A SEMANTICALLY-BASED CONFIDENCE MEASURE FOR SPEECH RECOGNITION

Stephen Cox and Srinandan Dasmahapatra

School of Information Systems, University of East Anglia, Norwich NR4 7TJ, U.K.

{sjc,sd}@sys.uea.ac.uk

ABSTRACT

In previous work, we have argued that it is beneficial to find confidence measures (CM's) that are not dependent on use of "side information" from a specific recogniser. Here, we extend this philosophy to include the use of semantic information in estimating the confidence that a word is correct. We are motivated by the observation that sometimes the recogniser outputs a word which can easily be spotted (by humans) as incorrect, because it bears no relation to the semantics of the rest of the decoded sentence. Latent semantic analysis (LSA) was used as a method for estimating semantic ”semantic similarity" between words in a text corpus. From these scores, an average semantic similarity of each decoded word to the other decoded words in an utterance could be estimated, and by thresholding this similarity measure, words were tagged as CORRECT or INCORRECT. We benchmarked the performance of this semantic CM against a tried-and-tested CM, the N-best CM. The precision of the semantic CM was inferior to that of N-best when the recall (the number of words considered) was high, but it out-performed N-best for low recall, and a combined classifier showed the benefits of using both techniques. An interesting and unexpected result was that the semantic CM was better at identifying correct words than incorrect words.

1. INTRODUCTION

There has recently been considerable research activity in the field of confidence estimation, for instance [3, 6, 9, 5]. There are several motivations for attaching a measure of confidence to the words output by the recogniser: it can be used to improve the efficiency of a speech dialogue understanding system by requesting confirmation or re-input only when necessary, for detection of out-of vocabulary (OOV) words, or to aid unsupervised speaker adaptation etc.

A variety of techniques for deriving confidence measures (CM's) for words has been proposed over the last few years. The most popular are measures derived from "side"-information generated by the decoder during the recognition process. Examples of these are likelihood ratios derived from a Viterbi recogniser [6], measures of competing words at a word boundary, [5]. These have sometimes been used together as a feature vector to increase performance further e.g. [2]. In a previous publication [4], we have argued that although such measures may work well for a particular recogniser, they have limited generality and our own experience has been that they can fail when implemented on a different decoder. Accordingly, we have tried to develop CM's that are less recogniser-specific. Our approach has been to attempt to decouple information from the acoustic and language-modelling components of the recogniser and then to examine this information separately and in combination to estimate a CM. This technique requires the use of a separate phone-loop recogniser (running in parallel with the word-level decoder) to provide independent information about acoustic matching.

In this paper, we discuss the use of a measure that requires neither decoder side-information nor an extra phone-loop decoder, but instead uses purely linguistic knowledge (that is independent of the linguistic knowledge in the decoder) to decide whether a decoded word is likely to be correct or not. The technique is based on the observation that humans can identify some words that are incorrect in a recogniser decoding on semantic grounds. Consider, for instance, a sentence decoded by our recogniser: "Exxon corporations said earlier this week that it replaced one hundred forty percent its violin gas production in nineteen eighty serve on", where the word "violin" is clearly incorrect because it is not semantically related to the other content words in the sentence (the correct transcription is "oil and"). Of course, not all incorrectly decoded words can be clearly identified on semantic grounds as in this example, but the occurrence of "semantic outliers" is not infrequent in recogniser decodings (see section 2 for some numbers from our own recogniser). One advantage of identifying incorrect words using this principle is that the information used to decide whether a word is a semantic outlier will be independent of any information about the quality of the acoustic matching in the decoder, and will also have some independence from measures derived from the decoder's language model. Hence it should be possible to combine the "semantic" information with other information to make an improved classifier, and in section 4.2 we describe how this was done.

The paper is organised as follows: we begin by describing a preliminary experiment in which we investigated the viability of the idea of estimating confidence in the correctness of a decoded word using semantic information. In section 3 we describe the latent semantic analysis procedure that enabled us to construct a table of "semantic scores" between words. Section 4 begins with a brief description of the data and the recogniser used and then describes the confidence measures that were derived from the semantic score measures and compares their performance with the "N-best" confidence measure. Experiments in which the "N-best" confidence measure was combined with the semantic confidence measure are also described. We conclude with a discussion and some ideas about how this idea can be extended.

