Proceedings of the Institute of Acoustics

MAKING USE OF SEMANTICS IN AN AUTOMATIC SPEECH RECOGNITION SYSTEM

S Short* & R J Collingham

University of Durham, Department of Computer Science, NLE Laboratory

1 INTRODUCTION

Automatic speech-to-phoneme recognition systems are unable to achieve a 100% recognition rate. Software has devised to build a word lattice, which maps strings of phonemes to patterns of the most similar words, or test-hypotheses. However every phoneme string may correspond to more than one test-hypothesis. For instance, the sentence "I recognise speech" has a similar string of phonemes to "I wreck a nice beach". This paper presents a method to determine which hypothesis should be chosen using semantic information.

2 THE AURAID SPEECH RECOGNITION SYSTEM

AURAID was developed at Durham University. Unlike other systems, it does not constrain the input's perplexity or domain. It makes use of an existing continuous speech phoneme recognition system as a front-end to a word recognition sub-system. The sub-system generates a lattice of word hypotheses using dynamic programming with robust parameter estimation obtained using evolutionary programming. Sentence hypotheses are obtained by parsing the word lattice using a beam search and contributing knowledge consisting of anti-grammar rules, that check the syntactic incorrectness of word sequences. and word frequency information. The system is described in [2].

Conventional parsing techniques can not be used for spontaneous speech because of the errors it contains such as repair and filled pauses. Moreover they are computationally expensive, and so would not be appropriate to use on the large search space contained within a word lattice. A statistical language model could be used to constrain this search space, but the way in which it would choose words would depend on the representativity and coverage of the corpora used to derive its expectation values. This is likely to increase the domain dependency.

3 WEAK SEMANTIC SELECTION

3.1 Semantic selection

Semantic selection is the use of the meaning of concepts to prune impossible interpretations of a possibly ambiguous input. An example is "green ideas sleep":

- The adjective 'green' cannot be applied to the noun 'idea', as it only applies to concrete concepts, and the latter is abstract.
- The verb 'sleep' requires an animate subject, which 'idea' is not.

*This research has been funded by Siemens Plessey and EPSRC

The semantic analyser used in LOLITA needs to have the grammatical relations between the words of the input sentences rendered explicit to build its semantic representation. This is achieved by a parser and a full complex grammar. Only then is the semantic representation checked for semantic selection errors. In the case of speech it is impossible to perform a full grammatical analysis, therefore a direct application of this approach is unfeasible.

3.2 Weak semantic selection

Just as AURAID could not include a grammar, so uses instead an anti-grammar, it cannot include full semantic selection, but includes weak semantic selection.

The basis of this heuristic is a pair of observations about English:

- Adjectives tend to precede the noun to which they are to be applied
- The subject and objects of a verb tend to be near it. Moreover the subject tends to preceed it, and the object to follow it.

'proximity' needs to be defined. For our purposes it is sufficient to say that it corresponds to inclusion in the series of words between the previous verb, and the next.

It is however possible to find exceptions to these heuristics, such as "the cat is fat". Thus they can only assign penalties to the test-hypotheses, rather than rejecting them. Moreover because the semantic selection algorithm does not have access to a complete grammatical parse, any noun near the verb is a candidate subject or object. Thus many sentences that would be rejected if the real noun were known, will be accepted by the weak semantic selection. For instance: "The man's camel owns a house" should be rejected because a camel cannot own anything, but the presence of the man will confuse the algorithm.

The anti-grammar of AURAID is fed a tagged input, which expresses the grammatical nature of each word of each test hypothesis. These tags must be used by the weak semantic selection to ensure it uses the same grammatical interpretation of grammatically ambiguous words (such as live, which can be both an adjective and a verb) as the anti-grammar. This information is used to determine which semantic selection rule should be applied (for adjectives, or verbs). It is also used determine the boundaries of the possible domain of the subject and object of each verb.

A special treatment is accorded to pronouns: they are replaced by the most general concept for which they can stand. For instance 'she' will refer to any female entity.

3.3 Conclusion

Although the weak semantic selection can provide useful penalties, allowing the test-hypotheses to be given an order of preference not only dependent on grammatical and word frequency sources of information, it is not a sufficiently strong heuristic to deal with all the cases it will encounter. However a more detailed semantic analysis is possible within the constraints considered. This will be used in conjunction with weak semantic selection.

