LARGE-SCALE LEXICAL SEMANTICS FOR SPEECH RECOGNITION SUPPORT

George Demetriou, Eric Atwell & Clive Souter
Centre for Computer Analysis of Language And Speech (CCALAS)
& Artificial Intelligence Division, School of Computer Studies
University of Leeds, Leeds LS2 9JT, UK
Tel. +44 113 233 6827, FAX: +44 113 233 5468, e-mail: george@scs.leeds.ac.uk

ABSTRACT

This paper presents a study on the use of wide-coverage semantic knowledge for large vocabulary (theoretically unrestricted) domain-independent speech recognition. A machine readable dictionary was used to provide the semantic information about the words and a semantic model was developed based on the conceptual association between words as computed directly from the textual representations of their meanings. The findings of our research suggest that the model is capable of capturing phenomena of semantic associativity or connectivity between words in texts and considerably reducing the semantic ambiguity in natural language. The model can cover both short and long-distance semantic relationships between words and has shown signs of robustness across various text genres. Experiments with simulated speech recognition hypotheses indicate that the model can efficiently be used to reduce the word error rates when applied to word lattices or N-best sentence hypotheses.

1. INTRODUCTION

Despite recent achievements in continuous speech recognition, the correct transcription of large vocabulary spoken input by computers remains very difficult, if not impossible, without the proper analysis, modelling and application of linguistic knowledge in the form of syntactic or semantic constraints. So far, stochastic language modelling in the form of N-grams ([1],[6]) has been the most 'standard' approach in dramatically reducing the word error rates of speech recognition systems. But as current trends move to very-large vocabulary recognition, enormous amounts of training texts and sophisticated 'smoothing' techniques are needed to overcome the sparse data problem and provide accurate statistics about N-gram probabilities in natural language. In addition, the need for more general language models which can capture both short and long distance dependencies between the words (something that N-grams often fail to do) as well as demonstrate decent performances across a variety of language domains or genres has become apparent over the last years.

In this paper, the use of wide-coverage semantic knowledge for the improvement of performance of speech recognizers is investigated. Approaches that have made use of semantic information in the past were intended for small to medium vocabulary systems and primarily to assist understanding. Examples of such approaches include semantic networks ([8]), semantic grammars ([13]), case frames ([4]), statistically-based approaches ([10]), unification-based approaches ([13]) and neural networks ([5]).

The work presented in this paper distinguishes itself in that it uses large-scale semantic information which can be fully acquired from reusable language resources such as machine readable dictionaries (MRDs) without the need for hand-coding or training procedures. In practice, it would be impossible to develop a model for all semantic dependencies between all words language in all probable contexts. Semantic knowledge in MRDs, however weak or incomplete, can in theory provide semantic constraints about all words in the language in a way that is very economical.

2. MEANING REPRESENTATIONS AND SEMANTIC ASSOCIATIVITY

The meanings of the words in language were acquired from the Longman Dictionary of Contemporary English (LDOCE), 1978 edition. The dictionary includes semantic information about 36,000 distinct root forms (lemmas) which can cover at least 80,000 words in English language (when inflected and derived forms are taken into account[1]). Three kinds of lexical semantic knowledge are provided in the dictionary:

- knowledge about the meanings of the words; a word meaning is represented by a limited set of conceptual attributes or primitives
- knowledge about the selection restrictions or preferences certain classes of words (verbs, nouns, and adjectives) have in language; such restrictions are realised via a semantic hierarchy of nouns which are classified into 35 categories
- knowledge about the specific discourse or domain a word can occur; such knowledge is expressed with a set of subject categories (125 main subject categories and 212 subdivisions)

To utilise the dictionary information for our purposes, all the individual meanings (sense definitions in the dictionary) of a

[1] This estimation was made with the use of a separate lemmatisation dictionary compiled from a number of reusable language resources.

word were merged into a single meaning representation. The end result was a meaning definition specified by a set of semantic primitives (including the codes about the selectional restriction and subject domain information). Duplicate semantic primitives were eliminated and a list of 36 very common function words (such as "a", "the", "of" etc.) were excluded from the meaning definition.