2. PRELIMINARY EXPERIMENT

We examined about 600 sentences decoded by our recogniser from the WSJCAM0 corpus and, without knowing the correct transcription, attempted to mark the words that we thought were incorrect on "semantic" grounds. This marking was done conservatively i.e. only words that seemed to be clearly wrong because they were incongruous were tagged as *I* ("Incorrect"). The confusion-matrix of our hand-marking is shown in Table 1. Ta-

6th International Conference on Spoken
Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

		CLASSIFIED	
		Unclassified	Incorrect
ACTUAL	Correct	7680	49
	Incorrect	2720	421

Table 1: Confusion-matrix for hand-tagging of words on semantic grounds

ble 1 indicates that $421+49=470$ words were classified as I out of the $2720+421=3141$ that were actually incorrect i.e. *Recall* = 15%. Of the 470 words tagged as I, 421 were correctly tagged, so *Precision* = 89.6% (N.B. these definitions of “Recall” and “Precision” are used in section 4.2). This experiment indicated to us that using a “semantic” criterion to identify incorrect words had the potential to identify only a small number of words, but with fairly high precision. These words were all nouns or verbs that were incongruous and it would not be possible to identify incorrectly decoded function words. However, non-function words, in most cases, bear more information.

3. APPLICATION OF LATENT SEMANTIC ANALYSIS

3.1. Background

Latent semantic analysis (LSA) is a technique which has been in use for some years in the field of information retrieval and has latterly been applied in speech recognition [1]. It is not proposed to describe the theory of LSA in detail here—for an introduction, see [8]. LSA is a technique for associating words that tend to co-occur within documents that are “semantically coherent” (examples of documents are entries in an encyclopaedia or stories in a newspaper). The assumption is that words that tend to co-occur within the same document are semantically linked.

The essential idea behind LSA is to form a very large matrix of word/document co-occurrences and then to represent the row and column vectors of this matrix in a greatly reduced sub-space using the technique of singular value decomposition (SVD). Because the matrix is very sparse and its rank is much lower than its dimensionality, it is possible to represent the vectors in a low dimensional space with relatively small error. The key property of LSA is that words whose vectors are “close” in the reduced space correspond to semantically similar words (also, documents whose vectors are close in the space convey similar semantic meanings). Hence it is possible to form a “semantic score-matrix” between words which can be used to provide an estimate of the likelihood of two words co-occuring within the same text.

3.2. Application of LSA

The 1994 subset of the Wall Street Journal (WSJ) (available in the North American News Text Corpus (NANT)) was used to form a word/document matrix W for latent semantic analysis. A “document” was defined to be a news-story (as demarcated by <TEXT> and </TEXT> in the files) and the text was pre-processed to remove punctuation and to spell out abbreviations, numbers, dates etc. The entries in W were formed from the raw counts c_{ij} of the number of times word i appears in document j according to the technique described in [1]. In [1], entry (i,j) of W is defined as $W_{ij}=G_iL_{ij}$, where G_i is a “global weight” for word w_i and reflects the fact that words occur with different frequencies through documents, and L_{ij} a local term that adjusts c_{ij} to take account of widely different values. There were 19685 different words and 19396 documents in the corpus and SVD of W was done using the MATLAB routine `svds()`. After some experimentation with different dimensionalities of reduced space, a reduced space dimensionality of 100 was used. Each word was then described by a 100-d vector, and the “semantic score” between words w_m and w_n in the lexicon, $S(m,n)$, was computed as the cosine of the angle between the vectors i.e.

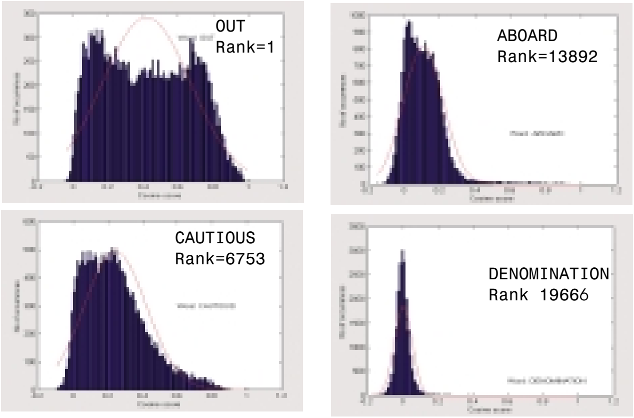

Figure 1: Distributions of “semantic scores” for four words

$$S(w_m,w_n)=\frac{\boldsymbol{\omega}_m\cdot\boldsymbol{\omega}_n}{|\boldsymbol{\omega}_m||\boldsymbol{\omega}_n|}, \tag{1}$$

where $\boldsymbol{\omega}_m$ is the 100-d vector associated with word w_m. It should be noted that $-1\le S(m,n)\le 1$ and a high positive value for $S(m,n)$ indicates that the words w_m and w_n have a high semantic correlation.

