Special Topic Presentation: Incremental Processing

Rebecca Myhre

What and Why?

- Most spoken dialogue systems wait for user to stop speaking before processing input and deciding how to react.
- Incremental processing uses results from partial phrase speech recognition to inform system decisions.
- Using incremental results can make system more responsive, but main motivation is to allow dialogue system to more closely mimic human conversation.
  - Allows for interruptions, overlapping dialogue, sentence completion, back-channeling, etc.

Issues, Open Questions

- There are a lot of partial results; which ones do you use?
- How do you deal with the instability and inaccuracy of partial ASR results?
- Where can incremental processing be best applied?

Ethan Selfridge, Iker Arizmendi, Peter Heeman, and Jason Williams. (2011). Stability and Accuracy in Incremental Speech Recognition. In *Proceedings of the 12th Annual SigDial Meeting on Discourse and Dialogue*, Portland, Oregon.

Overview

- Goal: devise method to identify stable and accurate partial phrase results for system to use.
- Approach: think about decoding process.
- Three types of partial results are defined:
  - Basic – most likely path through partially decoded Viterbi lattice.
  - Terminal – most likely path ends at a terminal node.
  - Immortal – all paths come together at a single, "immortal" node. This partial result is stable and *will* be the final ASR output for this span, whether or not it is accurate.

Data, Models

- Dataset: utterances from calls to CMU's "Let's Go!" system.
- Three LMs: two rule-based, one statistical:
  - RLM1 = street, neighborhood names from bus timetable database
  - RLM2 = neighborhood names
  - SLM = trigram model
- Tested on different sets; RLM test sets were designed to be 80% in-grammar.

	RLM1	RLM2	SLM
Num. Utts All	7722	5411	42620
Num. Utts MW	3213	1748	20396
Words/Utt All	1.7	1.5	2.3
Words/Utt MW	2.8	2.6	3.8
Utt. Acc. All.	50 %	60 %	62 %
Utt. Acc. MW	53 %	56 %	44 %

Frequency, Stability, and Accuracy

- Stability compares partial ASR result to final ASR result.
- Accuracy compares partial ASR result to transcription.
- Immortal > Terminal > Basic

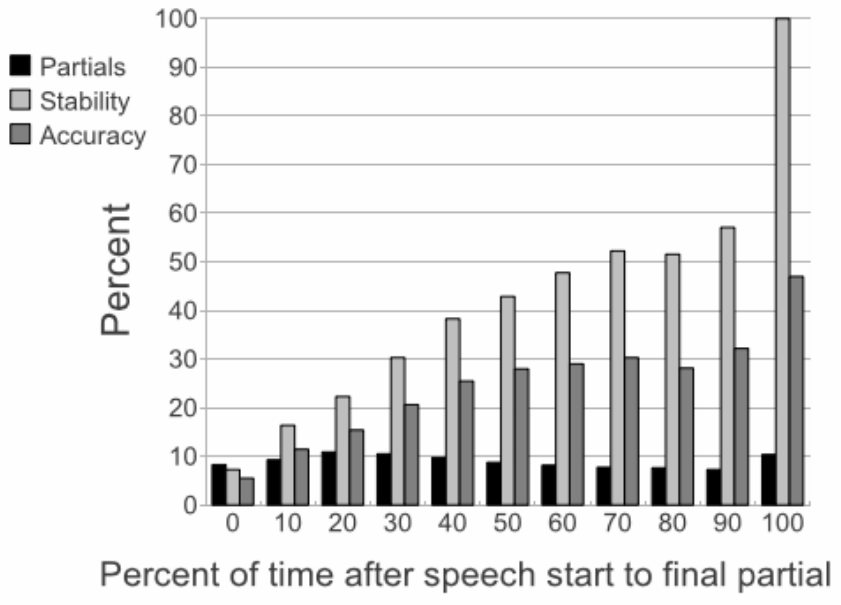
ISR	Group	RLM1	RLM2	SLM
Basic	All	12.0	9.9	11.6
	MW	14.6	12.3	29.7
Terminal	All	5.4	3.3	6.2
	MW	6.4	4.1	8.8
Immortal	All	0.22	0.32	0.55
	MW	0.42	0.67	0.63

ISR	Group	RLM1	RLM2	SLM
Stability				
Basic	All	10 %	11 %	7 %
	MW	14 %	15 %	9 %
Terminal	All	23 %	31 %	37 %
	MW	20 %	28 %	36 %
Accuracy				
Basic	All	9 %	1 %	5 %
	MW	11 %	13 %	6 %
Terminal	All	13 %	21 %	24 %
	MW	12 %	17 %	21 %
Immortal	All	91 %	93 %	55 %
	MW	90 %	90 %	56 %

Hybrid Approach: LAISR
(Lattice-Aware Incremental Speech Recognition)

- Recognizes both Terminal and Immortal results; checks for Immortal result first, then backs off to Terminal result.
- Produces a steady stream of partials with better (although not great) stability and accuracy.

