Proceedings of the Institute of Acoustics

WHAT DID I SAY... ? USING MEANING TO ASSESS SPEECH RECOGNISERS

S Short* & R J Collingham

University of Durham, Department of Computer Science, NLE Laboratory

1 INTRODUCTION

The current metric used to evaluate the performance of automatic speech recognition systems is very primitive. This paper attempts to redress this by presenting the structure of a new metric which uses semantic distance.

2 A CRITICISM OF THE EXISTING METRICS

The existing metric that is used to assess the performance of automatic speech recognition systems is to simply count the number of words correctly recognised. No account is taken of the importance of the words that are incorrectly recognised, nor the understandability of the resulting output of the recogniser. This paper describes an approach to speech recognition assessment using a measure of meaning. Such a measure allows one to say for example, that a spoken text is recognised with 75% words correct and 85% of the original meaning. This may appear to be an irrelevant measure to develop because the ultimate goal of automatic speech recognition is to achieve 100% word recognition. This may be true of 'clean' speech that contains no errors, such as read speech, but is not true for natural spontaneous speech which contains many filled pauses, part words and sentence repair. For example, given the spoken input: *"I err want the err ti time of the err first tr no the last train to err Newcastle"*, we would prefer our speech recogniser to come up with something like: *"I want the time of the last train to Newcastle* which could be said to have a word accuracy of 50% but a meaning measure of 100% compared to the original spoken input. The approach that is taken is to reduce the original data and the recognised data to their individual meaning representations which are then compared.

One approach to speech recognition evaluation that is used in the ARPA ATIS competition is semantic frame filling. Typically, a spontaneous sentence is processed by a speech recogniser and the best hypothesis passed to a partial parser which extracts certain information to fill semantic frames. For example, an ATIS system given the following sentence: *"i want a flight uh that arrives in boston let's say at 3pm"* would attempt to extract the information "flight", "arrive", "Boston", "3pm", and ignore the irrelevant parts of the sentence. The filled slots in a semantic frame that are output by the recogniser may either be compared to a prepared correct frame, or passed onto a database retrieval component whose answers are then compared to a standard response.

This approach has two disadvantages. Firstly, the semantic frames have to be specified in advance in addition to the types of words that may fill the slots in a frame, making the recogniser very domain dependent. Secondly, important information may be missed due to the fact that not all of

*This research has been funded by Siemens Plessey and EPSRC

the sentences that are spoken may be reduced to slots in a semantic frame, yet they may still carry significant semantic information.

What is needed, is a domain independent method of extracting essential meaning that can be used for general English and not information rich database-like queries.

3 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 concepts. One such property 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 and evaluation of similarity.

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 – ie they have a similar effect.

Definition: *Similarity is a measure of the interchangeability of two concepts in general.*

In this context, agents can be said to have understood the meaning of a possibly corrupted text if the behaviour they display thereafter is identical to that which they would display if they had understood the uncorrupted version. The desired evaluation metric measures the similarity of the meaning of two texts, and thereby the similarity of obtained behaviour. Thus it measures how interchangeable the texts are in general, and is equivalent to the semantic distance property of similarity.

3.1 Requirements on the Knowledge Source

The quality of the results of any algorithm depends on that of its input. Thus a specification of the quality and nature of this input is important to achieve good results.

Proposition 3.1 *The knowledge base must not be specially constructed for this purpose*
Argument: The semantic distance metric must be domain independent, if it is to evaluate the output of a domain independent speech recogniser. To produce specially adapted data would be expensive for the large scale envisaged. Moreover it would require knowing precisely what information would be needed for the evaluation of semantic distance before starting off.
If instead the knowledge base already exists and is used successfully by a wide range of natural language processing and inference algorithms, it can be argued that it captures a large number of the features of natural language. Indeed if many algorithms which successfully simulate human behaviour in natural language tasks use a large common subset of information, it is likely that this subset captures some of the essential features of the language. These features would be invaluable to any evaluation of semantic distance. This corresponds to an assumption which is the basis of much work at the Natural Language Engineering Laboratory at Durham University: there is a set of core features of natural language which can be exploited by building a core system and data set. This can then be used by various specific applications. If the assumption is correct, many applications can be built easily using the common core, and reducing the amount of effort and resources required

to build each specific application. If the assumption proves wrong, this is valuable information in itself. So far, in our work on the large scale natural language processing system LOLITA, it has proved successful and fruitful: the template analysis, translation and language tutoring applications have all exploited to a large extent the core system of LOLITA.
An algorithm exploiting such a large knowledge base would have an advantage over its competitors which use specially adapted knowledge bases: flexibility. Indeed, if a large range of information is available, it is far easier to improve the performance of the algorithm by rewriting it, than by rewriting the knowledge base. Moreover, the algorithm will benefit from any improvements to the core system due to other applications' needs.