4 SEMANTIC CONTEXT

4.1 Semantic distance

Semantic distance is the term used for a set of properties of concepts. These properties are derived by plausible reasoning techniques and express forms of similarities between the meanings of con-

cepts. Two such properties will be defined and described. At this stage it should be noted that this paper belongs to the field of Artificial Intelligence, which we define as the "simulation of successful human behaviour". From this viewpoint, the meaning of a concept corresponds to the behaviour it produces of the agent who uses it. The human behaviour (or property) we wish to model is the recognition of associativity.

Associativity corresponds to the number and specificity of any contexts in which two or more concepts may occur. One such context is that of things related to civilian airplanes. This is the context which gives a high degree of associativity to planes, runways, airports, and air-hostesses. This context is less specific than that of all the things my stick-insects get up to, and would therefore have a lower degree of associativity.

Definition: *Associativity is a qualitative and quantitative measure of the extent and frequence of two concepts belonging to a common context.*

4.2 Requirements

Proposition 4.1 *The semantic context algorithm must be fast to deliver results in real time.*
Argument: The semantic context algorithm must be fast to compensate for the lower degree of reliability of its results (these will be based on a less detailed analysis) than those provided by the complete analysis. Moreover its whole raison d'être is to be provide a fast alternative to the slow yet complete process of grammatical and semantic analyses of the test-hypotheses generated by AURAID. In this case speed refers to low complexity of the algorithm.

Proposition 4.2 *The knowledge base must not be specially constructed for this purpose.*
Argument: AURAID is to be a domain independent system. To produce specially adapted data to choose test-hypotheses would be expensive for the large scale envisaged. Moreover it would require knowing precisely what information would be needed for that purpose before setting out on it. The rest of this argument is analogous to that provided for **Proposition 3.1** in [1].

The constraints of speed, and yet generality of the knowledge base imply that the algorithm must be able to identify the information useful to its purpose within the context of large amounts of extraneous and irrelevant information.
Proposition 4.3 *The information must be structured in a representation allowing fast access to the relevant part using syntactic information. (see* **Proposition 3.2** *in [1]).*

4.3 LOLITA, the knowledge base

A knowledge base exists which has the required features: LOLITA. It is a large natural language system which has been used successfully for a wide range of applications, including dialogue analysis and query answering. Moreover it has an added advantage for us: there is on-call local expertise as it was created in the same laboratory. More details on the representation can be found in [1].

An important feature of LOLITA to stress is that every word is mapped to one of a large number of separate meanings, rather than reduced to a generic concept. Thus "to drink" and "to eat" each may have many meanings, rather than being mapped to a single concept of "ingesting". This is vital to obtain the associativity and similarity properties.

4.4 Reminder of two semantic distance properties

Two important semantic distance properties were defined in [1] and will be used in this paper.

Definition: *Specificity expresses how precisely a property of a concept can identify it from all other concepts.*

Proposition 4.4 *The specificity of a property $\mathcal{P}$ with respect to an entity concept $\mathcal{C}$ depends on the reduction of search space required to find $\mathcal{C}$ among all entity concepts when $\mathcal{P}$ is known with respect to that when $\mathcal{P}$ is not known.*

Definition: *Similarity is a measure of the interchangeability of two concepts in general*
Similarity corresponds to what extent a concept can be used instead of another. This is a general measure of similarity, and does not restrict itself to particular purposes, such as whether books are similar to stones when thrown at people in that they have a similar effect.

4.5 Properties of paths

In this paper we shall consider a reduced form of specificity. For the rest of this paper we shall use the following: Let the (factual) event (or relation) E have arcs $\mathcal{A}_n$ with respective targets $\mathcal{T}_n$. The targets of $\mathcal{T}_n$ can be chosen from its target range $\mathcal{TR}_n$, which is the target of E's prototype event arc $\mathcal{A}_n$.

Proposition 4.5 *The specificity of a concept x with respect to the event E and the arc $\mathcal{A}_y$ by which it is to be connected to E depends on:*

$$\frac{card(intersect(\mathcal{X}, \mathcal{TR}_y))}{card(\mathcal{TR}_y)}$$

where $\mathcal{X}$ is the set of x and all its children, and
$intersect(\mathcal{X}, \mathcal{TR}_y) =$ if $x \notin \mathcal{TR}_y$ then $\emptyset$ else $\mathcal{X} \cap \mathcal{TR}_y$.

