SENTENCE COMPREHENSION WITHOUT PROPOSITIONAL STRUCTURE

STEFAN L. FRANK

Nijmegen Institute for Cognition and Information
Radboud University Nijmegen, Montessorilaan 3
6525 HR Nijmegen, The Netherlands
E-mail: S.Frank@nici.ru.nl

Comprehending a sentence requires the construction of a mental representation of the situation the sentence describes. Many researchers assume that, apart from such a situational representation, there is a level of representation at which the propositional structure of the sentence is encoded. This paper presents a simple sentence comprehension model, consisting of a neural network that transforms sentences into representations of the events they describe. During training, the network develops internal representations of the sentences. An investigation of these representations reveals that they can encode propositional information without implementing propositional structure.

1. Introduction

The mental representation of discourse has generally been assumed to involve three distinct levels (Kintsch & Van Dijk, 1978; Van Dijk & Kintsch, 1983): the *surface text*, the *textbase*, and the *situation model*. The first of these consists of the text's literal wording. This gives rise to the second level, in which the text's propositional structure is encoded. For instance, the sentence *Bob plays soccer* may be represented in the textbase as a predicate relating two arguments: PLAY(BOB,SOCCER).[a] The highest level of representation, the situation model, includes inferences derived from the reader's background knowledge and experience. In the current example, this may be information about 'what it is like' when Bob plays soccer, for instance, that he is likely to be outside.

Although there is considerable empirical evidence for the existence of a situational representation in the cognitive system (e.g., Zwaan, Madden,

[a] Examples of surface text shall be printed in *italics*, while SMALLCAPS denote propositions. Situation model descriptions are printed in normal font.

Yaxley, & Aveyard, 2004), evidence supporting the existence of a textbase-level representation is weak. Data by Ratcliff and McKoon (1978) and Dell, McKoon, and Ratcliff (1983) has been taken as evidence for a propositional decomposition of text, but as argued by Zwaan (1999) and Frank, Koppen, Noordman, and Vonk (in press), it can also be interpreted differently. Since similar propositions usually describe similar situations, it is difficult to distinguish between propositional and situational representations.

An experiment by Fletcher and Chrysler (1990) is often assumed to indicate that text is indeed represented at three distinct levels. They showed that readers more often confuse two sentences that differ only in surface text than two sentences that also differ propositionally, which in turn are more difficult to tell apart than two sentences that also differ situationally. However, as will be explained in Section 3.3, Frank et al. (in press) have shown that these results do not require distinct levels of representation. They trained a neural network to convert sentences into situational representations of the described events. The network's *single* intermediate level of representation could then account for Fletcher and Chrysler's findings.

The situational representations used by Frank et al. (in press) were taken from the Distributed Situation Space (DSS) model (Frank, Koppen, Noordman, & Vonk, 2003). This paper shall first explain these situational representations. Next, Frank et al.'s (in press) simple sentence comprehension model is explained, and the nature of its intermediate representation of sentences is investigated.

2. Situational representations

The DSS model (Frank et al., 2003) simulates how readers apply their world knowledge to make inferences during story comprehension. The model represents stories at the situational level only: Its representations do not depend on the wording or structure of the text but only on the relation between story events and world knowledge. Since the amount of world knowledge readers have is simply too large to implement even a significant subset for use by a computational model, a simple *microworld* was developed, all knowledge of which is available to the model.

2.1. *The microworld*

In the microworld, there are two characters whose names are Bob and Jilly. All activities they can engage in, as well as all situations they may find themselves in, can be described using just 14 basic events. These include states like ‘Bob is tired’ and ‘Jilly is outside’, and actions such as ‘they play soccer’, ‘they play hide-and-seek’, ‘Bob plays with the dog’, ‘Jilly plays a computer game’, ‘Bob wins’, and ‘Jilly wins’. The basic events can be combined using the Boolean operators of negation, conjunction, and disjunction. However, not all combinations of basic events are as likely to occur. For instance, Bob and Jilly can only play one game at a time, and they never play a computer game outside or soccer inside. Also, they are more likely to be both outside or both not outside (i.e., inside) than to be at different locations.