Proposition 3.2 *The information must be structured in a representation allowing fast access to the relevant part using syntactic information.*

Argument: 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. An efficient method of achieving this is to provide syntactic information to direct the search process and thus to reduce the search space.

4 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.

Before discussing the proposed model of a metric of semantic distance, the knowledge base must be briefly considered. For interested readers, more details can be found in [2]. LOLITA's knowledge base takes the form of a hypergraph, or semantic net. The vertexes correspond to concepts and are partitioned into events and entities. The edges are directed and correspond to the structural relations between vertexes. Structural relations are those which form the representation in which the knowledge base is expressed. As such they define what information can be expressed. The representation is chosen such that the relevant information can be found and extracted very rapidly from LOLITA's large knowledge base. As such the design of the structural representation is a keystone in the development of a large scale system.

An important feature of LOLITA 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 basis of all the properties we group under the term semantic distance, namely specificity.

4.1 Entities

Entities correspond to partitions of the world which display stable behavioural patterns and are useful to the agent which creates them. It is their behaviour which differentiates them from the rest of the world: even being red is a behaviour, that of reflecting red light. This is reflected in LOLITA's semantic net where it is the relations between them that define their particular nature. Thus they do not have any internal structure. Examples range from knives to capitalism and to numbers.

4.2 Events

Events correspond to relationships between entities. These relationships are expressed by a detailed internal structure within LOLITA's semantic net. For instance, "*John is a man*" is an event. The

event's **agent** is *John*, **object** is *man*, and **action** is *is_a*. The words in **bold font** are expressed in the semantic net as arcs and correspond to the internal structure of an event, whereas those in *italics* are the targets of these arcs. Events only apply to certain agents and objects: For instance the *ownership* event can only take a human agent and non-human possession. Events can be further subdivided into stative and active categories. Stative events express the state of the world (such as John's maleness). Active events are those which change the state of the world. For instance, *"John killed Mary"* would have as **precondition** that Mary were alive, and as **postcondition** that Mary were dead. This example illustrates only part of the rich internal structure of events.

Events are divided into two classes:

1. A prototypical event defines the nature of a particular type of event, or relation: it defines what agents, what objects can participate in such an event, what time-scale that event usually takes, where it can occur, what preconditions are required for it to occur, and in what postconditions it results.
2. A factual event defines properties of entities and relations between them.

5 SEMANTIC DISTANCE: THE PROPERTIES

5.1 The specificity of entities

Specificity is the keystone to most of the semantic distance properties: as we shall see, it is used to control the search required when evaluating the result of a particular property. It is based on the idea that for a concept to have a meaning, it must be identifiable by its properties. If two concepts are indistinguishable by their properties, ie by their behaviour, their meaning is the same to you.

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

Proposition 5.1 *The specificity of a property $\mathcal{P}$ with respect to an entity concept $\mathcal{C}$ depends on the reduction in the 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*

Argument: This is a more precise formulation of the definition

Specificity fundamentally corresponds to relevance: the more specific a piece of information to a particular concept, the more relevant it is likely to be: the fact that this property does not apply to other concepts means that it is a feature particular to the concept, and thereby one by which the concept can be identified.

There are two sources of information for specificity. The first is prototypical events. These express what particular events the concept may participate in. The more an event can be applied to this concept only (or any of its specialisations), the more important any instance of this event will be in indicating the closeness of the context to that concept. The second is factual events. These express useful facts such as "birds have wings". If having wings reduces the search space of possible concepts vastly, then any instance of having wings is likely to refer in some way to birds.