Argument: If the property E is known to be joined by the arc $\mathcal{A}_y$ to an unknown concept a, then the specificity expresses the likelihood that $a = x$. If x is not included in $\mathcal{TR}_y$, this cannot be true, so the specificity must be 0. The reason all the children of x are included is that they form the part of the search space corresponding to x, whereas $\mathcal{TR}_y - \mathcal{X}$ forms the search space corresponding to all concepts which are not an x. Therefore the formula corresponds to the ratio of the search space corresponding to x versus the total search space expressed by $\mathcal{TR}_y$. If $\mathcal{X} = \mathcal{TR}_y$ the ratio is 1, ie maximum specificity. This ratio of search spaces corresponds to how precisely the property E of the concept x can identify it from all other concepts, which is the definition of specificity.

This form of specificity is reduced in that it only considers prototypical events, although richer sources of information exist in the net.

Proposition 4.6 *The likelihood the context changes when the event E is traversed depends on the specificity of the concept a wrt E and arc $\mathcal{A}_a$ and that of b wrt E and arc $\mathcal{A}_b$. $\mathcal{A}_a$ and $\mathcal{A}_b$ are the arcs connecting a and b respectively to E, through which E is traversed.*

Argument: An event will be highly indicative of a particular context if the target ranges of all its arcs are small with respect to the total number of concepts in the knowledge base which could be the target of such an arc. For instance the concept of "to mill" only takes as agent a miller and as object flour or grain. However if the target ranges are large with respect to the total number of possible concepts, as for the verb "to do" which takes any entity as agent and any event as target, the event does not indicate any particular context. Since the specificities of a and b indicate how much E can identify a and b from any other concepts, they also indicate how specific E is to a context containing a and b. The probability that E corresponds to a relation joining a concept of type a to a concept of type b expresses the specificity of E to the context formed by a and b. This depends on both the probability of a being the target of $\mathcal{A}_a$ (the specificity of a) and that of b being the target of $\mathcal{A}_b$ (the specificity of b). From probability theory, $P(A\&B) = P(A) * P(B)$, hence the specificity of E to the context formed by a and b is:

$$specr(a, b, \mathcal{A}_a, \mathcal{A}_b, E) = speci(a, \mathcal{A}_a, E) * speci(b, \mathcal{A}_b, E)$$

where $speci(x, A_x, E)$ is the specificity of the concept x wrt the event E and the arc A_x.
This is also the probability that the context stays the same when E is traversed as described above.

Definition: *A path is a unique chain of relations joining two concepts*
Definition: *The strength of a path expresses how strongly a path joins two concepts, and is an indication of how small the context is such that the concepts fit within it*

Proposition 4.7 *The strength of a path depends on the specificity of the relations to the terms they join, and on the number of these relations.*
Argument: The specificity of the relations express the likely change of context involved when a relation is traversed. Each traversal therefore has a certain uncertainty attached to it. As the number of relations in a path increases, so does the degree of uncertainty in the result, and thus so decreases the lower bound of the path strength. The lower bound is chosen as the final value of strength, as specificity is not a particularly reliable measure itself.

The probability of the context change for the nth relation r_n of the path $\mathcal{P}$ will be determined by $specr(r_n, \mathcal{P})$. If the probability of the context being the same for the $(n-1)$th relation is $P(n-1)$, then the probability that it is the same after the nth relation is $P(n) = P(n-1) * specr(r_n)$. $P(0)$, the probability that the context is the same before any relations are traversed is 1. Expanding this out, we obtain the following equation for path strength:

$$path\_strength(\mathcal{P}) = \prod_{x \in \mathcal{P}} specr(r_n, \mathcal{P})$$

Proposition 4.8 *The strength of a set of paths between two concepts depends on the strength importance of each of the paths, and their mutual differences*
Argument: Each path expresses how semantically close the two concepts are, ie how small a context is required to contain them. The more different relations bind two concepts together, the more ways these concepts are related within the world. This indicates that they fit in the many contexts. Moreover, the more specific and the more different the contexts, the higher the likelihood of a very important causal relation between them in the world.

Unfortunately the determination of path difference is beyond the scope of this paper.

4.6 Associativity
The associativity of entities and events differs in that entities are defined by the relations which connect them to other events, whereas events are defined by their internal structure. This point is explored in more detail in [1].