Before proceeding to use these scores for estimation of confidence measures, we were interested to examine their distributions for different words. The score between each word in the lexicon and all other words was computed to give a set of scores $\{L_i\}$ for word w_i. The mean score $\overline{L_i}$ was computed for each word, and the words were then ranked by their value of $\overline{L_i}$. It was noticeable that the words with high values of $\overline{L_i}$ were predominantly (but not exclusively) function words. The words with the lowest values were, without exception, rarely occurring nouns. Figure 1 shows the distribution of the $\{L_i\}$ for four words at different positions in the ranking. The implication of Figure 1 is that broad distributions are associated with words that occur with many other words and hence have a broad range of semantic scores, whereas the narrower distributions centred on zero represent words that occur with only a small fixed set of words and hence have a semantic score close to zero to most other words in the lexicon.

It seemed possible that these “semantic scores” were actually simply a reflection of the fact that a word that occurred often in the training data would naturally co-occur with many other words and hence have a high mean score, whereas the reverse would be true for infrequently occurring words. However, it was found that there was only a very weak correlation between the mean semantic score and the number of occurrences (or the ranking of the number of occurrences) of a word ($|r|=0.15$ for the latter case). For instance, the word “proving” has a medium number of occurrences in the training-data (91, rank 6992) but a high mean semantic score of 0.385 (rank 165), which indicates that it co-occurs with a diverse collection of words. Conversely, “gas” has quite a high number of occurrences in the training-data (1946, rank 579), but a very low mean semantic score of 0.02 (rank 19081), which indicates that it co-occurs only with a very specific set of other words.

4. CONFIDENCE MEASURES FROM LSA

The motivation for estimating semantic scores for each word in the lexicon was the possibility of identifying “semantically incorrect” decoded words i.e. words whose meaning and usage were not cognate with the other decoded words in the sentence. We first attempted to identify such words by using the pre-computed semantic scores to compute the mean semantic score for each decoded word. Suppose that the sequence of words decoded from an input utterance is $w_{U(1)}, w_{U(2)}, \ldots, w_{U(N)}$, where $U()$ maps from the number of the decoded word in the sentence to the number of the same word in the lexicon. The mean semantic score for the i'th decoded word is

$$MSS_i = \frac{1}{N}\sum_{j=1}^{N} S(U(i),U(j)). \qquad (2)$$

(Notice that if word w_i is the same as word w_j, $S(U(i),U(j)) = 1$, which increases the value of MSS_i. In practice, only common function words usually re-occur in a decoding, and as these words will be discarded after the application of a threshold (see section 4.1), this effect does not cause a problem.) We would expect *MSS* to be low for semantically incorrect words and high for words that are cognate. However, *MSS* is a poor indicator of semantically incorrect words. The reason for its poor performance is that most words have high scores to function words, and although a semantically incorrect word may have low scores to other content words in the decoded sentence, these low scores are masked by the “noise” from the higher scores to decoded function words.

4.1. Using a stop list to eliminate common words

This result suggested that it was necessary to discard decoded words that had high mean semantic scores to most other words in the lexicon, as these words had low semantic weight and contributed mainly noise to the value of *MSS* for other words. Accordingly, we experimented with discarding all decoded words whose value of $\underline{L}_i$ was above a threshold L_T—the list of discarded words is sometimes known as a *stop list* [7]. We then compute a confidence measure for each of the remaining words. The confidence measures we experimented with were as follows:

1. *MSS* as defined in equation 2
2. *MR*, the mean rank of the semantic scores to the decoded word w_i. MR_i was computed by finding the rank of each $S(U(i),U(j)$ in the set of semantic scores $\{L_i\}$ and then computing the mean.
3. *PSS*, the probability that the set of semantic scores from word w_i to the other decoded words was generated from the distribution of scores $\{L_i\}$. $PSS_i = \prod \Pr(L_i \leq S(U(i),U(j)))$, where L_i is a random variable whose distribution is estimated from $\{L_i\}$, the set of semantic scores for word w_i. We approximated the distributions shown in Figure 1 by 5 component Gaussian mixtures to estimate $\Pr(L_i \leq S(U(i),U(j)))$.

In practice, we found that all three of the above statistics gave very similar performance, with *PSS* marginally the best. The effect of varying L_T on the tagging accuracy of *PSS* is shown in Table 2.