5.2 Similarity

The similarity of events and entities must be distinguished. In the case of events, the internal structure of the prototype event expresses what information is important to the definition of an event. This is not true for entities, for which another source of such information must be established.

5.2.1 The similarity of entities:

Proposition 5.2 *The similarity of entity concepts depends on the number of properties they share, and on the specificity of these properties*

Argument: The relations which differentiate entity concepts from the rest of the world correspond to their properties. If two concepts are differentiated from the rest of the world in a similar way this indicates, that given what is important to the agent, there is little difference between them: they are similar. One indicator of the similarity of the differentiation is the number of shared differences: a book and a man share fewer properties than a man and a woman do.

Determining the number of common properties is not sufficient: if two concepts share 3 properties but differ by 10, they are not more similar than two concepts sharing 2 properties, but differing by 1. The number of common properties must therefore be compared to the total number of properties to express similarity. We are interested in the size of the intersection of the two sets of properties, versus that of their union.

However this is insufficient: lions and dogs share a large number of properties, such as being capable of living, rearing their young on milk, eating meat... If only the number of properties were to be considered, there would be little difference in the similarity of lions and dogs, and that of cats and dogs. In a large knowledge base, adding a property to a particular general concept could change substantially the measure of similarity, which affects its robustness. Moreover, people would not consider lions and dogs to be as similar as cats and dogs. Therefore there is a problem of relevance of the shared properties. The properties must not be so general that they express very little information. Thus the other indicator of similarity is the specificity of the shared properties to the concepts being compared. The common size of cats and dogs, their existence as pets are important specific properties which increase their degree of similarity.

Even more precise values of similarity can be obtained in cases where the properties which are not shared are considered: the similarity of the properties themselves can be evaluated and used to form weaker judgements of similarity in the same way as is done for common properties.

The ratio of the cardinality of the intersection of the sets $\mathcal{A}$ and $\mathcal{B}$ to that of their union is given by:

$$\frac{card(\mathcal{S})}{card(\mathcal{A}) + card(\mathcal{B}) - card(\mathcal{S})} \tag{1}$$

where $\mathcal{S} = \mathcal{A} \cap \mathcal{B}$. However this must be extended to take into account the fact that certain properties are more important to the measurement of similarity than others, importance expressed by specificity. Thus the similarity between a and b is

$$simil(a,b) = \frac{\sum_{x \in \mathcal{S}} min(speci(x,a), speci(x,b))}{\sum_{x \in \mathcal{A}} speci(x,a) + \sum_{x \in \mathcal{B}} speci(x,b) - \sum_{x \in \mathcal{S}} min(speci(x,a), speci(x,b))} \tag{2}$$

where $\mathcal{A}$ is the set of the entity a's properties. $\mathcal{B}$ is that of b; $\mathcal{S} = \mathcal{A} \cap \mathcal{B}$. $speci(x,y)$ is the specificity of property x to concept y: it has values ranging from 0 (x cannot be applied to y) to 1 (x can only be applied to y). $min(x,y) =$ if $x > y$ then y else x. The min function ensures that differences in specificity between common concepts are penalised. This formula can be further extended if we no longer consider a boolean measure of property sharing. In this scheme, the amount a property is shared is expressed by its communality. Thus the $commun(x,\mathcal{Y})$ is the normalised value (between 0 and 1) corresponding to the highest similarity between property x and any property in the set $\mathcal{Y}$. It can be seen that if communality is reduced to being equal either to 0 or to 1, that this equation is identical to the previous one.

$$simil(a,b) = \frac{intersection\_value(a,\mathcal{A},b,\mathcal{B})}{\sum_{x \in \mathcal{A}} speci(x,a) + \sum_{x \in \mathcal{B}} speci(x,b) - intersection\_value(a,\mathcal{A},b,\mathcal{B})} \tag{3}$$

where

$$intersection\_value(a, \mathcal{A}, b, \mathcal{B}) = \sum_{x \in \mathcal{A} \cup \mathcal{B}} min(speci(x, a) * commun(x, \mathcal{B}), speci(x, b) * commun(x, \mathcal{A}))$$

5.2.2 The similarity of events: Prototypical similarity which compares the types of the relations, and factual similarity which compares the particulars of two instances must be considered separately.