2.2. *Representing microworld knowledge and situations*

Knowledge about the microworld was not implemented directly. Instead, a sequence of 250 consecutive example situations was constructed, following the microworld constraints. In each of these examples, each of the 14 basic events is either stated to be the case or not the case. That is, each example situation can be represented by a 14-dimensional binary vector, which served as training input to a Self-Organizing Map (SOM; Kohonen, 1995).

During training, the map develops a representation for each basic event, four of which are shown in Fig. 1. For each basic event p, each SOM cell i has a unique value $\mu_i(p)$ between 0 and 1, which is indicated by the cell's grayness in Fig. 1. This value, called the membership value of i for event p, denotes the extent to which the cell is part of the representation of p.

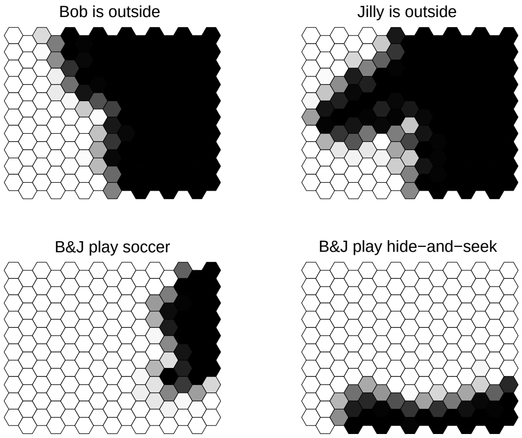

Figure 1. Self-organized maps of four basic events.

A useful property of these representations is that any microworld situation can be represented by combining the basic events using negation ($\neg$), conjunction ($\wedge$), and disjunction ($\vee$):

$$\begin{aligned}\mu_i(\neg p) &= 1 - \mu_i(p)\\ \mu_i(p \wedge q) &= \mu_i(p)\mu_i(q)\\ \mu_i(p \vee q) &= \mu_i(p) + \mu_i(q) - \mu_i(p)\mu_i(q).\end{aligned} \tag{1}$$

Since the SOM consists of 150 cells, $\mu(p) = (\mu_1(p), \ldots, \mu_{150}(p))$ forms a 150-dimensional vector representation of p. That is, each microworld situation corresponds to a point in 150-dimensional 'situation space'. Since there is no one-to-one relation between dimensions of this space and events in the microworld, this representation is distributed. It is the distributed nature of situation space that gives the DSS model its name.

2.3. *Belief values*

The SOM representations are not arbitrary but encode the probabilistic relations among the events. This is easy to see from the four patterns in Fig. 1. Although they are meaningless by themselves, the relations among them reflect relations among the represented events. The patterns representing 'Bob is outside' and 'Jilly is outside' overlap for a large part, because Bob and Jilly are more likely to be at the same place than to be at different places. Likewise, the representations of 'Bob and Jilly play soccer' and 'Bob and Jilly play hide-and-seek' do not overlap at all, because Bob and Jilly cannot play soccer and hide-and-seek simultaneously. Also, the activation pattern for 'they play soccer' falls completely within those for being outside, since soccer can only be played outside.

Frank et al. (2003) have shown that the conditional probability that event p occurs in the microworld, given that q does, can be estimated quite accurately from their representations. This estimation, denoted $\tau(p|q)$, is called the *belief value* of p in event q, and is a measure for the 'belief' the model has in the occurrence of p. It is computed by

$$\tau(p|q) = \frac{\sum_i \mu_i(p)\mu_i(q)}{\sum_i \mu_i(q)}.$$

Also, the a priori probability (belief value) of event p follows from its representation: $\tau(p) = \frac{1}{150}\sum_i \mu_i(p)$.