2.1 Computation of the Semantic Association

Having transformed the dictionary into a semantic knowledge base, a simple way to specify semantic affinities between two words is to find the conceptual overlap between their meanings as estimated by the number of semantic bonds or links the meanings have in common. This is computed with the use of a semantic association function which is defined as follows:

Let *x* and *y be* two words in language whose meanings are represented by the sets *X* and *Y* respectively. The *semantic association* between *x* and *y* is given by[2]

$$S(x,y) = m\{X \cup Y\} / m\{X \cap Y\} \quad (1)$$

i.e. the semantic association between the words *x* and *y* is the number of semantic primitives or attributes the meanings of *x* and *y* have in common divided by the total number of distinct attributes of these two meanings.

When a semantic association function is to operate upon larger word groupings such as phrases or N-best sentence hypotheses, the notion of pairwise semantic associations between words is extended to account for the latent semantic "activity" of the whole word sequence. The *semantic activity* of a word sequence *L* consisting of elements *L1, L2,..* is defined as

$$SA(L) = \frac{1}{k} \sum_{i,j} S(L_i, L_j) \quad i \neq j \quad (2)$$

where *S* is the semantic association function as defined in *(1)* and *k* is a normalisation factor. A typical value of *k* is $C_2^n = (n-1)n/2$ (i.e. the semantic activity of a word string can be interpreted as the average semantic association between pairwise word combinations in the string).

2.2 Estimating the Satisfiability and Constraint of Semantic Associations

Several experiments were conducted to assess the satisfiability of semantic associations and the constraint implied by the model. Due to space limitations they will only very briefly be mentioned here ([3] for more). It was found that the semantic associations between word combinations in natural language texts (3250 sentences taken from the British National Corpus - BNC) are significantly different than those expected when words are chosen at random from the dictionary at much better than 0.001 confidence level. The relative reduction in uncertainty in discriminating between "text" and "random" word combinations from the values of the semantic associations was better than 0.25. For a vocabulary of 80,112 words, the redundancy ([11]) of the semantic model was found 45.9 and the associated perplexity 446. Note that these figures were computed for single word pairs only i.e. when the prediction of a word depends on the semantic association with the previous word in the text. In the experiments with speech recognition hypotheses as described below, the semantic constraint was considerably stronger due to the fact that the semantic associations were computed between all pairwise combinations of words from beginning to end of the utterance.

2.3 Semantic Associations and Context Distance

One important finding of the experiments was that, contrary to some linguistic theories or studies that argue for local meaning relations between words, the semantic associations as captured by this model extend to much larger distances than expected and certainly beyond sentence boundaries. The results of the experiments suggest that the semantic associativity between two words is still much higher than average (i.e. expected at random) at distances larger than 100 words.

3. SEMANTIC ACTIVITY AND SPEECH RECOGNITION

To test the efficiency of the semantic association model for speech recognition, a large vocabulary speech recognition front-end was simulated with the use of phoneme confusion data from a speech recognizer ([7]) and a pronunciation lexicon 70,646 words. The model was tested on both N-best sentence hypotheses and word lattices (graphs) for an input text of 650 sentences (14491 words) selected from the BNC in 13 sets of 50 sentences, each set from a different BNC genre (*written text*: leisure, social science, world affairs, arts, imaginative, applied science, natural science, commerce/finance, belief/thought - *spoken text*: leisure, educative/informative, public/institutional, business). For each N-best sentence hypothesis a *semantic activity score* was computed and the hypothesis with the best score was assumed the most likely one.

The results for a total number of 637,176 N-best sentence hypotheses tested with the model suggest a considerable reduction of word error rates for a wide spectrum of recognition rates (fig. 1). Even when the baseline recognition accuracy without the application of the model was about 92% words correct, improvements were observed from 0.6 (for the written leisure text) to

[2] The notation m{ } is used to denote the number of elements in the expression within m{ }.

2.9% words correct (for the written belief/thought text). There was a variation in the recognition performance from genre to genre but not as much to suggest that the model is biased towards a particular genre.