Proposition 5.3 *The similarity of two prototypical events is obtained by determining for each type of arc, the similarity between the targets of each event.*

Argument: Prototypical events define the nature of an event. All the information they express is important, in that it differentiates the relations from each other. The specification of the possible agents and objects of an event must be obeyed: a factual event simply does not make sense if the agent it has is not either that of the prototypical event's agent, or its specialisation. Therefore for the prototypical events to be likely to be interchangeable, not only must their agents and objects be similar, but also they must have a non-empty intersection.

Equally restrictive are the pre- and post- conditions for an active event, as without satisfaction of both, the event could not occur. Thus two similar active events must share similar pre- and post-conditions. The conditions are divided into two sets (pre- and post- conditions), as it does not make sense to compare preconditions with postconditions. Each set of conditions is compared using **equation 2**: in this case specificity is determined by the inverse of the number of events which have the same condition. The reason for this is illustrated by the example "to die": its precondition is not specific in that it requires a living agent, but its postcondition is, as it requires a dead agent, which is very rare. Conditions are represented by factual events, such as "is_a alive_being".

A more advanced form of comparison of conditions actually compares pairs of conditions using measurement of factual event similarity: for two events a and b, with respective sets of conditions $\mathcal{A}$ and $\mathcal{B}$ respectively, **equation 3** is used: the specificity is defined as above, and $commun(x, \mathcal{Y})$ is defined as the maximum similarity obtained when comparing condition x with any of the conditions in $\mathcal{Y}$ – note that its value must be normalised to the range [0..1].

Static events have a similarity corresponding to pre- and post- condition similarity in active events. It is determined by comparing the sets of pre- and post- conditions each of the static events satisfy as done in **equation 1**: similarity is defined as the interchangeability of two concepts. No detailed analysis of the types of the common elements of the sets is performed: no more importance is assigned to one active event than to any other. This corresponds to the fact that the stative event is not dependent on the existence of any particular active event.

Other information such as prototypical duration can also be used, although it is far less strict. In LOLITA, this information is represented as factual, so must be compared using the algorithm to measure similarity between factual events.

The importance of a particular arc type depends on how strictly it must be obeyed for the event to occur. Thus the agent, object, pre- and post- conditions will be assigned equal importance in the calculation of similarity.

Proposition 5.4 *The similarity of two factual events depends on the similarity of their corresponding prototype events, and on the similarity between the targets of each event when matching arc types.*

Argument: The similarity of two factual events depends on how similar the types of relation they express are, ie the similarity of the prototype events: this ensures that "John eats an apple" and "John buys an apple" differ. Moreover the similarity of each of the elements of internal structure of the events determines similarity: thus "John eats an apple" differs from "Mary eats a pear". Factual events do not have pre- or post- conditions.

5.3 The similarity of two texts.

Two texts are represented within the semantic net as two sets of events. The number of events may differ, even if the two texts express similar information. An example of this occurring is when one of the texts is the summary of the other.

Proposition 5.5 *The similarity of the central concepts in texts is more important to the measurement of their similarity, than is those of the other concepts.*

Argument: When comparing a summary with an unabridged text, one checks whether the most important information in the text is written in the summary. Therefore the measure of similarity between two texts must ascribe more importance to the similarity of the important concepts within the texts.

Proposition 5.6 *The importance of a concept within a text is determined by the number of times the text refers to it.*

Argument: The central ideas of a text are usually the basis of some argument. Therefore the text tends to refer to them more often than to its other concepts. For instance, if a newspaper article is comparing the value for money of train versus plane travel between cities, the concepts of train, planes and cities will be referred to often.

The set of concepts used in a text are referred to by the internal structure of the events of the text. The importance of a particular concept of a text is then simply the number of events of the text which refer to it. This type of importance is called the interconnectivity of a concept.

