Evaluating the Potential Utility of ASR N-Best Lists for Incremental Spoken Dialogue Systems

Timo Baumann, Okko Buß, Michaela Atterer, David Schlangen

Department of Linguistics, University of Potsdam, Germany

Abstract

The potential of using ASR n-best lists for dialogue systems has often been recognised (if less often realised): it is often the case that even when the top-ranked hypothesis is erroneous, a better one can be found at a lower rank. In this paper, we describe metrics for evaluating whether the same potential carries over to *incremental* dialogue systems, where ASR output is consumed and reacted upon while speech is still ongoing. We show that even small N can provide an advantage for semantic processing, at a cost of a computational overhead.

Index Terms: dialogue systems, speech recognition, natural language understanding, incrementality

1. Introduction

By design, modern speech recognisers pursue all hypotheses about the input signal in one pass which are internally ranked for their quality [1]. Not always does this ranking reflect the right quality criteria, however, and in practice it does happen that more appropriate hypotheses are lower ranked (see e. g. [2]). It has often been tried to make use of these *n-best* lists in Spoken Dialogue Systems, under the assumption that contextual information can help to identify more appropriate candidate utterances (see Section 2 below).

In this paper we evaluate whether n-best lists could also be of use in *incremental* dialogue systems that process input while the speaker is still producing her utterance and that hence work with partial information from the ASR. We develop metrics for measuring this utility, looking both at the objective quality of the hypothesis as well as its utility for further semantic processing. More precisely, we measure whether temporal benefits (can hypotheses be established sooner?) and accuracy benefits (can correct hypotheses be established more often?) can be realised through the use of n-best lists, and how a decision for or against use of n-best lists can be made based on computational and accuracy trade-offs.[1]

The metrics developed here only explain how to measure whether using n-best lists in an incremental setting—given a particular ASR system—could *in principle* be advantageous. We leave the next steps, showing how to make use of such lists in practice and investigating to what extend techniques from the non-incremental case can be transferred, to future work.

In the remainder of the paper we cover related work in Section 2, describe measures for incremental n-best processing in Section 3, describe our setup in Section 4, and present our results in Section 5. We close with a general discussion and conclusions in Sections 6 and 7.

[1]The work reported here is an extension of [3], which deals with incremental one-best hypotheses only. See below for the substantive adaptation that were needed to cover n-best and semantic utility as well.

2. Related work

For the non-incremental case it has often been shown that there can be useful information at lower ranked positions in the n-best list produced by a speech recogniser for a given utterance. An often-tried method to get at this useful information is to use "higher-level" features of various sorts to re-rank the hypotheses, e. g. [2] use intra-utterance linguistic features to re-order the list; [4] use a limited form of dialogue context to judge the contextual plausibility of hypotheses; [5] additionally use pragmatic plausibility information to re-order and classify as "accept" / "reject" hypotheses in the list; [6] finally add information from different analysis-levels and target domains to the re-ordering process. More recently, statistical methods have been developed that can treat the whole n-best list as a belief distribution over observations, foregoing explicit re-ordering [7, 8].

To our knowledge, there is very little work on n-best lists in *incremental* speech recognition. [9] present an extension to a method for incremental NLU [10] to use n-best lists and show some improvement (8.49%); it is difficult though for us to evaluate these claims as the paper is only available in Japanese.[2]

In earlier work, we presented measures for capturing incremental performance of ASR systems, but only for the one-best case [3]. We focused on aspects such as the correctness of partial hypotheses, how quickly words are recognized, and how many intermittent word-hypotheses have to be withdrawn, and demonstrated a trade-off between the measures as well as that simple post-processing improves many of their measures. We extend this here to n-best hypothesis lists, and also generalise the evaluation metrics by using a hand-transcribed gold standard (instead of the final 1-best result); this allows us to evaluate the influence of using n-bests lists on the performance of the NLU module as well.

3. Measures for incremental n-best hypotheses

3.1. ASR measures

Following [3], we denote with w_{gold_t} the words in the gold transcription up to time t and similarly with $w_{hyp_t^N}$ the words in a hypothesis at time t, here additionally indexed with the rank in the n-best list (N). Instead of only evaluating whether a hypothesis at time t is "relatively correct" (i. e., it accords to the final best ASR hypothesis up to time t when processing has completed, as in [3]), we calculate the incremental word error rate at time t as the ordinary word error rate of the N'th-ranked hypothesis at time t and the gold standard up to this point: $WER_t^N = WER(w_{hyp_t^N}, w_{gold_t})$.

