$C^2D^2E^2$: Using Call Centers to Motivate the Use of Dialog and Diarization in Entity Extraction

Kenneth Church, Weizhong Zhu and **Jason Pelecanos**
IBM, Yorktown Heights, NY, USA
{kwchurch, zhuwe, jwpeleca}@us.ibm.com

Abstract

This paper introduces a deceptively simple entity extraction task intended to encourage more interdisciplinary collaboration between fields that don't normally work together: diarization, dialog and entity extraction. Given a corpus of 1.4M call center calls, extract mentions of trouble ticket numbers. The task is challenging because first mentions need to be distinguished from confirmations to avoid undesirable repetitions. It is common for agents to say part of the ticket number, and customers confirm with a repetition. There are opportunities for dialog (given/new) and diarization (who said what) to help remove repetitions. New information is spoken slowly by one side of a conversation; confirmations are spoken more quickly by the other side of the conversation.

1 Extracting Ticket Numbers

Much has been written on extracting entities from text (Etzioni et al., 2005), and even speech (Kubala et al., 1998), but less has been written in the context of dialog (Clark and Haviland, 1977) and diarization (Tranter and Reynolds, 2006; Anguera et al., 2012; Shum, 2011). This paper describes a ticket extraction task illustrated in Table 1. The challenge is to extract a 7 byte ticket number, "902MDYK," from the dialog. Confirmations ought to improve communication, but steps need to be taken to avoid undesirable repetition in extracted entities. Dialog theory suggests it should be possible to distinguish first mentions (**bold**) from confirmations (*italics*) based on prosodic cues such as pitch, energy and duration.

t0	t1	S1	S2
278.16	281.07	I do have the new hardware case number for you when you're ready	
282.60	282.85		*okay*
284.19	284.80	**nine**	
285.03	285.86	**zero**	
286.22	286.74	**two**	
290.82	291.30		*nine*
292.87	293.95	*zero two*	
297.87	298.24		*okay*
299.30	300.49	**M. as in Mike**	
301.97	303.56	**D. as in delta**	
304.89	306.31	**Y. as in Yankee**	
307.50	308.81	**K. as in kilo**	
310.14	310.57		*okay*
310.77	311.70		*nine zero two*
311.73	312.49		*M. D.*
312.53	313.18		*Y. T.*
313.75	314.21	*correct*	
314.21	317.28	and thank you for calling IBM is there anything else I can assist you with	

Table 1: A ticket dialog: 7 bytes (902MDYK) at 1.4 bps. First mentions (**bold**) are slower than confirmations (*italics*).

phone matches	calls	ticket matches (edit dist)
66%	238	0
59%	82	1
55%	40	2
4.1%	4033	3+

Table 2: Phone numbers are used to confirm ticket matches. Good ticket matches (top row) are confirmed more often than poor matches (bottom row). Poor matches are more common because ticket numbers are relatively rare, and most calls don't mention them.

Proceedings of EMNLP 2016 Workshop on Uphill Battles in Language Processing: Scaling Early Achievements to Robust Methods, pages 35–38,
Austin, TX, November 5, 2016.

In Table 1, "zero two" was 55% slower the first time than the second (1.7 vs. 1.1 seconds).

Much of Table 1 was produced by machine, using tools that are currently available for public use, or will be available soon. Words came from ASR (automatic speech recognition) and speaker labels (S1 and S2) from diarization.[1] We plan to label **bold** and *italics* automatically, but for now, that was done by hand.

It is remarkable how hard it is to transmit ticket numbers. In this case, it takes 39 seconds to transmit 7 bytes, "902MDYK," a mere 1.4 bps (bits per second).[2] Agents are well aware of the difficulty of the task. In Table 1, the agent says the first three digits slowly in citation form (more like isolated digits than continuous speech) (Moon, 1991). Citation form should be helpful, though in practice, ASR is trained on continuous speech, and consequently struggles with citation form.

After a few back-and-forth confirmations, the customer confirms the first three digits with a backchannel (Ward and Tsukahara, 2000) "okay," enabling the agent to continue transmitting the last four bytes, "MDYK," slowly at a byte/sec or less, using a combination of military and conventional spelling: in Mike," "D. as in delta," etc. When we discuss Figure 1, we will refer to this strategy as *slow mode*. If the agent was speaking to another agent, she would say, "Mike delta Yankee kilo," quickly with no intervening silences. We will refer to this strategy as *fast mode*.

