INCORPORATING CONFIDENCE MEASURES IN THE DUTCH TRAIN TIMETABLE INFORMATION SYSTEM DEVELOPED IN THE ARISE PROJECT

G. Bouwman, J. Sturm, L. Boves

University of Nijmegen
P.O. Box 9103
6500 HD Nijmegen, The Netherlands

ABSTRACT

The use of Confidence Measures (CMs) in Spoken Dialog System (SDS) applications to suppress the number of verification turns for 'reliably correctly recognised utterances' can greatly reduce average dialog length which enhances usability and increases user satisfaction [1]. This paper gives a brief but clear review of the method of CM assessment, which was presented in [2]. It proceeds by demonstrating how the Dutch ARISE (Automatic Railways Information Systems in Europe) SDS was equipped with this technology and shows in deep detail how the parameters involved are to be optimised. The evaluation reveals and explains a typical behaviour of this method with train timetable information-alike systems. This results in a set of conclusions that were not foreseen when the method was first developed for a directory information system. The paper ends with an outlook for solutions in new research directions.

1. INTRODUCTION

A number of telephone based travel information systems has been built since the (D)ARPA funded ATIS program and the advances offered in automatic speech recognition (ASR) technology. At this moment automatic train timetable information systems are operational in Switzerland and in the Netherlands. These systems are localised versions of a German prototype developed by Philips [2]. The most characteristic features of these systems are the use of mixed-initiative dialogue control and implicit verification; both meant to make the human-computer interaction faster and more natural.

Analyses of caller behaviour, both in the laboratory and in the field, have shown that many users have difficulty in grasping the concept of implicit verification. If a caller said "I want to go from Arnhem to Amsterdam", and the system replies with "When do you want to travel from Haarlem to Amsterdam?", many callers are confused by the combination of a verification question and a question for additional information. This has prompted research into alternative dialogue strategies, that avoid implicit verification, without incurring the cost of a much longer and more tedious interaction.

In the ARISE (Automatic Railways Information Systems in Europe) project we develop a train timetable information system that combines the mixed initiative option with explicit verification in the first part of a dialogue. In theory explicit verification would raise the number of turns, and therewith the expected duration of a typical dialogue.

The example in Figure 1 shows an excerpt from a real dialogue. System and user utterances are denoted by Sx and Ux respectively, followed by the spoken Dutch sentence and the English translation in *italics*.

S1	Van waar naar waar wilt u reizen?	*From where to where do you want to travel?*
U1	Ik zou graag naar Amsterdam reizen.	*I'd like to travel to Amsterdam.*
S2	Wilt u naar Amsterdam?	*Do you want to go to Amsterdam?*
U2	Ja, dat klopt.	*Yes, that's right.*
S3	Waar vandaan wilt u vertrekken?	*Where do you want to leave?*
U3	Uit Haarlem, alstublieft.	*From Haarlem, please.*
S4	Wilt u uit Arnhem vertrekken?	*Do you want to leave from Arnhem?*
U4	Nee, Haarlem.	*No, Haarlem.*
...	...	...

Figure 1. Example dialogue showing explicit verification

It is not hard to see that implicit verification could shorten the information-providing part of the dialogue by half the number of turns, as long as users are not confused by an incorrect recognition. As dialogue length is a critical factor for the usability of an SDS [1], it would be highly desirable to have some kind of confidence measure at dialogue level to get an indication whether the speech recogniser has confidence in the correctness of a certain information item or not. This gives the ability to verify explicitly only in the cases where the system has insufficient confidence in a correct recognition.

A method to deduce such confidence measures was proposed in [3]. In contrast to other methods that use acoustic word score only, the basic idea of this method is the following: if all scored sentence hypotheses within a predefined score distance from the first best sentence show consensus about a particular information item A, then item A is assumed to be reliable.

This paper will examine the suitability of this method for the ARISE system. It is organised as follows: in section 2 a review of the method to deduce the measures is given. In the third section we will show how we optimised the most important parameters

for particularly our system. Section 4 presents the results of the tests we did and typical problems for travel information systems using this method. In section 5 we draw conclusions for both this confidence measure and the ARISE system.

2. CONFIDENCE MEASURES BASED ON SENTENCE PROBABILITIES

Before the method can be applied, the *word graph* output of the recogniser needs to be processed. The word graph is a compact representation of all word hypotheses within an utterance. For every word also an acoustical score is provided. The word graph is submitted to an application dependent grammar, which builds so-called concepts from meaningful word sequences. Each grammar rule defining a concept, is responsible to deduce at least one *attribute*. Attributes are the most elementary information items and are to be used to fill in the final database query. For the words that don't comply with any concept, so-called filler arcs are created. Figure 2 and Figure 3 show an example of a word graph and its corresponding concept graph.

