# 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. Popular applications include automatic attendants, stock quotation and travel information systems. 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 with the Dutch ARISE system. 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 this 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 attached to the output of the recognition and interpretation module. Utterances with a very high confidence need not be verified at all. If dialogue control can be reasonably confident that recognition was correct, it may use implicit verification. In all other cases explicit verification or rejection is called for.

A method to compute 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.

The method of [3] was successfully applied to shorten dialogues in an automatic attendant system. This paper will examine the suitability of this method for the more complex travel informa-

tion domain. The paper is organised as follows: in section 2 a review of the method to compute confidence measures is given. In section 3 we explain how we optimised the most important parameters for our particular 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 by 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 that are used to fill in the final database query form. 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.

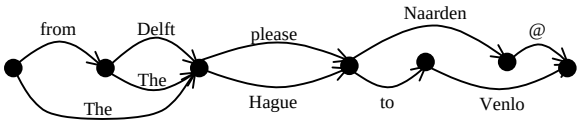


**Figure 2.** Example word graph.

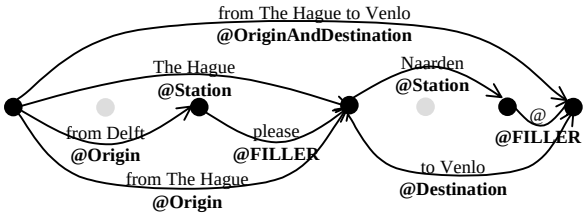


**Figure 3.** Corresponding concept graph.

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 scaling causes an exponential redistribution of the probabilities, making the score less sensitive to small changes in the threshold:

$$sc_{scaled} = \alpha \cdot sc\,, \tag{1}$$

where $\alpha$ is typically positive and smaller than 1. Since the scores are negative logarithms of probabilities, they must be transformed to sentence probabilities. Because of the fact that only the first $n$ sentences will be considered, the scores scaled by $\alpha$ must be normalised by a second factor $\beta$ to sum to 1:

$$p_s = \beta \cdot e^{-sc_{scaled}}\,, \tag{2}$$

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

Now that the scores are actual probabilities, every attribute $a$ in the first best sentence is assigned an attribute probability ( $p_a$ ):

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

with $\delta_{a,s} = 1$ for all sentences containing attribute $a$, and 0 otherwise, as long as the concepts responsible for setting the attributes can be unified. 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]` $(= 1 \cdot p(H_0) + 0 \cdot p(H_1))$
`destination = venlo [prob. = 1.00]` $(= 1 \cdot p(H_0) + 1 \cdot p(H_1))$

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 the score of an attribute exceeds a certain threshold, it is considered to be reliable. In this paper the extreme threshold value of $p_a = 1.00$ was chosen. In all other cases an attribute was marked as unreliable. For this, two parameters and a strategy need to be determined:

1. 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.
2. the ***pruning threshold*** of the recogniser that is responsible for the size of the word graphs and therefore the number of competing sentences;
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 Preferences in case of ambiguous parses

Another optimum that had to be found concerns the preference strategy of the stochastic attributed concept grammar. Because the method only allows to sum probability values of identical attributes (and attributes set by different concepts are considered as different), ambiguous parses should be avoided. This can be accomplished by the use of so-called *preference* rules, which suppress multiple ambiguous parses in favour of one best parse.

For instance, in the grammar of the original ARISE system an utterance recognised as "Groningen to Amsterdam" would be parsed as both an @OriginAndDestination-concept and a @Station- plus a @Destination-concept. The confidence method, however, requires to prefer the first one over the second, because an `origin`-attribute can not be matched with a `station`-

attribute, even though the same station name is concerned. So this would make the first unreliable in all cases. Our preferences were therefore set subject to the rule of 'take the least abstract parse'. We will come back to the consequences in section 4.

