 Experiments

4.1 Experiment 1: Non-Incremental vs. Incremental

4.2 Results

4.3 Experiment 2: Incremental vs. Incremental-Adaptive

4.4 Results

References

Gregory Aist, James Allen, Ellen Campana, Lucian Galescu, Carlos A. Gomez Gallo, Scott Stoness, Mary Swift, and Michael Tanenhaus. Software architectures for incremental understanding of human speech. In *Interspeech 2006*, pages 1922—-1925, 2006.

Gregory Aist, James Allen, Ellen Campana, Carlos Gomez Gallo, Scott Stoness, and Mary Swift. Incremental understanding in human-computer dialogue and experimental evidence for advantages over nonincremental methods. In *Pragmatics*, volume 1, pages 149–154, Trento, Italy, 2007.

Layla El Asri, Romain Laroche, Olivier Pietquin, and Hatim Khouzaimi. NASTIA: Negotiating Appointment Setting Interface. In *Proceedings of LREC*, pages 266–271, 2014.

Timo Baumann, Michaela Atterer, and David Schlangen. Assessing and Improving the Performance of Speech Recognition for Incremental Systems. In *Proceedings of NAACL-HLT 2009*, Boulder, USA, June 2009.

Jens Edlund, Joakim Gustafson, Mattias Heldner, and Anna Hjalmarsson. Towards human-like spoken dialogue systems. *Speech Communication*, 50(8-9): 630–645, 2008.

Casey Kennington, Spyros Kousidis, and David Schlangen. Interpreting Situated Dialogue Utterances: an Update Model that Uses Speech, Gaze, and Gesture Information. In *Proceedings of SIGdial*, 2013.

Casey Kennington, Spyros Kousidis, and David Schlangen. InproTKs: A Toolkit for Incremental Situated Processing. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 84–88, Philadelphia, PA, U.S.A., 2014. Association for Computational Linguistics.

David Schlangen and Gabriel Skantze. A general, abstract model of incremental dialogue processing. *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics on EACL 09*, 2(1):710–718, 2009.

David Schlangen and Gabriel Skantze. A General, Abstract Model of Incremental Dialogue Processing. In *Dialogue & Discourse*, volume 2, pages 83–111, 2011.

Gabriel Skantze and Anna Hjalmarsson. Towards Incremental Speech Production in Dialogue Systems. In *Word Journal Of The International Linguistic Association*, pages 1–8, Tokyo, Japan, September 1991.

Gabriel Skantze and David Schlangen. Incremental dialogue processing in a micro-domain. *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics on EACL 09*, (April):745–753, 2009.

Michael J Spivey, Michael K Tanenhaus, Kathleen M Eberhard, and Julie C Sedivy. Eye movements and spoken language comprehension: Effects of visual context on syntactic ambiguity resolution. *Cognitive Psychology*, 45(4):447–481, 2002.

Michael Tanenhaus. Integration of Visual and Linguistic Information in Spoken Language Comprehension. *Science*, 268:1632–1634, 1995.

Gokhan Tur, Li Deng, Dilek Hakkani-Tür, and Xiaodong He. Towards deeper understanding: Deep convex networks for semantic utterance classification. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pages 5045–5048. IEEE, 2012.