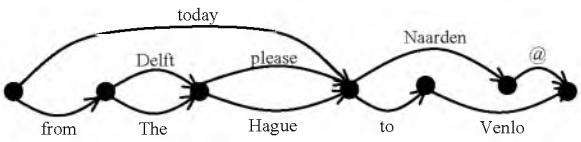

Figure 2. Example word graph

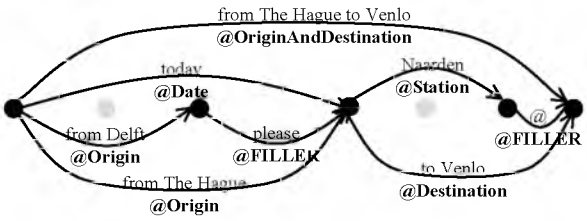

Figure 3. Corresponding concept graph

Every path through a concept graph represents a sentence hypothesis, provided that the attributes deduced from concepts within one path do not contradict each other. Every sentence gets a score based on the acoustical scores of its words, its language model probability, and the concept grammar probability. The N-best list of possible sentences ranked by their scores forms the starting point of the method.

First, an empirically established factor scales the distribution of the sentence scores. Because of the Log-Likelihood nature of the scores, this causes an exponential redistribution of the probabilities, making the score less sensible for small changes in the threshold:

$$sc_{scaled} = \alpha \cdot sc, \qquad (1)$$

where α is typically positive and smaller than 1. Since the scores are negative logarithms of probabilities, they can be restored again to represent sentence probabilities. Because of the fact that only n sentences are under consideration and the scores were scaled by α, they are normalised by a second factor to sum up to one:

$$p_s = \beta \cdot e^{-sc_{scaled}}, \qquad (2)$$

such that $\sum_{s=1}^{n} p_s = 1$, where p_s is the sentence probability.

Now that the best scores correspond with the highest probabilities, every attribute a in the first best sentence is assigned an attribute probability (p_a):

$$p_a = \sum_{s=1}^{n} p_s \cdot \delta_{a,s}, \qquad (3)$$

with $\delta_{a,s}$ being 1 for all sentences containing attribute a, and 0 otherwise, irrespective of the concept that was responsible for setting it. For example, if the first and only two best sentence hypotheses would be:

H_0: "From The Hague to Venlo" [$p(H_0) = 0.85$]
H_1: "From Delft, please, to Venlo" [$p(H_1) = 0.15$],

then the concepts @OriginAndDestination (for H_0), @Origin and @Destination (both for H_1) set the attributes

`origin = the hague [prob. = 0.85]` (= 1·$p(H_0)$ + 0·$p(H_1)$)
`destination = venlo [prob. = 1.00]` (= 1·$p(H_0)$ + 1·$p(H_1)$)

In order to determine whether an attribute is reliable or not, only the sentences within a predetermined score range of the first-best sentence are considered, limited by a constant maximum number of sentences. If all these sentences set the same attribute, i.e. p_a = 1.00, the value is considered to be reliable. In all other cases it marked as unreliable. For this, two parameters and a strategy need to be determined:

1. the ***pruning threshold*** of the recogniser that is responsible for the size of the word graphs and therefore the number of competing sentences;
2. the ***preference strategy*** which is a set of rules to parse the word graphs non-ambiguously in such a way that competing concepts can be compared; and
3. the above-mentioned score range or ***score distance*** which is directly responsible for the precision and recall of the measure.

3. PARAMETER ASSESSMENT

3.1 Pruning threshold

The system used to take the first best sentence only, but because of the fact that the method needs sentence hypotheses which differ only in the attributes, the speech recognition component has to supply sufficiently large word graphs; this increases the probability that if the first-best hypothesis (H_0) is incorrect, the correct one is at least among the competitors, causing the attributes of H_0 to be unreliable. This property makes a reconsideration of the pruning threshold necessary.

A single path word graph (SPWG) is a word graph that consists of just one path, in other words, yields only one hypothesis. Restricted to reaction utterances to the first question ("From where to where do you want to travel?"), the baseline settings of our recogniser generated a SPWG in 31.3% of the cases. Over a 94% of them were literally correctly recognised. Manual checking showed that the remaining 5.9% would partially succeed at concept level, i.e. yield the correct attributes. Overall, we had successful understanding of 96.1% of the utterances. Table 1 shows these percentages for different pruning thresholds.