We define the (anti)-*oracle* WER as that of the best (worst) hypothesis among those in the n-best list:

[2]Our summary is based on the English abstract of the Japanese paper.

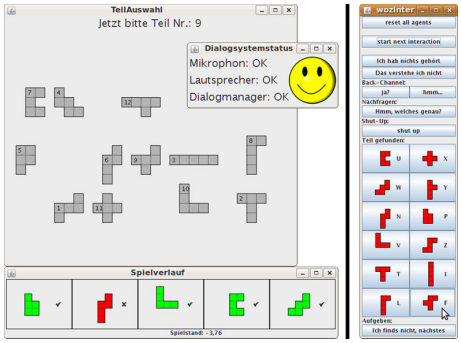


Figure 1: Setting of the Wizard-of-Oz experiment used for obtaining our data with the user interface to the left and the wizard interface to the right.

$oWER_t = \min(WER_t^{1..N})$, ($aWER_t = \max(WER_t^{1..N})$); this gives an indication of the possible gain if we identify the best hypothesis, or loss should we strike on the worst. We are also interested in their respective positions in the n-best lists: $oPOS_t = \arg\min_{n \in 1..N} WER_t^n$, ($\arg\max$ for $aPOS_t$).

3.2. NLU measures

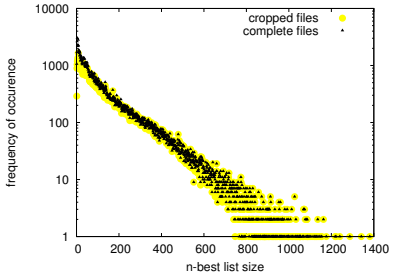
Similar to WER for ASR output, we use CER (concept error rate) as our base measure for incremental NLU metrics. In principle, the gold standard for evaluating NLU hypotheses changes (or rather, expands) over time just as that for ASR: the more of the utterance has been processed, the more is knowable about its meaning. In our experiments explained below, however, the NLU task is simplified to filling just one slot, and we assume that filling the slot should occur as early as possible; because of this, we do not need temporal alignment of semantic information.

Unlike in the non-incremental case, the semantic slot can be unfilled at times during the processing of the utterance. We only consider this to be an error toward the end of an utterance, when some meaning should have been extracted; to express this idea, we define an incremental uncertainty-adjusted concept error at time t for each slot. It is 0 if the slot is correctly filled, 1 if the slot is incorrectly filled, and $\alpha \frac{t}{t_{max}}$ if the slot is unfilled (while it is filled in the gold standard, t_{max} is the utterance duration). α (with $0 \leq \alpha \leq 1$) denotes how much better no answer is than a wrong answer. We can derive oracle results and positions in the n-best list for the uncertainty-adjusted CER in the same way as we did with WER above.

We adapt the Edit Overhead (EO) from [3] to count the number of edits between all adjacent n-best lists (only counting additions and deletions between the lists, not changes in position) over the course of the utterance compared to the number of edits that would have been necessary (1 in our case as there is one concept to be filled per utterance). Also, we define *concept first correct* (CFC) to be the percentage into the utterance at which the referent was first (oracularly) correctly resolved.

4. Setup and corpora

We use the Sphinx-4 speech recognition framework [11] for our experiments, using the built-in LexTree decoder which we extended to provide incremental n-best hypotheses. The N best results are constructed at each time step (10 ms) from the list of best ranking tokens provided by the token-pass algorithm. We built German acoustic models based on a small corpus of spontaneous instructions in a puzzle building domain,[3] and the Kiel corpus [12]; our statistical language model was trained on the puzzle domain data.



size	freq
0	57013
1	32383
2	7505
3	4341
4	3051
5	2077
6	1084
7	920
8	421
9	85

Figure 2: Distribution of n-best list sizes for ASR and NLU

We collected evaluation data in a Wizard-of-Oz setting with 11 participants producing in total 255 utterances, 2364 words and approximately 18 1/2 minutes of speech. The subjects were shown 12 pentomino pieces[4] on the screen (see Figure 1), one of which was marked as the one to describe. The subjects were prompted by synthesized speech to instruct the (wizard-driven) system which one to pick. The utterances were transcribed manually and word timings were produced using forced alignment with the MAUS tool [13]. The semantic annotation (the intended referent) was derived automatically from the instruction to the user and the wizard's selection.

Due to computational limitations we had to restrict Sphinx's active list to 100,000 entries (including acoustic variations) and a large (but not infinite) relative beam width. We analyzed ASR n-best list sizes by utterance time and found the averaged sizes to be remarkably stable over time. Exceptions to this were utterance-final and initial silences where there were fewer hypotheses on average. This most likely stems from the fact that understanding silence is much easier than understanding speech and varying leading and trailing silence would vary results (regardless of the actual recognition process). We thus decided to ignore ASR results during leading and trailing silences (according to the gold alignment) for further analyses. Notice also that WER would not be defined before the first word in the gold alignment starts, because of the normalization by the number of words in gold.

