EVALUATING SPEECH RECOGNITION IN THE CONTEXT OF A SPOKEN DIALOGUE SYSTEM: CRITICAL ERROR RATE

Géraldine Damnati

France Télécom R&D, DIH/IPS
2 av. Pierre Marzin
22307 Lannion Cedex, France
geraldine.damnati@rd.francetelecom.com

ABSTRACT

Evaluating a speech recognition system is a key issue towards understanding its deficiencies and focusing potential improvements on useful aspects. When a system is designed for a given application, it is particularly relevant to have an evaluation procedure that reflects the role of the system in this application. Evaluating continuous speech recognition through word error rate is not completely appropriate when the speech recognizer is used as spoken dialogue system input. Some errors are particularly harmful, when they concern content words for example, while some others do not have any impact on the following comprehension step. The attempt is not to evaluate natural language understanding but to propose a more appropriate evaluation of speech recognition, by making use of semantic information to define the notion of critical errors.

1. INTRODUCTION

Assessment in speech processing is now accepted in the speech community as a key issue for technology improvements. Evaluation campaigns have proven to be the occasion of significant technical improvements. Those campaigns have also raised the importance of a suitable evaluation criterion which allows objective and reproducible experiments to be run.

In the spoken dialogue domain, interaction between speech recognition and speech understanding is so tight that the classical speech recognition measure (word error rate) is not exactly appropriate.

The attempt of this paper is to provide an evaluation procedure for speech recognition systems as the input of a human-machine spoken dialogue system. The global word error rate indicates the intrinsic recognizer's performance but is not representative of the system's performance with respect to the task it has been designed for. In a spoken dialogue system the recognizer's output is to be interpreted by the comprehension module and it is obvious that every recognition error does not have the same impact for message understanding.

In fact, the natural language interpreter does not use directly the recognized word sequence but first transforms it into a concept sequence from which the comprehension task is achieved.

The interpreter in Artimis system [Sadek,96] achieves the word/concept transformation in three steps.

- First, a distinction between meaningless (empty) words and meaningful (non-empty) words is achieved from an predefined list of empty words. The interpretation task is therefore initialized by a filtered sequence of non-empty words.
- As a word can lead to different concepts depending on the context, several concept sequences can be derived from the filtered word sequence. The possible concept sequences are then given in a N-best list.
- Disambiguation is processed in the third step by making use of the dialogue history. It leads to the concept sequence that is to be interpreted. This step does not use only *a priori* information but makes use of dynamic information provided by the dialogue context.

Several approaches towards a more realistic evaluation paradigm have been proposed. In order to evaluate the natural language understanding task, concept accuracy [Boros,98] allows the error rate to be computed at the concept level. Several problems are related to this procedure such as reference annotation and reproducibility. [Hanrieder,98] raises the problem of manual concept annotation of the reference and proposes to use the output of the interpreter from the real utterances as the reference. Beyond the reference annotation, this approach is practically problematic and does not correspond to our attempts. In fact, making use of the dialogue history for the interpretation results in non reproducible results. Changing the speech

recognition system implies different word sequences, different system answers and necessarily a different dialogue progress. The only way to thoroughly evaluate concept accuracy would be to collect data with the dialogue system but without speech recognition, an operator capturing the utterance for the interpreter input. Afterwards, each utterance can be independently processed by a speech recognition system and submitted to the interpreter with the reference history given by the real utterances. In such conditions, the concept sequence can be compared to the reference concept sequence of the acquisition step. This framework is rather constraining. In fact, if the interpreter is to evolve, data can no longer be used for evaluation of different recognition systems. What's more, data are usually collected whether through a complete system or through a WOZ that simulates the dialogue system.

The interest of an evaluation method is not only to compute a system's performance at a given time but also to observe the performance evolution when a parameter is changed at any level of the process. In particular, in order to evaluate the impact of speech recognition improvements on the natural language interpreter input, one has to be able to run the evaluation on the same data before and after changing the speech recognition system.

In order to have a correct and reproducible evaluation paradigm for speech recognition we propose an hybrid evaluation which approximates concept accuracy. It focuses on errors that are potentially harmful for the following interpretation step. Such errors are called critical errors.

The next section presents the way word sequences have to be processed in order to obtain the critical error rate. Then a short section presents the dialogue application and the data that were used to illustrate the interest of the assessment method. Experiments are finally presented in section 4. The first experiment illustrates the critical error rate definition and further experiments illustrate the evolution of word error rate and critical error rate in various conditions. This second illustration show how complementary the two measures are.