Note that Equation 1 preserves the relation between belief values and probabilities in the microworld. Also note that the DSS model's representation of world *knowledge* cannot be separated from the representation of world *situations*. Instead, the vector representations of situations implement world knowledge. This unity of situation and knowledge shows that the DSS model indeed represents story events at the situational level.

3. A simple sentence comprehension model

Frank et al. (in press) argue that the goal of sentence comprehension is the construction of a situation model and that an intermediate, possibly propositional representation may only arise to the extent that it is useful for reaching this goal. They trained a simple recurrent neural network (Elman, 1990) to convert sentences from a *microlanguage* into the DSS model's vector representation of the corresponding events in the microworld. The intermediate representation that developed in the network's hidden layer forms an alternative type of textbase-level representation.

3.1. *The microlanguage*

The microlanguage consists of 15 words: *Bob, Jilly, and, plays, is, wins, loses, soccer, hide-and-seek, a_computer_game, with_the_dog, outside, inside, tired, awake.* Note that both *a_computer_game* and *with_the_dog* are considered one word. By combining the 15 words following the grammar of Table 1, 328 different sentences can be constructed.

Table 1. Grammar of the microlanguage.

S	→	NP VP
NP	→	*Bob* | *Jilly* | *Bob and Jilly* | *Jilly and Bob*
VP	→	*plays* Game [Place | *and is* State | *and* Result]
	→	*is* Place [*and plays* Game | *and* State | *and* Result]
	→	*is* State [*and plays* Game | *and* Place | *and* Result]
	→	Result [*and plays* Game | Place | *and is* State]
Game	→	*soccer* | *hide-and-seek* | *a_computer_game* | *with_the_dog*
Place	→	*outside* | *inside*
State	→	*tired* | *awake*
Result	→	*wins* | *loses*

3.2. *The model*

A simple recurrent network was trained to transform 290 of the 328 microlanguage sentences into the DSS representation of the described events. The network consisted of 15 localist input elements, one for each word, and processed sentences one word at a time. The hidden layer had six elements and received activation from the input layer as well as a copy of its own previous activation. The output layer consisted of 150 elements, one for each DSS vector element. Seven such networks were trained, each with a different initial random weight setting.

During training, the difference between the desired and the actual output was backpropagated to adjust the connection weights. After training, however, the adequacy of the network's output was measured by using belief values. The idea is that comprehension of a sentence describing event p should result in an output vector $X(p)$ in which p's belief value $\tau(p|X(p))$ is larger than its a priori belief value $\tau(p)$. In the ideal case, when $X(p) = \mu(p)$ so $\tau(p|X(p)) = \tau(p|p)$, the 'amount of comprehension' is defined to equal 1. If $\tau(p|X(p)) = \tau(p)$, the network has not learned anything about the sentence and the amount of comprehension is defined as 0. It is also possible for the sentence to be misunderstood, in which case $\tau(p|X(p)) < \tau(p)$ and the amount of comprehension is negative. Formally, the amount of comprehension of sentence p equals

$$\text{compr}(p) = \frac{\tau(p|X(p)) - \tau(p)}{\tau(p|p) - \tau(p)}.$$

After training, the networks' average comprehension scores lie significantly above 0, both for training and test sentences. This shows that the network learned its task and could extrapolate to novel sentences.

3.3. *Probing the hidden layer*

After processing a sentence, the activations of the hidden-layer elements form an intermediate vector representation of the sentence, called 'the sentence's vector'. Frank et al. (in press) investigated the nature of this representation. It was found that, on average, the euclidean distance between vectors representing sentences that differ at the surface text level but not propositionally is shorter than the distance between vectors representing sentences that do differ propositionally. Also, vectors of sentences that differ propositionally but describe identical situations were, on average, closer together than vectors of sentences describing different situations. Assuming vectors that are closer together are harder to tell apart, these results correspond to Fletcher and Chrysler's (1990), discussed in the introduction.