Figure 2 shows the size distribution of incremental n-best lists limited to words for cropped and complete utterances. It shows that very large n-best lists are rare: while there is one n-best list in our corpus with 1381 different hypotheses, more than 95 % of the n-best lists have less than 433 entries, and the median length of n-best lists is 85. In comparison, utterance final n-best lists are much shorter on average: The maximum length is 293; over 95 % are shorter than 210 entries; the median length is below 37. In other words, incrementally analyzing all available n-best hypotheses not only incurs additional computational cost because at each frame in time N hypotheses have to be considered, but also because the in-utterance N are much higher than utterance-final N. In order to limit the computational cost, we limit N to 200 for our analyses (completely covering more than 77 % of the n-best lists, while truncating the rest). One goal of our experiments described below is to find an even lower plausible threshold for N.

Non-incremental WER for our ASR is 62 % (oracle: 57 %

[3] Available from `http://www.voxforge.org/home/downloads/speech/`

[4] All geometrical shapes that can be formed by attaching 5 squares by their edges (irrespective of symmetry or orientation).

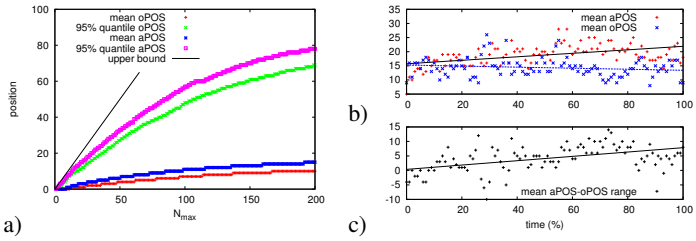


Figure 3: Positions of (anti)-oracle WER with different N_{max} and means of oPOS, aPOS and their difference over time.

anti-oracle: 76 %).[5]

Our domain is limited to finding one among twelve puzzle pieces, so we are only dealing with reference resolution although our incremental semantic component (described in detail in [14]) from which CER are computed, is capable of more complex semantic extraction. The grammar which is used with the component was developed using approximately 150 sentences from a larger domain, where pieces could not only be selected but also flipped, turned and moved to a number of places; the current domain is a sub-domain, where we only select pieces and are so far only interested in one of the slots of the frame semantics, the *object:name*-slot.

Figure 2 also shows the distribution of n-best list sizes for the NLU component given the ASR's input. It never finds more than 9 different meanings among the ASR hypotheses and most frequently does not find a meaning at all (which may be the right thing to do, as the content words for the referent may not yet have been said). In fact, empty hypotheses more often occur in the beginning of the utterance and are less common towards the end (cf. Section 5).

5. Experiments and results

In this section we first analyze incremental ASR measures against limited n-best lists of size N_{max}. Then we analyze incremental NLU measures depending on variants of ASR output and try to draw conclusions on this.

Figure 3 (a) shows the mean positions of the (anti)-oracle WERs in the n-best lists for varying N_{max}. Both the best and the worst WERs occur on average quite early into the n-best lists, and are rare to be at the end the lists, as can be seen from the position of the 95 % quantiles. Thus, from the point of view of ASR, using a large n-best list does not seem to buy one much. Figure 3 (b) plots the mean positions of (anti-)oracle WERs over time (i. e., percentage of the utterance). oPOS and aPOS are at roughly the same position (on average) towards the beginning of the utterance and oPOS seems to decrease while aPOS increases as illustrated by the linear regression of their respective values. This tendency can be seen more clearly in Figure 3 (c) which plots the difference in position between anti-oracle and oracle WER over time / percentage of utterance. We can conclude from this that using higher N in the beginning of utterances promises a higher profit but is connected with a higher risk, while we could reduce N towards the end of the utterance and at the same time be more certain to still keep available the objectively best hypothesis, while reducing the danger of choosing a bad hypothesis.

Figure 4 (a) shows that n-best ASR produces a considerably

[5]Note that our corpus contains spontaneous speech of unknown speakers that contains many disfluencies; we also used relatively little training data in building our acoustic models and did not tweak the ASR parameters for maximum performance.