The similarity of two sets of events is calculated using an equation analogous to **equation 3**. In this case, the importance of the particular event is determined by the interconnectivity (*interc*) of all the elements of its internal structure. The amount the events are shared, or communality, is determined as follows: $commun(x, \mathcal{Y})$ is defined as the maximum similarity obtained when comparing the event x with any of the events of $\mathcal{Y}$. Again, $commun(x, \mathcal{Y}) \in [0, 1]$

$$text\_simil(\mathcal{A}, \mathcal{B}) = \frac{intersection\_value(\mathcal{A}, \mathcal{B})}{\sum_{x \in \mathcal{A}} interc(x, \mathcal{A}) + \sum_{x \in \mathcal{B}} interc(x, \mathcal{B}) - intersection\_value(\mathcal{A}, \mathcal{B})} \quad (4)$$

where

$$intersection\_value(\mathcal{A}, \mathcal{B}) = \sum_{x \in \mathcal{A} \cup \mathcal{B}} min(interc(x, \mathcal{A}) * commun(x, \mathcal{B}), interc(x, \mathcal{B}) * commun(x, \mathcal{A}))$$

It should be noted that the events of the texts should not be used in the evaluation of the texts' similarity, to avoid them from contaminating the results.

6 REFINEMENTS TO SIMILARITY

The metric described in this paper is crude in two respects.

First, the proposed measurement of interconnectivity only takes into account concepts rather than contexts. This means that an important idea is only recognised if it is always referred to by the same concept. Obviously this occurs rarely: repetitions are usually considered a mark of bad style. Thus if a text contains references to trains, railway stations, locomotives, but none of these words are repeated sufficiently often, the importance of the idea of trains will be totally missed. The metric used at Durham University copes with this problem by extending the measurement of interconnectivity to determine the important contexts by using another semantic distance property: associativity. Interested readers may find more information about this topic in the paper [1].

The second problem is that the measurement of similarity may be biassed if LOLITA has more knowledge about some domain than about another. This will occur because the measurement of similarity depends on specificity, and specificity depends on the number and type of properties which a concept has or can have. We will assume the amount of knowledge about prototypical events to be homogeneous over the whole net. This is justified in that if the meaning of a relation is defined, it must be defined completely to be useful for reasoning. Thus its internal structure will be complete. This cannot however be said of the factual events: the number of these will vary depending on what type of task LOLITA had previously been used for. For instance, if LOLITA is exclusively used for template analysis of financial texts, its knowledge base will contain a lot of factual information about finance, but not so much on the migratory habits of butterflies. This distortion can be corrected for by determining the ratio of the amount of factual information versus that of the prototypical information. The metric used at Durham University takes full account of this by using relevant properties, which are unfortunately beyond the scope of this paper.

Moreover the scenario we describe is simplified. For instance events can have many elements for each arc type, for instance many agents.

7 COMPARISON WITH OTHER METRICS OF SEMANTIC DISTANCE

Semantic distance arises a lot of interest (for instance [4] and [5]), although few papers present any actual means of realising it. However those which do, such as [3], rely on the existence of a specialisation hierarchy and assume that the conceptual change in granularity expressed by each of its levels is constant. Thus they simply calculate the semantic distance by finding the least subtype which subsumes the two types to compare, and adding the distance from each of these to the subsuming type. However this assumption has not been justified. A counter-example is provided by numbers: These can be divided into natural and rational numbers. It can be shown that although both sets are infinite, the cardinality of the set of the natural numbers is smaller than that of the rational numbers. Thus, although there is a least subtype, "numbers", the change in conceptual size from it to its children varies. The measure of similarity defined in this paper does not rely on such an assumption but uses the information defining the concepts. This results in a behaviour which varies with the amount of information available. At worst, if two concepts cannot be distinguished by their properties, they will be assumed identical. This is analogous to what a blind person might know of the difference between green and red if he did not know in what circumstances they usually appear.

8 REFERENCES

[1] SHORT S et al., 'Making Use of Semantics in an Automatic Speech Recognition System', *IOA 1994 Autumn Conference Proceedings* (1994)

[2] LONG D et al., 'Reasoning by Analogy and Causality: a Model and Application', *Ellis Horwood*, (December 1993)

[3] FOO N et al., 'Semantic Distance in Conceptual Graphs', *Conceptual Structures: Current Research and Practice: Ellis Horwood* (1992)

[4] DELUGACH H, 'An Exploration Into Semantic Distance', *Seventh Annual Workshop On Conceptual Graphs Proceedings* (1992)

[5] MYAENG S H, 'Conceptual Graphs as a Framework for Text Retrieval', *Conceptual Structures: Current Research and Practice: Ellis Horwood* (1992)