Here, we shall probe the hidden layer in more detail. In particular, we are interested in finding out if particular sentence vectors encode mostly textual, propositional, or situational information, and how this relates to comprehension scores. This is done by computing distances between vectors of triplets of sentences, such as those labelled A, B, and C_1 in Table 2.

Table 2. Four microlanguage sentences, constituting two test triplets, and their propositional structures.

	sentence	proposition
A	*Bob is tired and inside*	INSIDE(BOB) ∧ TIRED(BOB)
B	*Bob is tired and plays soccer*	PLAY(BOB,SOCCER) ∧ TIRED(BOB)
C_1	*Bob is outside and tired*	OUTSIDE(BOB) ∧ TIRED(BOB)
C_2	*Bob plays a_computer_game and is tired*	PLAY(BOB,COMP) ∧ TIRED(BOB)

Which two of the three sentences A, B, and C_1 are most similar to each other depends on the representational level one looks at. When concerned only with the surface text, A and B are most similar since they have a four-word sequence in common (*Bob is tired and*). At the propositional level, A and C_1 are most similar because they differ by only one predicate (INSIDE vs OUTSIDE). Situationally, however, B and C_1 are most similar since soccer must be played outside which excludes being inside.

Replacing sentence C_1 by C_2 results in another triplet. Again, A and B are most similar textually. For this reason, (A,B) is called the 'textual pair'. Propositionally, B and C_2 have most in common, while A and C_2 describe similar situations. Therefore, pairs (A,C_1) and (B,C_2) are the 'propositional' pairs, while (A,C_2) and (B,C_1) are 'situational'. By computing distances between sentence vectors of these pairs, the triplet's level of representation can be revealed. For instance, a relatively short distance between the textual pair indicates that the triplet's representation strongly depends on surface text. Below, we shall probe the networks' intermediate representations to find out what type of information is encoded and how this relates to comprehension scores.

3.4. *Results*

Two of the test triplets used to probe the intermediate representations are shown in Table 2. Thirty additional triplets were formed by replacing *Bob* by *Jilly*, by swapping the clauses on either side of *and*, and by replacing the clause about being tired by being awake, winning, or losing. The 32 test triplets were processed by the seven trained networks, and distances between all three sentence vectors of each triplet were computed. These were normalized, such that the distance between the textual pair (the so-called 'textual distance'), the propositional distance, and the situational distance sum up to 1 for each triplet. Table 3 shows these distances, averaged for each network, as well as the networks' average comprehension scores for the 64 different sentences from the test triplets.

Table 3. Textual, propositional, and situational distances, averaged per network and sorted by the networks' comprehension scores on the triplets' sentences.

	distance		
compr	text	prop	sit
.327	.357	.379	.264
.310	.378	.350	.272
.293	.329	.354	.317
.273	.339	.350	.311
.272	.367	.373	.260
.270	.231	.403	.366
.238	.257	.384	.358
av.	.323	.370	.307

Sentence vectors mainly encode situational and textual information: The average propositional distance is significantly larger than the textual ($t_{446} = 5.50; p < .0001$) and situational distances ($t_{446} = 8.38; p < .0001$). The difference between textual and situational distances is close to significant ($t_{446} = 1.74; p = .08$). Although propositional distances are larger on average, it was found to be the shortest in 11.6% of the cases. In these cases, the main level of representation was propositional.

The results in Table 3 indicate that situational representations lead to higher comprehension scores than textual or propositional representations. Indeed, there is a large negative correlation of $-.69$ between the networks' comprehension scores and the average situational distance. However, due to the limited number of data points, it is not significant ($t_5 = 2.16; p = .08$). Likewise, the correlation between comprehension scores and textual distances is positive but not significant ($r = .66; t_5 = 1.97; p = .11$). The small correlation between comprehension scores and propositional distances is far from significant.