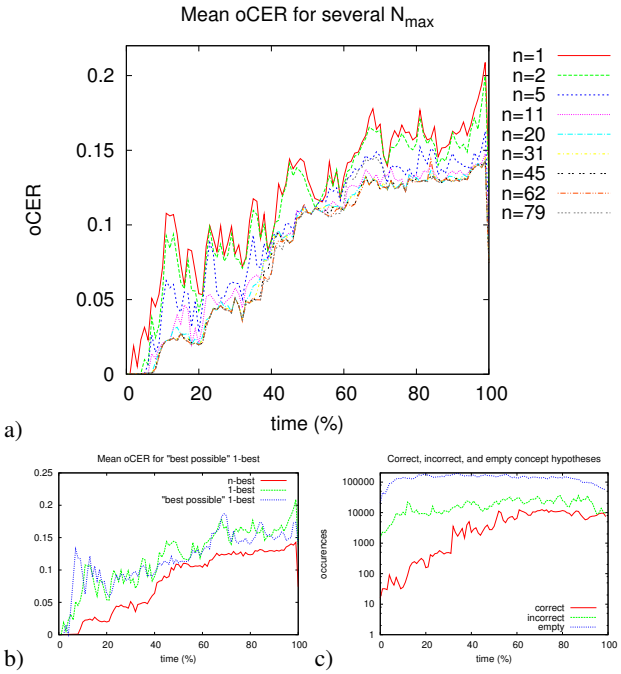


Figure 4: NLU performance: oracle CER over time in different settings (a and b), and result distributions in the output (c).

lower concept error rate than the mean best concept error with decreasing N. This is even for small N and particularly so at the beginning of utterances . We can conclude that in our setting using n-best ASR is promising (in terms of CER) especially from the perspective of incremental processing considerations and for filtering early false positives. Towards the utterance end the difference is less pronounced, but still present.

To find out whether optimizing the ASR's output would improve NLU performance, we compare 1-best performance with the performance using the "best" (in terms of WER) hypotheses. Figure 4 (b) shows that this does not result in an improvement. It seems that re-ranking is inferior to using N hypotheses, which confirms the strategy used by [6] and extends their results to the incremental domain.

We noticed that the mean number of concept hypotheses found in the entire n-best list approaches 1 over time. This is important because an n-best approach that yields a greater number would be more difficult to post-process. With this observation we feel confident that post-processing of n-best ASR hypotheses from a NLU point-of-view is merited.

In terms of timing, the point of first correctness (CFC), i. e. the point in the utterance at which a correct hypothesis first occurs in an n-best list compared to the 1-best case, is an indication of whether n-best lists allow a measurable timing advantage (resolving speaker intentions sooner). The mean CFC for 1-best was at 51.2 % utterance completion, while for the n-best case it was 41.0 %, a relative improvement of 20 %.

To determine the overall utility of a potential gain, however, the edit overhead (cf. Section 3) of n-best list processing must be considered. The 1-best case exhibits a mean of 256 edits while the n-best case results in 39368 edits (spread over an average 117,708 incremental results per utterance). Thus, a potential 20 % gain in CFC is accompanied by a 150-fold increase in EO. This discrepancy could be addressed by post-processing similar to [3], which we will investigate in future work.

In an initial look at how this can be accomplished, Figure 4 (c) describes the distribution of empty, incorrect and correct NLU hypotheses for the ASR n-best list that provided both the 20 % gain in CFC and the increase in edit overhead. Utterance-initially the number of correct hypotheses is far lower than that of the incorrect ones,[6] and always than the empty case. With greater utterance completion, the correct and incorrect values converge sharply. It thus illustrates that post-processing must occur in conjunction with timing measures, but also what these timing measures may be in the future.

6. Discussion and Future Work

Our effort thus far has been to demonstrate *potential* timing and accuracy gains in WER and CER, to compare how gains in the former may effect gains the latter and to establish their overall utility. The experimental results point towards potential gains in timing and accuracy performance as well as challenges in controlling utility trade-offs. Moreover, we were able to deduce that a post-processing approach of n-best ASR output must be informed not by properties of the recognition results itself, but by other measures, most likely timing and semantic ones, since a gain in CER could not be directly correlated with a gain in WER.

Finding an optimal operating point for N in terms of accuracy and edit overhead is relatively straightforward affair. Figure 4 (a) points towards a large gain in CER at an optimal N somewhere between 10 and 20, above which accuracy gains become smaller and the computational trade-off bigger. The mean edit overhead of limiting N to 11 is 15974, 60 times that of the 1-best baseline, however much more acceptable than the 150-fold increase seen with a practically limitless N (or at least the maximum observed in our experiments).