In table 2, e_{Rec} is the error-rate of the recogniser on the retained words and e_{CM} is the tagging error-rate (i.e. the error-rate of tagging words as ’C’ or ’I’) when using the confidence-measure. A “guessing” confidence measure would have $e_{Rec} = e_{CM}$, and the final column gives the improvement of using *PSS* as a confidence-measure over guessing. It is interesting that e_{Rec} at first decreases, probably as more function-like words (which tend to have a higher error-rate than function words) are discarded. However, when 78% of words are discarded, e_{Rec} begins to rise again and continues to rise. Examination of the words retained when $L_T \leq 0.15$ shows a greatly increased proportion of single letter words, mostly “u”, “p” and “s”. These words are almost always incorrect insertions by the decoder, and hence the error-rate increases.

Threshold L_T	% words discarded	e_{Rec}	e_{CM}	% improvement
0.45	0	0.303	0.288	5.0
0.4	5	0.296	0.282	4.7
0.35	25.7	0.286	0.274	4.2
0.3	45.9	0.260	0.247	5.0
0.25	53.6	0.238	0.221	7.1
0.2	64.6	0.230	0.222	3.4
0.15	78	0.243	0.242	0.4
0.1	88	0.276	0.285	-3.2
0.05	97	0.296	0.294	0.7

Table 2: Effect of varying the threshold L_T

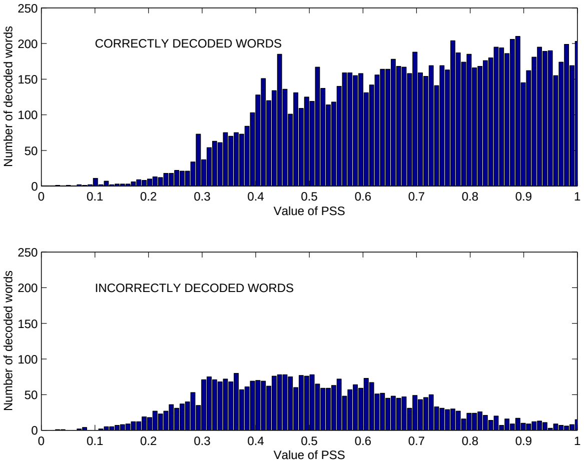

Figure 2: Distributions of values of *PSS* for correct and incorrect words

Table 2 does not show the predicted increase in tagging accuracy when commonly co-occuring words are eliminated. An examination of the distribution of the values of *PSS* was revealing. In Figure 2, the values of *PSS* for correctly (top) and incorrectly (bottom) decoded words are shown for $L_T = 0.25$ i.e. with 53.6% of decoded words excluded. The histograms have a large overlap showing that *PSS* is unable to separate these classes very effectively. However, it is interesting to examine the “tails” of the distributions. Our hypothesis is that a semantic confidence measure should be effective at identifying incorrect words, and so we would expect to see a high probability of low values of *PSS* for *incorrect* words, and a low probability of low values of *PSS* for *correct* words. In fact Figure 2 shows that the probabilities of low values of *PSS* are very similar for both correct and incorrect words. However, *high* values of *PSS* are significantly more probable for correct words than for incorrect words. Hence *PSS* is deriving any discrimination it has by identifying *correctly* decoded words. Analysis revealed that the words associated with high values of *PSS* were predominantly words that commonly occurred in the WSJ data (numbers, financial terms etc.) that were highly cognate with each other. Inspection of the decoded words

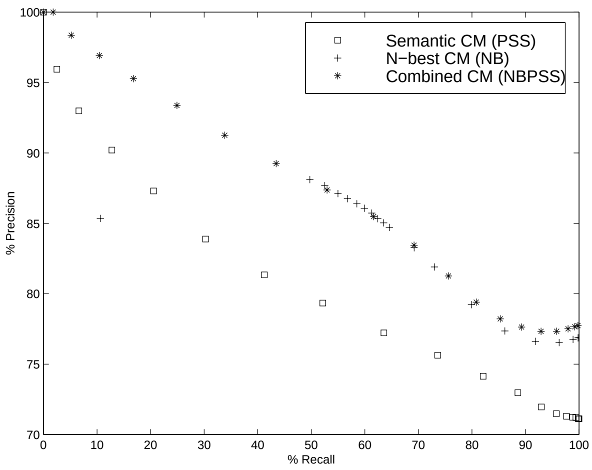

Figure 3: Recall/precision curves for *PSS*, *NB* and *NBPSS*

that had very low values of *PSS* associated with them showed that some of these were very common words that had been correctly decoded. It is possible that the corpus used for making the LSA analysis does not have enough material to capture the large set of words that these common words co-occur with, and the decoded utterances in the test-set contain previously unseen co-occurrences that lead to a low semantic score for these words.