Proposition 4.9 *The associativity of entity concepts depends on the strength of the set of paths which connect them.*
Argument: Associativity measures of the extent and frequence of two concepts belonging to a common context. The strength of the set of paths connecting them does the same.

The associativity of entities could be mistaken for similarity: in the context of pets, cats and dogs hold both a great degree of similarity and of associativity. However this does not hold the other way round: planes and runways are associated but never similar.

Definition: *The associativity of two events expresses how much they can affect each other, ie whether they can belong to a common context.*

Proposition 4.10 *The associativity of two prototypical events depends on whether they share pre- and post- conditions, on the associativity of their target ranges when matching arc types, on the specificity of the events to their target ranges, and on the event part hierarchy.*
Argument: If the postconditions of one event affect the other, or if two events share pre- and post- conditions then they are more likely to occur in similar circumstances, or belong to a common context. This is illustrated by the likelihood that "to shoot" and "to die" are associated. Similarly, if their agent and object ranges (each represented by a concept and its children) are associated, the events are more likely to be associated. For instance "to reap" and "to mill" are associated. Moreover a pragmatic heuristic can be added to express the fact that concepts belonging to a common context tend to occur at similar times and in similar places. If an arc has a small prototypical target range, the event is more specific to the concepts connected by this arc, and therefore the associativity of the target of this arc is given a greater importance. Finally events are also divided into other shorter events. For instance fetching an object can be divided into going to the place where it stands, taking it, and coming back while carrying it. Each of these events can be further subdivided, until a part hierarchy is obtained. This is also used: if many events correspond to part of a larger event, it is likely to be their common context. The larger event need not even be mentioned: "I put the plant's roots in a pot, added some earth and water."

Proposition 4.11 *The associativity of two factual events depends on the associativity of their prototypes, and on the associativity of their targets when matching arc types.*
Argument: The associativity of two factual events depends on how associated the types of relation they express are. This is the associativity of their prototypes: Factual events do not have pre- and post- conditions, so this ensures that events with associated pre- and post- conditions are recognised as such. Moreover, if two events belong to a common context, the targets of their events will have a high degree of associativity. The same rules apply when matching factual event arcs as when matching prototypical event arcs.

5 BASIC MODEL

The model is based on the assumption that people only talk about a few subjects at the same time. Thus, it is possible to extract the particular subject area from the terms they use. This in turn enables a preference to be assigned to the various test-hypotheses generated by the speech recogniser. The various consequences of this assumption and the manner in which they can be exploited will be discussed.

If the assumption is true, the conversation can be divided into a few specific contexts. For instance a conversation about travel will mention words to do with travel quite regularly. This context can be determined by using associativity. However there may be more than one such context. For instance, an application of AURAID at Durham University is to help deaf students attend lectures. In a course about software maintenance, one of the lecturers uses examples of badly made swings to illustrate the importance of requirement documents. If AURAID limited itself to one context, these analogies would be misinterpreted. Therefore the semantic context algorithm must allow for more than one context.

However careful analysis of a wide range of newspaper articles shows that similarity also plays a role. For instance, in a financial article three topics were mentioned: financial concepts, numbers and dates. These accounted for more than half the words, the rest being either context non-specific, or used as metaphors. The topics of numbers and dates were determined by similarity and not by associativity.

Therefore each word with semantic content produced by the speech recogniser must be classified into a particular topic area. Moreover each test hypothesis must be given a score which corresponds to how well all of its words fit the available topic areas. Thus the test hypothesis which fits the available topics best will be chosen, and its words will be added to the relevant topic. The easiest way of testing whether a word fits a particular topic is to measure it with the property (similarity or associativity) appropriate to the topic, with respect to all the words in that topic, and take average value of the obtained results: thus we obtain the semantic distance to the "centre" of the topic. The closer this value, the better.

For the classification of words into topics to work, the measures of associativity and of similarity must give results which can sensibly be compared. One simple solution is to assume that the difference can be compensated for by weighting all measurements of associativity by a constant w. As we have pointed out before, the knowledge base we are using forms part of a large natural language processing system, LOLITA. Because LOLITA performs a full syntactic and semantic analysis on all input texts, it determines as much as the information in the text allows the precise meaning and structure these texts will take in the semantic net. The semantic context algorithm can be applied to these same texts, and its predictions can be fine-tuned automatically by comparing its predictions with the correct solutions provided by LOLITA.