As mentioned in the introduction, we have ignored the problem of actually *finding* the 1-best result among our n-best hypotheses here and leave this for future work. An appropriate approach must however consider the computational impact (such as edit overhead) and, most importantly, some kind of measure for ruling out adverse hypotheses. These can take several forms (e. g. timing or symbolic ones).

The relationship between correct and incorrect hypotheses generated by enabling n-best recognition, captured by the (anti-)oracle WER statistics and CER Figure 4 (d), already point towards initial timing measures. Our experiments detail time as percentage of completion. These can easily be revised in absolute terms, which in turn can become timeout parameters (e. g. "do not accept semantic hypotheses before they are t milliseconds old") or break-points at which costly n-best processing can be turned on or off. Same goes for the first-correctness measure discussed above.

The ratio of semantically empty to non-empty recognition hypotheses is a further measure that can be explored, as well as the lexical/phonetic density of a semantic hypothesis (how many words/units were used to produce a hypothesis.)

7. Conclusions

We defined and explored timing, accuracy and utility measures for evaluating n-best lists in an incremental SDS both from a ASR accuracy and more holistic system point of view.

Our findings point towards potential gains in semantic performance of the SDS through the incremental use of n-best lists. Moreover, they illustrated that gains in performance of the incremental speech recognition achieved through n-best list re-ranking did not necessarily result in better semantic performance, from which we conclude that post-processing should not focus on re-ranking alone. Lastly, relatively low values of N seem to achieve the best trade-off in accuracy gain and computational overhead.

Some of the results observed are easily translated into real performance gains for the SDS, such as timeout values. In the process of exploring these issues we discovered several possible solutions to the question of how to actually identify the best hypothesis in the list.

Acknowledgements
This work was funded by a DFG grant in the Emmy Noether programme.

8. References

[1] J. Odell, V. Valtchev, P. Woodland, and S. Young, "A One Pass Decoder Design For Large Vocabulary Recognition," in *Proc. of the ARPA Workshop on Human Language Technology*, 1994.

[2] M. Rayner, D. Carter, V. Digalais, and P. Price, "Combining knowledge sources to reorder n-best speech hypothesis lists," in *Proc. of the ARPA Human Language Technology Workshop*, Plainsboro, USA, 1994.

[3] T. Baumann, M. Atterer, and D. Schlangen, "Assessing and improving the performance of speech recognition for incremental systems," in *Proc. of NAACL-HLT*, Boulder, USA, 2009.

[4] A. Chotimongkol and A. Rudnicky, "N-best speech hypothesis reordering using linear regression," in *Proc. of Eurospeech*, Aalborg, Denmark, 2001.

[5] M. Gabsdil and O. Lemon, "Combining acoustic and pragmatic features to predict recognition performance in spoken dialogue systems," in *Proc. of ACL*, Barcelona, Spain, 2004.

[6] M. Purver, F. Ratiu, and L. Cavedon, "Robust interpretation in dialogue by combining confidence scores with contextual features," in *Proc. of Interspeech*, Pittsburgh, USA, 2006.

[7] J. D. Williams, "Exploiting the asr n-best by tracking multiple dialog state hypotheses," in *Proc. of Interspeech-ICSLP*, Brisbane, Australia, 2008.

[8] C. Lee, S. Jung, and G. G. Lee, "Robust dialog management with n-best hypotheses using dialog examples and agenda," in *Proc. of ACL-HLT*, Columbus, USA, 2008.

[9] N. Miyazaki, M. Nakano, and K. Aikawa, "Robust speech understanding using incremental understanding with n-best recognition hypotheses," *Joho Shori Gakkai Kenkyu Hokoku*, vol. 10, 2002.

[10] M. Nakano, N. Miyazaki, J. Hirasawa, K. Dohsaka, and T. Kawabata, "Understanding unsegmented user utterances in real-time spoken dialogue systems," in *Proc. of ACL*, College Park, USA, 1999.

[11] W. Walker, P. Lamere, P. Kwok, B. Raj, R. Singh, E. Gouvea, P. Wolf, and J. Woelfel, "Sphinx-4: A flexible open source framework for speech recognition," Sun Microsystems Inc., Tech. Rep. SMLI TR2004-0811, 2004.

[12] IPDS, "The Kiel Corpus of Read Speech," CD-ROM, Kiel, Germany, 1994.

[13] F. Schiel, "Maus goes iterative," in *Proc. of LREC*, Lisbon, Portugal, 2004.

[14] M. Atterer and D. Schlangen, "Rubisc – a robust unification-based incremental semantic chunker," in *Proc. of 2nd International Workshop on Semantic Representation of Spoken Language*, Athens, Greece, 2009.

[6] Note that the baseline is $1/12$, which relativises our error rate.