2. CRITICAL ERROR RATE

The problematic step to reach the concept level being the disambiguation, we avoid it by focusing on words that can be associated to only one concept. For such words, the first two interpretation steps are sufficient to determine the corresponding concept.
As the contextual disambiguation step is avoided, required semantic information is restricted to *a priori* information available from the natural language interpreter.

Both the reference and the recognizer's output are automatically transformed into an hybrid sequence of words and concepts. The conceptual transformation being done as far as possible without any knowledge of the dialogue history. The alignment for error rate evaluation is computed over the transformed sequences.

2.1. Empty words processing

Empty words are first filtered in both the reference and the recognized sequence. Two solutions can be adopted leading to different alignments. One possibility is to replace an empty word by a unique symbol <EMPTY>. The second solution consists in suppressing directly empty words.

Each strategy has a different impact on alignments and on error counting. The following example is adapted from a real example. The reference sequence is the succession of the five words $\{W_A\ W_B\ W_C\ W_D\ W_E\}$ and the decoded sequence is composed of the three words $\{W_F\ W_G\ W_D\}$. Given that W_B and W_C are empty words, the first filtering would lead to the following alignment (as provided by the NIST evaluation toolkit):

Ref:	W_A	<EMPTY>	<EMPTY>	W_D	W_E
Dec:	*	W_F	W_G	W_D	*

The second solution leads to the following alignment:

Ref:	*	A	D	E
Dec:	F	G	D	*

In the first case, the alignment provides two deletions and two substitutions and in the second case, the alignment provides one insertion, one deletion and one substitution.

In what follows we adopt the second solution which is more directly interpretable. The first solution implies that errors such as empty word deletions or empty word insertions are counted in the error rate and would have to be discounted. On the other hand, the second solutions focuses on relevant words and is closer to our objectives.

However, suppressing empty words implies that the number of reference words is modified. Therefore, initial word error rate and critical error rate should not be compared directly. That is why the absolute number of errors will be given as a complement.

2.2. Conceptual processing

Conceptual information can be difficult to reach from the speech recognition point of view when the correspondence between a word and its concept is ambiguous. However some correspondences are unique and no ambiguity disturbs concept identification. In such cases, the concept attribution only makes use of *a priori* knowledge of the natural language interpretation system. This *a priori* knowledge can be easily exploited to improve evaluation. In other words, the evaluation criterion can reach the conceptual level for those words which can be associated to a single concept.

Thus, the second step concerns non empty words that are associated to a single concept. These words are simply replaced by their concept. Doing this, substitutions between two words that are associated to a same concept are not counted in the global error rate. Such errors can occur when words are similar (for example two words derived from a same lemma).

The resulting sequence can then contain words as well as concepts. Once the reference sequence is applied the same treatment, the two filtered sequences are aligned in order to compute an hybrid error rate which only reflects critical errors from the dialogue point of view.

The result does not completely reflect reality as some critical errors can turn out to be harmless as the context can help to find the good interpretation despite recognition errors. Anyway, critical error rate is rigorous and easy to obtain. It is a good intermediate between inappropriate word error rate and theoretical concept error rate. It provides an upper bound of the errors that can be harmful for speech understanding.

3. DATA DESCRIPTION

The approach has been used on the PlanResto demonstrator developed at France Telecom R&D. This application allows the user to find a restaurant in Paris according to some criteria such as location, price or specialities.

The lexicon which contains 1916 entries has to be exhaustive for several domains (such as subway stations). 497 lexicon entries are empty words and 1064 lexicon entries are associated to a unique concept.

This application has been recently developed and yet only a low amount of data is available. The training corpus is composed of around 4500 utterances leading to around 20.000 word occurrences and the test corpus contains 942 utterances corresponding to 3841 word occurrences. Training data are only used to train the language model, no acoustic adaptation as been performed so far over our generic spontaneous speech acoustic model.

21.4% of the test set occurrences are empty words and 13.0% of words in the test are non-empty words that are associated to several concepts. Finally the filtered reference contains 3018 items, 16.6% of which being words and 83.4% of which being concepts.

4. EXPERIMENTS

Speech recognition experiments are run with the France Telecom recognition software PhilSoft. The recognition model includes a multigaussian acoustic model and a bigram language model smoothed by hierarchical smoothing. Hierarchical smoothing was introduced in [Damnati, 00] and consists in using several word-class mappings extracted from a same hierarchical classification tree of the lexicon. Each mapping (corresponding to a different level in the tree) leads to a class-based model. The closer to the root the mapping is extracted, the more general the model is. Those models can be combined in the classical framework of absolute discounting, the first level corresponding to the most precise level extracted from the tree and the backing off being achieved successively towards the more general models. This smoothing technique have turned out to be very efficient as it takes full advantage of the hierarchical structure of the classification tree.