So far, the model assumes a static world in which topics are already determined and do not wander. This can be corrected by assigning a limited lifetime to all the concepts that have been mentioned. This lifetime must depend on the last time the topic into which the concept was classified was mentioned, and on the last time the topic itself was mentioned: If the topic had not been mentioned for some time, but is suddenly referred to, then all its concepts will be of contextual relevance. For instance after one of the lecturer's examples about swings, the context will return to software engineering. Moreover the life level of each concept can be used to modulate the importance of its contribution to the distance between new words and the topic in which it occurs. This life level allows the topic area to move slowly from one semantic area (or context) to another.

Topics also have an internal structure upon which the importance of each concept depends: certain entity concepts are closer to the centre of the topic. These central concepts are given more importance. They are the concepts for which the associativity or similarity (depending on the topic type) to each other concept of the topic is maximal. Any relations mentioned in the input are primed to increase path strength and specificity measurements used in the calculation of the semantic distance, and render the topic more attractive to new words. Moreover although relations may be grouped together into their own topics, far greater preference is given to mixed groups of entities and relations. It should be noted that the specificity of relations used in this algorithm is more complete than that described in this paper. The average strength of the paths (for associativity), or the average values of similarity, are used to determine the binding strength of the topic: this corresponds to how thinly spread a topic is.

Our assumption states that people only talk about a few subjects at the same time. Therefore the number of topics at any one time must be restricted. It also indicates that the topics should be wide enough as to allow whole subject areas to be discussed, but not so large as no longer to be able to make a clear cut decision of which test hypothesis to choose. This means that a combination of the life level, the size, and the binding strength of each topic determines its fitness. Only the n fittest topics are allowed to survive.

Finally two groups of words (and the test hypotheses from which they came) must be maintained, and assigned the lifetime corresponding to their group: words that did not fit any topic, and all the recent words which were assigned to a topic. These concepts must be constantly checked against the

possibility that they would form a better new topic. Thus we allow a limited form of backtracking to correct bad choices of topic which can occur when a new topic is introduced, and has not yet a sufficient number of concepts to be recognised as a new topic.

Therefore the model of semantic context is based on competing principles: the number of topics is restricted, all the topics compete for new words with which to increase their lifetime, and new topics are constantly checked for and if necessary the algorithm is prepared to backtrack a limited distance.

The complexity of this algorithm may seem high at first glance: Must every new concept be compared with all the concepts mentioned so far? No. The lifetime of concepts is the first limiting factor. Moreover, the importance assigned to central concepts of topics means that for each topic, only the most important concepts need be checked. However the check for new concepts does involve checking all the concept in the groups with the new concept. This is not as bad as it may seem as techniques based on inheritance and on speculative topic forming can reduce the search space effectively. It should be noted that this algorithm does not only produce the best test hypotheses, but also provides the best word senses: the semantic distance properties operate at the conceptual level.

6 CONCLUSION

Semantic information can be used to improve the recognition of spontaneous speech. Weak semantic selection and semantic context are complementary methods of reducing the number of test-hypotheses and simultaneously disambiguating the input into unambiguous concepts. The use of semantic context in other areas is being investigated. For instance, it could be used to assign preference to certain meanings of the words of an input sentence, to reduce the number of parses required before the correct interpretation is determined. It is also a means of plausible disambiguation and of determining context. The model of associativity it uses is less ad-hoc than those of models which use spreading activation constrained by "magic numbers" ([3],[4], [5]). Moreover its results can be improved by a more complete account of specificity which uses more of the information available within the semantic net.

7 REFERENCES

[1] SHORT S et al., 'What did I say... ? – Using Meaning to Assess Speech Recognisers'. *IOA 1994 Autumn Conference Proceedings* (1994)

[2] COLLINGHAM R J et al., 'Using Anti-Grammar and Semantic Categories for the Recognition of Spontaneous Speech'. *Proceedings of Eurospeech* (1993)

[3] COLLINS A et al., 'A spreading-activation theory of semantic processing'. *Psychological Review* (1975)

[4] CHARNIAK E, 'Passing Markers: A Theory of Contextual Influence in Language Comprehension'. *Cognitive Science* (1983)

[5] HIRST G, 'Semantic interpretation and the resolution of ambiguity'. *Cambridge University Press*, (1987)