Finally, the customer ends the exchange with another backchannel "okay," followed by a quick repetition of all 7 bytes. Again we see that first mentions (**bold**) take more time than subsequent mentions (*italics*). In Table 1, the **bold** first mention of "902MDYK" takes $12.1 = 286.74 - 284.19 + 308.81 - 299.30$ seconds, which is considerably longer than the customer's confirmation in *italics*: $2.4 = 313.18 - 310.77$ seconds.

Ticket numbers are also hard for machines. ASR errors don't help. For example, the final "K" in the final repetition was misheard by the machine as "T."

[1]The ASR tools are currently available for public use at: https://www.ibm.com/watson/developercloud/text-to-speech.html, and diarization will be released soon.

[2]The estimate of 1.4 bps would be even worse if we included opportunities for compressing tickets to less than 7 bytes.

t0	transcript
344.01	and I do have a hardware case number whenever you're ready for it
348.86	hang on just one moment
353.65	okay go ahead that will be Alfa zero nine
358.18	the number two
359.85	golf Victor Juliet
363.55	I'm sorry what was after golf
366.46	golf and then V. as in Victor J. as in Juliet
370.28	okay
371.86	Alfa zero niner two golf Victor Juliet that is correct Sir you can't do anything else for today

Table 3: An example with a retransmission: 1.7 bits per second to transmit "A082GVJ"

After listening to the audio, it isn't clear if a human could get this right without context because the customer is speaking quickly with an accent. Nevertheless, the confirmation, "correct," makes it clear that the agent believes the dialog was successfully concluded and there is no need for additional confirmations/corrections. Although it is tempting to work on ASR errors forever, we believe there are bigger opportunities for dialog and diarization.

2 Communication Speed

The corpus can be used to measure factors that impact communication speed: given/new, familiarity, shared conventions, dialects, experience, corrections, etc. In Table 1, first mentions are slower than subsequent mentions. Disfluencies (Hindle, 1983) and corrections ("I'm sorry what was after golf") take even more time, as illustrated in Table 3.

Figure 1 shows that familiar phone numbers are quicker than less familiar ticket numbers, especially in slow mode, where each letter is expressed as a separate intonation phrase. Agents speed up when talking to other agents, and slow down for customers, especially when customers need more time. Agents have more experience than customers and are therefore faster.

Agents tend to use slow mode when speaking with customers, especially the first time they say the ticket number. Table 1 showed an example of slow mode. Fast mode tends to be used for confirmations, or when agents are speaking with other agents. Figure 1 shows that fast mode is faster than slow mode, as one would expect.

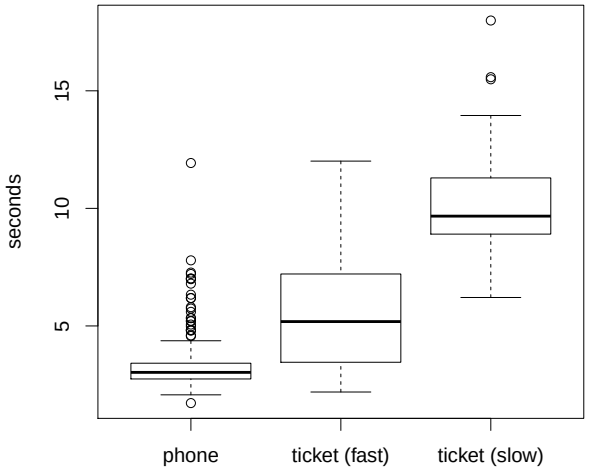

Figure 1: Time to say phone numbers and tickets, computed over a sample of 552 simple/robust matches. The plot shows that phone numbers are faster than ticket numbers. Ticket numbers are typically spoken in one of two ways which we call *fast mode* and *slow mode*. The plot shows that fast mode is faster than slow mode, as one would expect.

Figure 1 gives an optimistic lower bound view of times. The figure was computed over a small sample of 552 calls where simple (robust) matching methods were sufficient to find the endpoints of the match in the audio. Tables 1 and 3 demonstrate that total times tend to be much longer because of factors not included in Figure 1 such as prompts, confirmations and retransmissions.

Shared knowledge helps. Phone numbers are quicker than tickets because everyone knows their own phone number. In addition, everyone knows that phone numbers are typically 10 digits, parsed: $3+3+4$. Communication slows down when phone numbers are expressed in unfamiliar ways such as "double nine" and "triple zero," common in Indian English and Australian English, but not American English.

3 Materials

We are working with a call center corpus of 1.4M calls. Table 4 shows call duration by number of speakers. The average call is 5.6 minutes, but most calls are shorter than average, and a few calls are much longer than average. The 50th, 95th and 99th percentiles are 4, 15 and 31 minutes, respectively. The longer calls are likely to involve one or more transfers, and therefore, longer calls tend to have more speakers.