Pruning Threshold	20,000	30,000	50,000
# word graphs	3140	3140	3140
# with single path	1503 (=47.8%)	984 (=31.3%)	237 (=7.6%)
# of which correct	1314 (=87.4%)	926 (=94.1%)	231 (=97.5%)
# correct concepts	1372 (=91.2%)	946 (=96.1%)	235 (=99.2%)

Table 1. Accuracy effects of pruning word graphs

Taken in consideration that the first utterance in a dialog usually is the most complex and therefore the most difficult to recognise, a pruning threshold of 30,000 gives enough confidence that the correct hypothesis is within the score distance range of H_0.

3.2 Preferences in case of ambiguous parses

Another optimum that had to be found concerns the preference strategy of the stochastic attributed concept grammar. Because of the method's pre-assumption that differences among competing attributes should cause the associated sentences to differ in score, the parses of a word graph should be subject to *preference* rules, which suppress one parse in favour of another. This prevents two different concept parses of the same word graph path to reinforce or weaken the individual scores of the involved attributes, which may bias the probability.

Originally, an utterance recognised as "Groningen to Amsterdam" would be parsed as both an @OriginAndDestination-concept and a @Station- plus a @Destination-concept. At dialogue management level we were free to choose the most likely one. If, for instance, the `origin`-attribute was already known and confirmed in the dialogue's history to be "Amsterdam", the H_0-hypothesis could not be unified with be current belief. H_1 is still able to supply new information about the possible destination, namely "Groningen". The confidence method however, requires to prefer the first one over the second, because an `origin`-attribute is not a `station`-attribute, which would make the first unreliable in all cases. Our preferences were therefore set subject to the rule of 'take the least abstract parse'.

3.3 Score distance

The maximal allowed score distance from the best to competing sentences is obviously the key parameter of this method: on the one hand, if it is too large then most best sentences will have competitors within range, resulting in a high *false rejection rate*. The false rejection rate is the number of correctly recognised attributes marked to be unreliable divided by the total number of attributes. On the other hand, a small value causes an incorrect H_0-hypothesis to be *falsely accepted*, because a competing (correct) alternative hypothesis H_1 might be out of range. The false acceptance rate, also known as *false alarm rate* is the ratio of incorrectly recognised attributes marked to be reliable and the total number of attributes. Therefore, the parameter was experimented with to determine the false alarm vs. false rejection rates, resulting in an Receiver Operation Characteristic (ROC) curve, see Figure 4

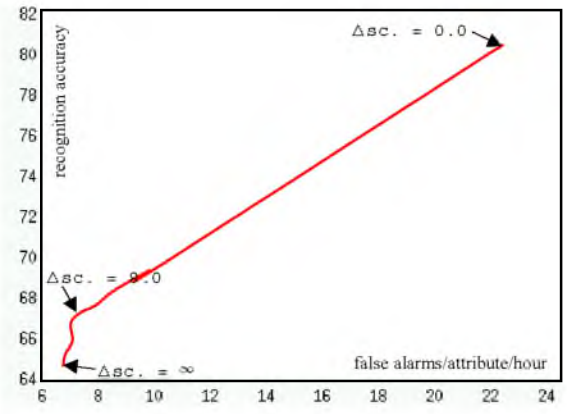

Figure 4. Receiver Operation Characteristic (ROC) curve

Accuracy refers to the percentage of attributes resulting from correctly understood concepts which were also marked to be reliable. False alarms per attribute per hour denote the false alarm rate. The point in the upper right corner concerns score distance 0.0, the baseline performance: H_0 is always accepted. Increasing the distance up to 8.0 gives a linear behaviour: the numbers of false rejection and false acceptance increase in equal proportion here. At this point (8.0) the score distance is taken to be the optimal one, because of the relative low cost of a rejection (extra turn in the dialog) versus the high cost of false acceptance (possibility of confusion).

4. TESTING/EVALUATION

The dialog strategy was adjusted in such a way that reliably recognised concepts are implicitly verified by repeating the information only, followed by a new question. Unreliable information is still verified in an explicit way. An example is shown in Figure 5.

S1	Van waar naar waar wilt u reizen?	*From where to where do you want to travel?*
U1	Ik zou graag naar Amsterdam reizen.	*I'd like to travel to Amsterdam.*
S2	Naar Amsterdam. Waar vandaan wilt u vertrekken?	*To Amsterdam. From where do you want to leave?*
...	...	...