4.2. Combining semantic and *N*-best confidence measures

The semantic confidence-measure based on the value of *PSS* is a weak indicator of correctly or incorrectly decoded words, but the information it provides is largely independent of similar information from the decoder itself. It seemed possible that combining *PSS* with a CM derived directly from the decoder would be useful. We used the *N*-best CM (*NB*) to provide this. In our formulation of this CM, the value of *NB* for the *i*'th decoded word in the top decoding w_i is estimated by noting the proportion of times that w_i occurs in the same position in the top *N* decodings (we used $N = 100$). If we assume that NB_i and PSS_i, the values of *NB* and *PSS* for word w_i, are independent estimates of the probability that word w_i is correct, the product of these values ($NBPSS_i$) gives a further estimate of this probability. Figure 3 shows receiver operating curves formed from the three CM's *NB*, *PSS* and *NBPSS* by varying the threshold above which utterances were classified as correct (N.B. this was run on *all* decoded utterances). Figure 3 shows that *PSS* (squares) is generally a poor indicator of the status of a decoded word: if all decoded words are examined, its performance is no better than chance, but as the proportion of decoded words examined drops, its accuracy increases to close to 100%. *NB* (crosses) is better for high values of recall, but the maximum precision *NB* is capable of is 87.5% at a recall of about 50%. The single *NB* point to the left of this point is made from the values of *NB* that are exactly 1.0 (i.e. words that occur in all top 100 decodings), and the proportion of such words that are actually correct is about 85.5%. When *PSS* is combined with *NB* to form *NBPSS* (asterisks), performance with high recall is slightly better than *NB* alone *NBPSS* retains the ability of *PSS* to give high precision for low recall. This is useful if it is desired to identify a small number of decoded words as correct with a high confidence.

5. DISCUSSION

This experiment has shown that using information about how well a decoded word relates semantically with the other decoded words in an utterance can provide useful information about whether the word is correct. This information is largely independent of information derived from a "side information" based confidence measure, such as *N*-best, and hence complements the latter. Although our original motivation was that a semantically-based CM should be able to identify incorrectly decoded words that were not cognate with the other decoded words in an utterance, we found unexpectedly that it was better at predicting correctly rather than incorrectly decoded words. This seems to be due to two effects: firstly, the presence of a small set of commonly occurring words in the WSJ data that are highly cognate (e.g. numbers, financial terms): secondly, the fact that it is impossible during training to capture all contexts in which very common function words occur.

One problem with the LSA approach is the very large number of words that are typically encountered in training. A possible way of countering this is to use *stemming*, where morphological variants of the same word are collapsed into a single root form. This would lower the number of different words in the training-corpus and hence increase the effective training-set size. Finally, although LSA is a convenient tool for examining semantic relations between words, there are other techniques which may be equally powerful and which are worthy of investigation.

ACKNOWLEDGMENT

This work was funded by a grant from the UK Engineering and Physical Sciences Research Council.

6. REFERENCES

[1] J.R. Bellegarda. A multispan language modeling framework for large vocabulary speech recognition. *IEEE Transactions on Speech and Audio Processing*, 6(5):456–467, September 1998.

[2] M.C. Benitez et al. Word verification using confidence measures in speech recognition. In *Proc. 5th International Conference on Speech Communication and Technology*, pages 1082–1085, November 1998.

[3] L. Chase. Word and acoustic confidence annotation for large vocabulary speech recognition. In *Proc. 5th European Conference on Speech Communication and Technology*, pages 815–818, September 1997.

[4] S.J. Cox and S. Dasmahapatra. A high-level approach to confidence estimation in speech recognition. In *Proc. 6th European Conf. on Speech Communication and Technology*, pages 41–44, September 1999.

[5] S.J. Cox and R.C. Rose. Confidence measures for the SWITCHBOARD database. In *Proc. IEEE Conf. on Acoustics, Speech and Signal-processing*, pages 511–515, 1996.

[6] L. Gillick, Y. Ito, and J Young. A probabilistic approach to confidence estimation and evaluation. In *Proc. IEEE Conf. on Acoustics, Speech and Signal-processing*, April 1997.

[7] D. Jurafsky and J.H. Martin. *Speech and Language Processing*. Prentice Hall, 2000.

[8] T.K. Landauer and S.T. Dumais. A solution to Plato's problem: representation of knowledge. *Psychological Review*, 104:211–240, 1997.

[9] T. Schaaf and T Kemp. Confidence measures for spontaneous speech recognition. In *Proc. IEEE Conf. on Acoustics, Speech and Signal-processing*, April 1997.