Partials per Utterance			
	RLM1	RLM2	SLM
All	5.6	3.5	6.7
MW	6.7	4.5	9.6
Stability Percentage			
All	24 %	33 %	40 %
MW	24 %	35 %	41 %
Accuracy Percentage			
All	15 %	23 %	26 %
MW	16 %	22 %	24 %



Stability and Confidence Measures

- They built Stability Measure and Confidence Measure classifiers, trained with logistic regression, for Basic ISR, Terminal ISR, and LAISR.
- Features used for all three ISRs:
  - Raw Watson confidence score, features that affect the confidence score, normalized cost, normalized speech likelihood, likelihoods of competing models, best path score in word confusion network (WCN), length of path in WCN, worst probability in WCN, and length of N-best list.
- For LAISR, additional features:
  - Three binary indicators of whether partial is Terminal, Immortal, or Terminal following an Immortal, and the percentage of words in the hypothesis which are immortal.

Results

		All			Multi-Word		
Stability Measure (SM) Equal Error Rate							
		RLM 1	RLM 2	SLM	RLM 1	RLM 2	SLM
Basic	WATSON Score	13.3	13.3	12.8	15.6	16.4	15.2
	Regression	**10.7**	**11.3**	**12.3**	**13.2**	**15.2**	15.1
Terminal	WATSON Score	24.3	29.1	34.4	26.6	26.0	34.1
	Regression	**19.7**	**26.5**	**26.5**	**23.0**	24.3	**24.7**
LAISR	WATSON Score	24.7	29.3	35.0	24.0	27.0	35.3
	Regression	**19.2**	**25.6**	**25.0**	**18.4**	**23.3**	**22.7**
Confidence Measure (CM) Equal Error Rate							
Basic	WATSON Score	11.3	11.7	9.9	14.1	14.0	11.6
	Regression	**9.8**	**9.8**	9.7	**12.3**	**12.9**	**11.0**
Terminal	WATSON Score	15.1	21.1	30.6	15.7	17.4	29.3
	Regression	**11.7**	**16.8**	**20.8**	**12.1**	**14.5**	**18.4**
LAISR	WATSON Score	15.8	21.8	32.3	18.4	19.5	31.8
	Regression	**11.6**	**16.6**	**21.0**	**11.6**	**14.2**	**18.7**

Conclusions

- LAISR's hybrid approach addresses the problem that many partials are unstable.
- LAISR outperforms Terminal ISR, especially for multi-word utterances.
- Can produce better stability and confidence scores that raw recognition score.
- Possible applications:
  - News broadcast transcription
  - More flexible SDS that can interrupt user (for instance, if input so far is likely to be stable and inaccurate)
  - Develop intention-level stability and accuracy measures

Kenji Sagae, Gwen Christian, David DeVault, and David Traum. (2009). Towards Natural Language Understanding of Partial Speech Recognition Results in Dialogue Systems. In *Proceedings of HLT-NAACL*.

David DeVault, Kenji Sagae, and David Traum. (2009). Can I finish? Learning when to respond to incremental interpretation results in interactive dialogue. In *The 10th Annual SIGDIAL Meeting on Discourse and Dialogue (SIGDIAL 2009)*, London, UK.

Overview

- Ultimate goal:

  Incorporate partial ASR results into NLU module to enable an agent that could initiate overlapping speech and complete utterances (a common event in human dialogue)

- Dataset: a corpus of utterances said by people playing the role of the captain in a negotiation scenario:

  User (Army captain) negotiates with the head of an NGO clinic and a local village elder to relocate a medical clinic from the marketplace somewhere else, ideally the US military base.

- System has to be robust to high out-of-vocabulary and word error rates.

- Handles this in part because it targets utterance meaning.