## 3.2 Pruning threshold

A single path word graph (SPWG) is a word graph that consists of just one path, in other words, yields only one hypothesis. The probability of this hypothesis scales to 1; thus, all attributes in a SPWG are considered as fully reliable. For the answers 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 94% of these were completely correctly recognised. Manual checking showed that a part of the remaining 5.9% was correct at concept level, i.e., yields the correct attributes. Overall, we had successful understanding of 96.1% of the utterances. Thus the remaining 3.9% will yield at least one incorrect attribute without a competitor. Because of the property that an attribute is only unreliable if there *is* a competing hypothesis that sets a different (or no) attribute value, 3.9% * 31.3% = 1.22% of all hypotheses will surely yield a *false alarm* with this method. False alarm is a situation where an attribute is marked as reliable, but is in fact incorrect. 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%) |
| basic false alarm | 4.21% | 1.22 % | 0.06% |

**Table 1.** Accuracy effects of pruning word graphs

Since the first utterance in a dialog usually is the most complex and therefore the most difficult to recognise, a pruning threshold of 30,000 will prevent the correct hypothesis from being pruned off the word graph in almost all cases for all answer types.

## 3.3 Score distance

The maximal allowed score distance from the best sentence is obviously a key parameter of this method: If it is too large then most best sentences will have competitors within the range, resulting in a high *false rejection rate*. (The false rejection rate is the proportion of correctly recognised attributes marked as unreliable.) However, a small score distance may cause an incorrect best hypothesis to be *falsely accepted*, because a competing (correct) alternative hypothesis might be out of range. Therefore, the distance parameter was experimented with to determine the false alarm vs. false rejection rates, resulting in the Receiver Operation Characteristic (ROC) curve shown in Figure 4.

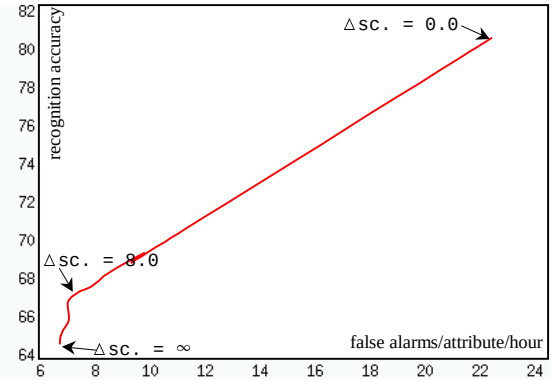


**Figure 4.** Receiver Operation Characteristic (ROC) curve.

Accuracy refers to the percentage of attributes resulting from correctly understood concepts which were marked as reliable. False alarms per attribute per hour is a measure of the false alarm rate. The point in the upper right corner concerns score distance 0.0, the baseline performance: the best sentence hypothesis $H_0$ 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. The score distance is taken as the optimal one, because from here the number of false alarms increases only marginally, while the accuracy drops substantially. The relatively low cost of a rejection (extra turn in the dialog) versus the high cost of false acceptance (possibility of confusion) makes this the best choice.

# 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 6.

| | | |
|---|---|---|
| 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 6.** Example dialogue showing implicit verification.

Theoretically the first part of a dialogue, 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 that the forced choice to reject certain parses of a path in a word graph sometimes gave undesirable results. 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 (because 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 should not be made at concept parsing level, where there is no knowledge about dialogue history. In the automatic attendant application problems with ambiguous parses never occurred. In the more complex travel information domain they seem difficult to avoid, even though the confidence measure computation has difficulty coping with ambiguities. However, the eventual impact of this problem remains to be established, if only because our experiment was confined to the processing of answers to the first question of the system, which tends to elicit the most complex answers.

Another problem of the method is related to the maximum number of hypotheses that is considered. Suppose we have the n-best sentence list shown in Figure 7:

| # | 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 7:** 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 sets of 'acoustically highly confusable' words (Delft, Elst, Best, …) that are at least as large as this maximum number of sentences. Not only will all words in these sets be marked to be unreliable most of the time, which would be justifiable, but other concepts in a sentence (Hengelo in the example), which might be found unreliable in other circumstances, may be falsely accepted because all their competitors are pushed beyond the maximum number of sentences.

The problem with confusable sets did not surface in [3, 4], probably because it is highly 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, it seems to be dependent on the application domain. One of the assumptions implicitly made by the method is that competing sentence hypotheses differ only in the values of the attributes. As a consequence, the method forces the grammar to give single parses of one sentence, because otherwise it might cause spurious competition. For a mixed-initiative application like ARISE, however, the fact that the true attribute value sometimes can only be established at the level of the dialogue control (and not during the concept parsing) requires maximal flexibility and therefore freedom of parsing. The compromises needed to meet these contradictory requirements had serious impacts on the usefulness of the confidence measures.

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], as a first attempt in that direction.

## 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.