Figure 5. Example dialogue showing implicit verification.

Theoretically the first part of a dialog, where the user provides information, could go with only half the number of turns. In order to be sure not to issue wrong train connection information, we still verify the last item(s) in an explicit way.

During our tests it became apparent the forced choice to reject certain parses of the same path in a word graph gave unwanted results in certain cases. An utterance like "Amsterdam Amstel" should be parsed as one @Station-concept, because it is the name of a station. However, the preference rule that an @OriginAndDestination-concept is better than a @Station-concept, resulted in an undesirable parse where "Amsterdam" is the origin and "Amsterdam Amstel" is the destination (for reasons that many people just say "Amstel" to refer to this station). When, for instance, an origin station is already verified, it should be up to the dialog management component to decide that a @Station-concept refers to a destination station. This choice is not to be made at concept parsing level, where there's no knowledge about dialogue history.

Another problem we had with the method revealed itself at the 'computational side' of the method. Suppose we have the following n-best sentence list followed by their score distance to the first-best. See Figure 6:

#	Sentence hypothesis	Score distance
H_0	I want from Delft to Venlo	0.000000
H_1	Elst	0.003411
	...	...
H_{m-1}	Best	0.055956
H_m	...	...
H_{n-1}	Delft Hengelo	0.960032
H_n	Elst	0.963443
	...	...
H_k	Best	1.013842

Figure 6: Example n-best hypotheses list

Now, the thick solid line denotes the maximal score distance (this example: 1.0). For computational reasons, the method also requires to set a parameter for the maximum number (m-1) of sentences considered. Now, suppose the system knows a number of 'acoustically highly confusable' words (Delft, Elst, Best, ...) that is at least as much as this maximum number of sentences. Not only all these words will always be marked to be unreliable, which could be justified, but also other concepts in the sentence (Hengelo), which might be found unreliable in other circumstances, are a potential danger for false acceptance.

The fact that several values for the maximum number of sentences were tested and found to yield equivalent results in the application of the original paper is not sufficient, because this problem is completely vocabulary dependent.

5. CONCLUSIONS AND OUTLOOK

This work showed that the use of Confidence Measures based on Sentence Probabilities is heavily dependent on having a non-ambiguous grammar and a low confusability in the lexicon, in other words, the application domain. One of the assumptions implicitly made by the method is that competing sentence hypotheses differ only in the values of the information items. As a consequence, the method forces the grammar to give single parses of one sentence, because otherwise it might cause false competition. For an application like ARISE however, the occurrence of partly specified information requires maximal flexibility and therefore freedom of parsing. In this way it stays possible to leave semantic related decisions to the next knowledge level, the level of dialogue management. The compromises needed to meet these contradictory requirements had serious impacts on the final result.

An important consideration is the fact that as a consequence of organisational reasons, the evaluations described are from a corpus which contained only users' reactions to the first utterance of the system. People tend to give a lot of information in complex sentences, which makes the recognition task much more difficult.

An intuitive solution to the problem described above would obviously be in looking at the actual word score of the involved information item, rather than deducing a measure from the score of the whole sentence. In the near future we will incorporate a new Confidence Measure, described in [5], which is based on a posteriori word probabilities. Within the framework of a Ph.D. project we will also start a research to propagate the 'real' acoustical CMs which are used during the recognition stage for Utterance Verification and Out Of Vocabulary word detection to the application level.

6. REFERENCES

[1] Bouwman G. and Hulstijn J. "Dialogue Strategy Redesign with Reliability Measures". *Proceedings of the First International Conference on Language Resources and Evaluation,* Granada, Spain, 1998, pages 191-198.

[2] Boves L., Herberts I. and Russel A. "Localisation and field test of a Dutch Train Time Table Information System". *Proceedings of the IEEE Third Workshop Interactive Voice Technology for Telecommunications Applications*, Granada, Spain, 1996, pages 89-92.

[3] Kellner A., Rüber, B., Seide F. and Tran, B.-H. "PADIS – an automatic telephone switchboard and directory information system". *Speech Communication,* 23:95--111, Oct, 1997

[4] Rüber B. "Obtaining Confidence Measures from Sentence Probabilities". *Proceedings of ESCA Eurospeech97,* Rhodes, Greece, 1997, pages 739-742.

[5] Wessel F., Macherey K. and Schlüter R. "Using Word Probabilities as Confidence Measures". *IEEE International Conference on Acoustics, Speech, and Signal Processing.* Seattle, May, 1998.