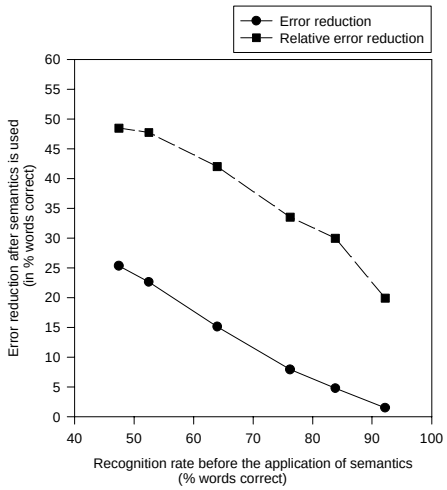

Fig. 1*: Error reduction and relative error reduction for different baseline accuracies*

3.1 Semantic Activity and Word Recognition

The semantic activity scores for the N-best hypotheses were correlated with the corresponding recognition scores. It was found that there exists positive correlation between semantic activity and word recognition accuracy at better than 0.01 significance level (Pearson correlation coefficient=0.33961 for 81 paired observations of grouped recognition scores). This indicates that the semantic associativity (as measured by this method) and the word correctness go hand-in-hand and the model is capable of consistently discriminating between more and less accurate hypotheses.

3.2 Error Reduction and Amount of Context

Since the semantic associations between words were computed from the beginning to the end of the utterance, it would be logical to assume that the longer the input utterance, the higher the probability of having semantically related words in the context and the higher the semantic activity. It was found that, the more the context that is taken into account, the larger the error reduction.

For short utterances (<5 words), there was observed a decrease in recognition performance on the average, which was due to the fact that short utterances contain very few content words so that it was more difficult for the algorithm to pick up meaning. For longer utterances, the model performed quite well with the error reductions ranging from 5 to about 30% and there was a clear correlation between context width and error reduction (Pearson correlation coefficient=0.834 for 15 pairs of grouped observations).

3.3 Inclusion of the Correct Hypothesis within the Top N%

It would be useful to know the probability of having the correct sentence hypothesis within the top N% semantically scored hypotheses. The inclusion rate of the correct hypothesis is shown in fig. 2.

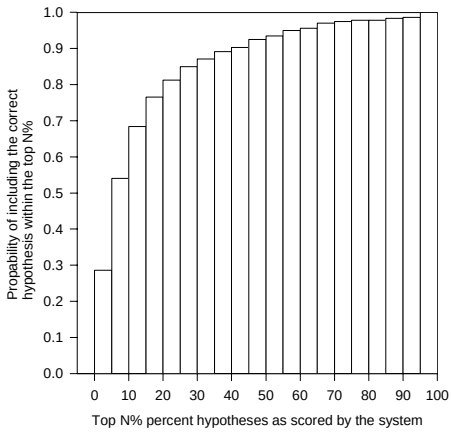

Fig. 2: *Probability of including the correct hypothesis within the top N%*

As it can be seen from fig. 2, the probability of having the correct hypothesis within the top 25% is larger than 80%. This indicates that when the set of N-best sentence hypotheses is not too large, there is a high degree of confidence that the semantic model would provide the correct hypothesis within a list of manageable size (the top N% values for 90% and 95% probability are 42.5% and 53.2% respectively).

3.4 Semantic Associativity and Lattice Parsing

To test the efficiency of the model for parsing word lattices or graphs, 650 lattices were produced. The average number of different paths through the lattices was estimated to be 1.38×10^{20}. With so many possible paths, the use of an exhaustive search procedure was out of the question and two algorithms for more efficient lattice search were developed. The first algorithm, called the Meaning-Driven Search Algorithm (MDSA), is based on the *uniform-cost* principle ([9]) and performs a kind of Dynamic Programming lattice search. The MDSA makes use of the history of recognised words so that the recognition of the next word in the graph depends on its semantic associativity with the previously recognised words (within the same utterance). Although the algorithm was found capable of increasing the recognition performance from about 15% to about 40%

on the average, it was found that often the correct path was lost within the first few stages of the search. When that happened, it was impossible for the algorithm to recover from errors later on during parsing. For this reason, a second algorithm, the Meaning Driven Look Ahead Search Algorithm (MDLASA) was designed.