Defining the comprehension of a triplet as the average of the comprehension scores of its three sentences, the number of data points can be increased by looking at the correlation between comprehension and distances per triplet. Although these correlations are lower than those presented above, they are significant for textual ($r = .19; t_{222} = 2.85; p < .005$) and situational ($r = -.26; t_{222} = 3.99; p < .0001$) distances.

3.5. *Discussion*

The more situationally represented sentence triplets are, the better they are comprehended by the network. This is not surprising since comprehension scores measure the quality of the *situational* output representation. If the intermediate representation already encodes the described situation, the output can be expected to be closer to the desired output compared to sentences whose vector is less situational.

During training, the network learns to produce situational output representations of the sentences. This task will make the network tend towards also developing situational *intermediate* representations, which help produce the desired output. As a result, vectors encode mainly situational information. Also, it does not come as a surprise that propositional information in vectors is rare. During training, the network is provided with both textual input and the desired situational output, while a propositional representation is something it has to develop by itself. It is likely that a more complex microlanguage and microworld, consisting of several entities that can be combined in different ways, will result in more propositional representations, since it may be the need for comprehending such complexities that drives the development of propositional intermediate representations.

4. Conclusion

The mental representation of text is usually assumed to include both a propositional textbase level and a situational level. The DSS model shows how world knowledge and a situational representation of text can be implemented. Frank et al.'s (in press) simple sentence comprehension model is a first attempt to extend the DSS model with a textbase-like level. However, the intermediate representations developed by the model are very different from the propositional structures that are traditionally thought to constitute the textbase.

First of all, there are no predicate-argument structures in the intermediate representation. Second, although sentence vectors encode mainly propositional information in a few cases, they are never purely propositional. Textual and situational information also affects the intermediate representation. Third, the construction of a textbase is not assumed to be a goal of text comprehension. Instead, there is only one comprehension goal: the construction of a situational representation. Even if some propositional information is encoded in the network's hidden layer, it is only there to the extent that it is useful for accomplishing this goal.

References

Dell, G.S., McKoon, G., & Ratcliff, R. (1983). The activation of antecedent information during the processing of anaphoric reference in reading. *Journal of Verbal Learning and Verbal Behavior*, 22, 121-132.

Elman, J.L. (1990). Finding structure in time. *Cognitive Science*, 14, 179-211.

Fletcher, C.R., & Chrysler, S.T. (1990). Surface forms, textbases, and situation models: recognition memory for three types of textual information. *Discourse Processes*, 13, 175-190.

Frank, S.L., Koppen, M., Noordman, L.G.M., & Vonk, W. (2003). Modeling knowledge-based inferences in story comprehension. *Cognitive Science*, 27, 875-910.

Frank, S.L., Koppen, M., Noordman, L.G.M., & Vonk, W. (in press). Modeling multiple levels of text representation. In F. Schmalhofer & C.A. Perfetti (Eds.), *Higher level language processes in the brain: inference and comprehension processes.* Mahwah, NJ: Erlbaum.

Kintsch, W., & Van Dijk, T.A. (1978). Toward a model of text comprehension and production. *Psychological Review*, 85, 363-394.

Kohonen, T. (1995). *Self-Organizing Maps.* Berlin: Springer.

Ratcliff, R., & McKoon, G. (1978). Priming in item recognition: evidence for the propositional structure of sentences. *Journal of Verbal Learning and Verbal Behavior*, 17, 403-417.

Van Dijk, T.A., & Kintsch, W. (1983). *Strategies of discourse comprehension.* New York: Academic Press.

Zwaan, R.A. (1999). Embodied cognition, perceptual symbols, and situation models. *Discourse Processes*, 28, 81-88.

Zwaan, R.A., Madden, C.J., Yaxley, R.H., & Aveyard, M.E. (2004). Moving words: dynamic representations in language comprehension. *Cognitive Science*, 28, 611-619.

Acknowledgements

The research presented here was supported by grant 451-04-043 of the Netherlands Organization for Scientific Research (NWO). I would like to thank Mathieu Koppen and Leo Noordman for their comments on earlier versions of this paper.