WORKING NOTE AC104

Alan Creak

30 September 1996

SENTENCES THAT MEAN "X", MORE OR LESS

In thinking about how to select sentences with a desired meaning from a collection of sentences stored in a database, I was helped by a geometrical analogy. This is it.

While composing another note[1] about selecting sentences from a fixed pool, I wanted to get some idea of how a constraint on the meaning of a sentence would affect the proportion of sentences which could be accepted. In thinking around this idea, I used a somewhat fanciful analogy which was slightly illuminating. I eventually decided that it was rather too fanciful to include in the original note, which was at least intended to be fairly serious, but I rather like it, so here it is, preserved for posterity.

CAUTION.

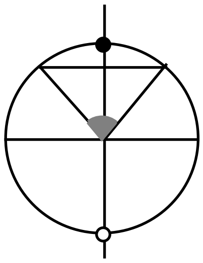
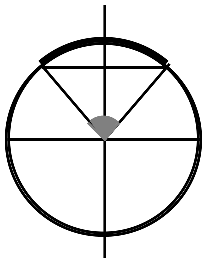
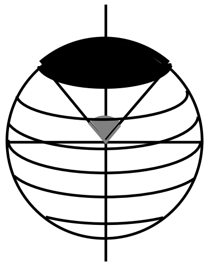
I hope it is unnecessary for me to state (but, just in case, here it is) that I do not attach any direct significance whatever to any of these diagrams, number, hypotheses, or whatever. My only excuse for introducing them is that I have nothing better to offer, and that they lead me to some questions and conjectures which I think are more worthy of attention.

A GEOMETRICAL MODEL OF MEANING.

Suppose that any sentence can be represented by a unit vector in some semantic space of very high dimension, in which the coordinates are principal axes of meaning. The set of all sentences could be represented in this space, and I know of no reason why it shouldn't be homogeneous.

The set of interesting sentences is less likely to be homogeneous, and will probably be clumpy. The clumpiness is inevitable unless you can define effective principal axes of meaning which are semantically orthogonal (I don't even know what that means in any but the vaguest sense), because some pairs of topics which one might expect to find represented by axes are more closely related than others; there are likely to be more sentences about both chickens and eggs than about both chickens and interstellar space. The set of all sentences which contain a particular word will similarly be clumpy, but I'll assume that there's no particular system about the clumps so that a homogeneous approximation is no more unreasonable than the rest of the model. In any case, I don't know what else I can do.

I then assume that the aim of the selection is to find a sentence with a semantic vector fairly close to the vector of sentence you really want. With this definition, I can reduce the problem to geometry, some of which is within my range of ability. Here is a sample of results for semantic spaces of ridiculously low dimension. The figures in the table are the proportion of the whole set of sentences with semantic vectors which fall within the listed acceptance angle of the desired vector. (My geometry gives out above three dimensions.)

Dimension	1	2	3
Diagram (The shaded angle is twice the acceptance angle; the dark region represents the selection.)			
Acceptance angle	Proportion selected		
π / 3 (60º)	0.500	0.333	0.250
π / 4 (45º)	0.500	0.250	0.147
π / 6 (30º)	0.500	0.167	0.067

COMMENTS.

The numbers presented are highly unlikely to be very significant, but perhaps their relationships are, and - particularly - the way these relationships change with increasing dimension.

The one-dimensional case is pathological; it applies to a universe in which the only statements are "chickens" and "not chickens". Nevertheless, it acts as a base case for the effect of dimensions, and helps to show that the effect of a change in dimension is not always what you expect. It's also a convenient hook on which to hang a question about the "semantic space" : what about sentences like "maybe chickens", "sometimes chickens", "perhaps not chickens" ? You could either remove the restriction to unit vectors so that each of these uncertain statements could be represented by a vector shorter than one unit in length, or - as I've supposed - regarded doubt as a different dimension. Given the basic improbability of the model, it doesn't seem worth following up, but any attempt at a more refined guess might require some treatment of these questions.

Two trends are noticeable, and I think that these are more likely to point to useful conclusions.

First, consider the effect of increasing the number of dimensions : keeping the acceptance angle constant - which is, in some sense, requiring that at least a certain component of the sentence conforms to the desired pattern - the proportion of acceptable sentences decreases fairly quickly as the dimension increases. Transferred to the sentence domain, that might suggest that finding an acceptable sentence will rapidly become harder as a more general coverage of topics is attempted. This corresponds to my guess[1] that constraining the topic is likely to be a fruitful approach.

Second, consider the effect of angle of acceptance when the dimension is kept constant. Once again, the proportion of acceptable sentences decreases rapidly as the angle is decreased. Interpreting, the suggestion is that if you want to make a statement rather precisely, it will be hard to find a sentence which will be acceptable (and it will get worse as you try to provide for more general conversation); conversely, if you don't much care what you say about the topic, finding a sentence might be quite easy.

EVALUATION.

Both these conclusions conform to reasonable expectation, so to that extent the model is useful. I thought that it helped me to make sense of the expectations when I was thinking about the problem of selection. Having got this far, though, I don't really see what else it can do; but if it fits any other sorts of question in the context of selecting sentences from collections thereof, it might be worth trying.

REFERENCES.

1 : G.A. Creak : *Analysing WordKeys*, unpublished Working Note AC103 (September, 1996).