Text sample	Recognition rate (words % correct)		
	Baseline	MDSA	MDLASA
wri_leisure	12.5	36.2	44.2
wri_soc_science	15.6	41.1	60.1
wri_world_affairs	16.3	51.9	60.5
wri_arts	16.8	50.8	61.7
wri_imaginative	16.3	43.9	52.7
wri_app_science	14.3	37.5	49.7
wri_commerce/finan.	14.9	43.4	48.9
wri_theology	14.8	42.9	43.9
spo_business	13.1	41.8	37.5
spo_public/institut.	15.6	38.8	50.6
spo_leisure	15.5	47.4	54.3
spo_educative/infor.	17.6	46.2	62.4
wri_nat_science	14.0	39.7	52.5
ALL	15.3	42.7	53.8

Table 1: *Performance of the meaning driven search algorithms*

The MDLASA uses a heuristic function to evaluate the promise of a node with respect to the concepts that may be encountered next in the graph. During the initial stages of the search, when the context is short, the algorithm relies more on the promise rather than the semantic score for a partial path. In the later stages of the search, when the largest part of a complete path has been determined ,the algorithm relies more on the semantic associations between the already recognised rather than the promise of the path. The results of the MDLASA indicate a relative increase of about 26% (from 42.7 to 53.8% words correct) on the average over the MDSA (see table 1).

4. CONCLUSIONS

We have presented a linguistic model for wide coverage, very-large vocabulary speech recognition support. The model is based on lexical semantic knowledge from an online dictionary and was found capable of capturing phenomena of semantic associativity between words in natural language and efficient in reducing the word error rates in large vocabulary recognition tasks. In summary, the advantages of the model are:

- very large vocabulary
- it requires no training and it is economical and easy to implement (the semantic associations between the words need not be stored - they can be computed in runtime)
- it can effectively model long-distance semantic relationships between words; as such it could also be used as a cache-based model
- no sense disambiguation or language understanding are needed; only the meaning associations are essential
- in theory, depending on the particular implementation, the model could also be used to provide top-down control of search

It is not one of the intentions of this paper to compare this model with other more established language models in the area of large vocabulary speech recognition, such as N-grams. It is rather the case that due to the distinctive differences between the two approaches (qualitative vs. quantitative information, long-distance vs. short-distance constraints, domain-independence vs. domain-dependence) one could see them as complementary. For example, a bigram or trigram language model could be used to model local constraints between the words and the semantic association model could provide clues about more global semantic dependencies.

5. REFERENCES

[1] Bahl L. R., Jelinek F. and Mercer R. (1983) A Maximum Likelihood Approach to Continuous Speech Recognition. In: IEEE Trans on PAMI, vol. 5, 179-190.

[2] Demetriou G. C. (1993), Lexical Disambiguation Using CHIP. In: EACL'93, Utrecht, 431-436.

[3] Demetriou G. C. (1997), Lexical Semantic Information Processing for Large Vocabulary Human-Computer Speech Communication, PhD Thesis, Artificial Intelligence Division, School of Computer Studies, University of Leeds.

[4] Hayes P. J., Hauptmann A., Carbonell J.and Tomita M. (1986), Parsing Spoken Language: a Semantic Caseframe Approach. In: COLING-86,588-592.

[5] Jain A. N. and Waibel A. H. (1990), Robust Connectionist Parsing of Spoken Language. In: ICASSP-90, 593-596.

[6] Jelinek F. (1990), Self-Organised Language Modeling for Speech Recognition. In: Readings in Speech Recognition, A. Waibel and Kai-Fu Lee (eds).

[7] McInnes F. R., McKelvie D. And Hiller S. M. (1990), The Structure, Strategy and Performance of a Modular Continuous Speech Recognition System. In: Proceedings of the Institute of Acoustics, vol. 12, part 10, 173-180.

[8] Niedermair G. T., M. Streit and H. Tropf (1990), Linguistic Processing Related to Speech Understanding in SPICOS II. In: Speech Communication 9, 565-585.

[9] Nilsson N. (1971), Problem-Solving in Artificial Intelligence, McGraw-Hill.

[10] Pieraccini R. and E. Levin (1992), Stochastic representation for semantic structure for speech understanding. In: Speech Communication 11, 283-288.

[11] Shannon C. E. (1948), A Mathematical Theory of Communication. In: Bell System Tech. Journal, vol. 27. no. 3 , 379-423 & 623-656.

[12] Thurmair (1988), Semantic Processing in Speech Understanding. In: Recent Advances in Speech Understanding, H. Niemann, M. Lang and G. Sagerer (eds), NATO ASI Series, vol. F46, Springer Verlag, 397-419.

[13] Tomabechi H. and M. Tomita (1988), The Integration of Unification-based Syntax/Semantics and Memory-based Pragmatics for Real-Time Understanding of Noisy Continuous Speech Input. In AAAI-88, vol 2, 724-728.