4.1. Baseline experiment

The baseline word-error rate is 22.1% that is 850 errors. A first evaluation concerns non-empty words. Both the reference and the recognized sequences are filtered, which results in a 19.6% error rate over the 3018 non-empty words occurrences (591 errors). When including correspondences for non-empty words that are associated to a single concept the critical error rate reaches 17.5% (527 errors).

	All	Non-empty	Critical
Nb. Words	3841	3018	3018
Nb. Errors	850	591	527
Error rate	22,1 %	19,6%	**17,5%**
Correct rate	80,9%	83,3%	85,4%

Table 1. Word error rate and critical error rate evaluation.

Considering the absolute number of errors, it appears that 69,5% of the errors are made on non-empty words and 10,8% of these errors are in fact substitutions of words that correspond to the same concept. The number of errors in the last column indicates, relatively to the total amount of errors, that at most 62% of the errors can lead to a comprehension problem.

4.2. Comparative experiments

In order to verify that critical error rate is not fully correlated to word error rate, which would be of low interest, we have run a set of experiments in various conditions. First, the pruning conditions are modified and second the language model is changed. Observing the evolution of the two error rates leads to interesting conclusions and show that they can be complementary.

In the next table, the pruning parameters are chosen from a very restrictive pruning to a permissive pruning. Of course the word error rate is degraded in restricted conditions while the real time factor also decreases.

Three levels are reported in the table, L1 corresponding to the most restrictive conditions. The table focuses on the number of errors that are made on non empty words and the number of critical errors. The proportion of such errors relatively to the total number of errors is also given.

Number of errors	All	Non-empty	Critical
L1	996	696 *(69.9%)*	636 *(63.9%)*
L2	927	642 *(69.3%)*	580 *(62.6%)*
L3	850	591 *(69.5%)*	527 *(62.0%)*

Table 2. Number of errors in various pruning conditions.

The proportion of errors made on non empty words does not evolve significantly with the pruning conditions. On the other hand, the proportion of critical errors is more important in restrictive conditions (63.9%). This result illustrates the fact that in less restrictive conditions the decoding process can select words that are acoustically similar to the reference even if the bigram probability is low. If the similar words relate to a same concept, the error is not harmful for comprehension.

For the second illustration, the language model has been changed. A poor smoothing technique, known as threshold smoothing, is applied to the model. Performance are worse than with hierarchical smoothing but it is interesting to note the proportion of critical errors in the same pruning conditions for different smoothing techniques.

Number of errors	All	Non-empty	Critical
Threshold	1103	781 *(70.8%)*	721 *(65.4%)*
Hierarchical	850	591 *(69.5%)*	527 *(62.0%)*
Reduction	**22.9%**	**24.3%**	**26.9%**

Table 3. Number of errors for threshold smoothing and hierarchical smoothing.

These results show that the hierarchical smoothing technique improves the overall performance (22.9% reduction) but also that the relative improvement is higher on potentially harmful errors (26.9%). This is particularly interesting from the application point of view.

CONCLUSION

The evaluation method proposed in this paper gives an upper bound of the speech recognition errors that can be harmful for natural language understanding in the context of a spoken dialogue system. The interest of this method is that the speech recognition module is evaluated with respect to the application it is designed for. It is simple to apply as it only makes use of *a priori* knowledge available in the interpretation module. What's more it is reproducible which means that it can be used to compare different models on the same data in order to observe the impact on critical errors.

REFERENCES

[Boros, 96] M. Boros, W. Ecker, F. Gallwitz, G. Görz, G. Hanrieder, H. Niemann, "*Towards Understanding Spontaneous Speech: Word Accuracy vs. Concept Accuracy*", Proceedings of ICSLP'96, pp. 843-846, Philadelphie, USA, 1996.

[Damnati, 00] G. Damnati, "*Modèles de langage et classification automatique pour la reconnaissance de la parole continue dans un contexte de dialogue oral homme-machine*", PhD Thesis, University of Avignon, France, 2000.

[Hanrieder, 98] G. Hanrieder, P. Heisterkamp, T. Brey, "*Fly with the EAGLES: evaluation of the "ACCeSS" spoken language dialogue system*", Proceedings of ICSLP'98, pp. 503-506, Sidney, Australia, 1998.

[Sadek, 96] D. Sadek, A. Ferrieux, A. Cozannet, P. Bretier, F. Panaget and J. Simonin, "*Effective Human-Computer Cooperative Spoken Dialogue : the AGS Demonstrator*". Proceedings of ICSLP'96, pp. 546-549, Philadelphia, USA, 1996.