Speakers	Calls	Seconds/Call
0	565	20
1	405	61
2	5021	342
3	837	533
4	107	986
5	22	1121
6+	13	1166

Table 4: Most of our calls have two speakers, a customer and an agent, though some have more speakers and some have less. The duration of the call tends to increase with the number of speakers. These counts were computed from a relatively small sample of nearly 7k calls that were manually transcribed.

A relatively small sample of almost 7k calls was transcribed by a human transcription service, mainly to measure WER (word error rates) for recognition, but can also measure diarization errors. Unfortunately, ground truth is hard to come by for entity extraction because we didn't ask the service to extract phone numbers and tickets.

Heuristics are introduced to overcome this deficiency. The first 4-5 bytes of the ticket are predictable from side information (timestamps), not available to the dialog participants. Edit distance is used to match the rest with tickets in a database. Matches are confirmed by comparing phone numbers in the database with phone numbers extracted from the audio. Table 2 shows good ticket matches (top row) are confirmed more often than poor matches (bottom row).[3] Given these confirmed matches, future work will label **bold** and *italics* automatically. An annotated corpus of this kind will motivate future work on the use of dialog and diarization in entity extraction.

[3]The phone matching heuristic is imperfect in a couple of ways. The top row is far from 100% because the customer may use a different phone number than what is in the database. The bottom row contains most of the calls because the entities of interest are quite rare and do not appear in most calls.

4 Conclusions

This paper introduced a deceptively simple entity extraction task intended to encourage more interdisciplinary collaboration between fields that don't normally work together: diarization, dialog and entity extraction. First mentions need to be distinguished from confirmations to avoid undesirable repetition in extracted entities. Dialog theory suggests the use of prosodic cues to distinguish marked first mentions from unmarked subsequent mentions. We saw in Table 1 that first mentions (**bold**) tend to be slower than subsequent confirmations (*italics*).

It also helps to determine who said what (diarization), because new information tends to come from one side of a conversation, and confirmations from the other side. While our corpus of 1.4M calls cannot be shared for obvious privacy concerns, the ASR and diarization tools are currently available for public use (or will be available soon). While much has been written on given/new, this corpus-based approach should help establish more precise numerical conclusions in future work.

The corpus can be used to measure a number of additional factors beyond given/new that impact communication speed: familiarity, shared conventions, dialects, experience, corrections, etc. Table 3 shows an example of corrections taking even more time ("I'm sorry what was after golf"). Figure 1 shows that familiar phone numbers are quicker than less familiar ticket numbers, especially in slow mode, where each letter is expressed as a separate intonation phrase. Agents speed up when talking to other agents, and slow down for customers, especially when customers need more time. Agents have more experience than customers and are therefore faster.

References

Xavier Anguera, Simon Bozonnet, Nicholas Evans, Corinne Fredouille, Gerald Friedland, and Oriol Vinyals. 2012. Speaker diarization: A review of recent research. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(2):356–370.

Herbert H Clark and Susan E Haviland. 1977. Comprehension and the given-new contract. *Discourse production and comprehension. Discourse processes: Advances in research and theory*, 1:1–40.

Oren Etzioni, Michael Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S Weld, and Alexander Yates. 2005. Unsupervised named-entity extraction from the web: An experimental study. *Artificial intelligence*, 165(1):91–134.

Donald Hindle. 1983. Deterministic parsing of syntactic non-fluencies. In *Proceedings of the 21st annual meeting on Association for Computational Linguistics*, pages 123–128. Association for Computational Linguistics.

Francis Kubala, Richard Schwartz, Rebecca Stone, and Ralph Weischedel. 1998. Named entity extraction from speech. In *Proceedings of DARPA Broadcast News Transcription and Understanding Workshop*, pages 287–292. Citeseer.

Seung-Jae Moon. 1991. An acoustic and perceptual study of undershoot in clear and citation-form speech. *Phonetic Experimental Research at the Institute of Linguistics University of Stockholm XIV, University of Stockholm, Institute of Linguistics*, pages 153–156.

Stephen Shum. 2011. *Unsupervised methods for speaker diarization*. Ph.D. thesis, Massachusetts Institute of Technology.

Sue E Tranter and Douglas A Reynolds. 2006. An overview of automatic speaker diarization systems. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1557–1565.

Nigel Ward and Wataru Tsukahara. 2000. Prosodic features which cue back-channel responses in english and japanese. *Journal of pragmatics*, 32(8):1177–1207.