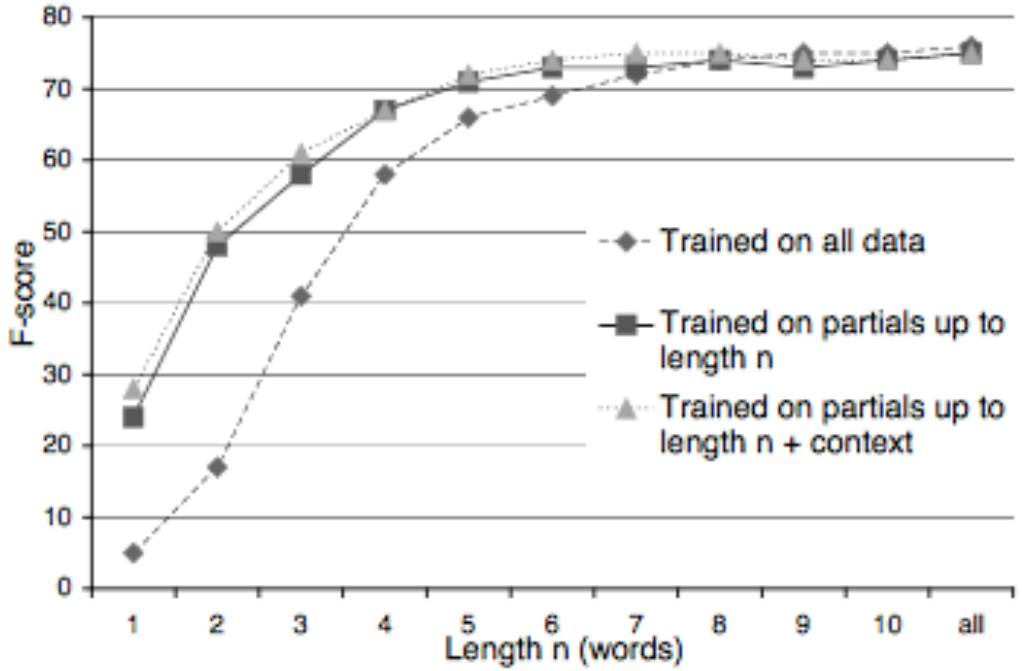
NLU module

- Maximum entropy classifier (mxNLU) trains the NLU module.
- ASR output is used as features: bag of words, bigrams, pairs of every two words in the input, number of words in input string
- Training set has 3,500 utterances and 136 unique frames, including 1 garbage frame (15% of utterances).
- Evaluate precision and recall at the level of attribute-value pairs outputted by the classifier:

  Precision = 0.78, Recall = 0.74, F-score = 0.76

Now with Incremental Processing

- Obtained partial ASR results for all utterances, then trained classifiers – 10 different models for utterances of different lengths (judged by number of words)
- Want to identify strategic points at which interpretation is not likely to significantly improve later in the sentence:



Identifying Viable Partial Results

- Second classifier, MAXF, is trained to learn when a partial ASR result is likely to have achieved an NLU F-score at least as high as if the entire utterance had been completed.
- Features:
  - K = number of partial results that have been received
  - N = length (word count) or current partial utterance
  - Entropy of probability distribution assigned to alternative output frames (low entropy = more focused distribution)
  - P_{max} = probability of most likely output frame
  - NLU = most probable output frame
- Label = MAXF(GOLD)
  - Boolean: F score of partial result ≥ F-score of final utterance
- Trained with a decision tree, 10-fold cross-validation evaluation
  - Precision over Recall

Intrinsic Evaluation

- Evaluated several different aspects of the model:
  - K_{MAXF}: first partial for which MAXF = TRUE
  - MAXF classifier output (TRUE or FALSE)
  - $\Delta F(K)$: loss associated with using partial utterance rather than complete utterance
  - T(K): remaining length (seconds) in the user utterance
- Results:
  - K_{MAXF} found in 79.2% of utterances
  - mean $T(K_{MAXF})$ is 1.6 seconds (if K_{MAXF} is found)
  - $\Delta F(K_{MAXF}) = 0$ 62.35% of the time
    $= -1$ 10.67% of the time
    $= 1$ 2.52% of the time

Extrinsic Evaluation

Prototype implementation of utterance completion:

Partial utterance: *we need to*
Predicted completion: *move your clinic*
Actual completion: *move the clinic*

Partial utterance: *I have orders*
Predicted completion: *to move you and this clinic*
Actual completion: *to help you in moving the clinic to a new location*

Partial utterance: *the market*
Predicted completion: *is not safe*
Actual completion: *is not a safer location*

Partial utterance: *we can also*
Predicted completion: *give you medical supplies*
Actual completion: *build you a well*

Discussion TIme

Thoughts, Discussion

- All papers recognize that *some* method of judging whether incremental results are useable is necessary.
- Focus on application of incremental results towards NLU rather than ASR appears to be a way to remain robust to some instability.
- These concepts are implementable, as (Sagae et al., 2009) and (DeVault et al., 2009), in particular, demonstrate.
- Would have been interesting to see oracle results using manually transcribed data– how much of error is attributable to ASR?
- What are your impressions of these approaches and techniques? Where do you think incremental processing can be best leveraged? Are there other ways incremental processing can be used that haven't been mentioned?

References

Ethan Selfridge, Iker Arizmendi, Peter Heeman, and Jason Williams. (2011). Stability and Accuracy in Incremental Speech Recognition. In *Proceedings of the 12th Annual SigDial Meeting on Discourse and Dialogue*, Portland, Oregon.

Kenji Sagae, Gwen Christian, David DeVault, and David Traum. (2009). Towards Natural Language Understanding of Partial Speech Recognition Results in Dialogue Systems. In *Proceedings of HLT-NAACL*.

David DeVault, Kenji Sagae, and David Traum. (2009). Can I finish? Learning when to respond to incremental interpretation results in interactive dialogue. In *The 10th Annual SIGDIAL Meeting on Discourse and Dialogue (SIGDIAL 2009)*, London, UK.