Incremental Construction of an Associative Network from a Corpus

Benoît Lemaire (Benoit.Lemaire@upmf-grenoble.fr)
L.S.E., University of Grenoble 2, BP 47
38040 Grenoble Cedex 9, France

Guy Denhière (denhiere@up.univ-mrs.fr)
L.P.C & C.N.R.S. Université de Provence
Case 66, 3 place Victor Hugo
13331 Marseille Cedex, France

Abstract

This paper presents a computational model of the incremental construction of an associative network from a corpus. It is aimed at modeling the development of the human semantic memory. It is not based on a vector representation, which does not well reproduce the asymmetrical property of word similarity, but rather on a network representation. Compared to Latent Semantic Analysis, it is incremental which is cognitively more plausible. It is also an attempt to take into account higher-order co-occurrences in the construction of word similarities. This model was compared to children association norms. A good correlation as well as a similar gradient of similarity were found.

Introduction

A computational model of the human semantic memory may be valuable for its ability to mimic the human semantic representations, but also for its ability to mimic the *construction* of these representations over a long period of time. Not all models possess both features. For instance, symbolic formalisms like semantic networks had proven to be interesting for representing human knowledge but they do not tell us how human beings build such representations over their life. Several computational models of both the representation and construction of the human semantic memory have been proposed in the recent years. Some of them are based on a general common mechanism that rely on a huge input, composed of examples of associations between words. The statistical analysis of the occurrences of each word within well-defined units of context leads to a computational representation of association links between words. The representation of word meanings per se is not of significant interest, it is rather their association links which combined will form a model of the long-term semantic memory.

These models can be distinguished along six features:

1. the kind of input they are based on (either a corpus or word association norms);
2. the knowledge representation formalism (either vector-based or network-based);
3. the way a new context is added to the long-term semantic memory (incrementally or not);
4. the unit of context in which co-occurrence information is considered (either a paragraph or a sliding window);
5. the use or not of higher-order co-occurrences;
6. compositionality: the way the meaning of a text can be inferred from the meaning of its words.

After a description of existing models, we will discuss these features, present our model and describe an experiment that aims at comparing our model to human data.

Existing computational models of the construction of semantic representations

Latent Semantic Analysis

LSA (Landauer, 2002) takes as input a corpus of free texts. The unit of context is the paragraph. The analysis of the occurrences of each word within all paragraphs leads to a representation of the meaning of words as vectors, which is well suited for drawing semantic comparisons between words. The underlying mechanism (singular value decomposition of the word-paragraph occurrence matrix) implicitly takes into account higher-order co-occurrences (Kontostathis & Pottenger, 2002). Compositionality in this model is straightforward: the meaning of a text is a linear combination of the meaning of its words. There is however no way of updating the semantic space with a new unit of context without redoing the whole process. LSA's semantic representations have been largely tested in the literature (Foltz, 1996 ; Wolfe et al., 1998). This model can account for some mechanisms of the construction of knowledge (Landauer & Dumais, 1997).

Hyperspace Analogue to Language

HAL (Burgess, 1998) is also a model of the semantic memory. It is similar to LSA except that (1) it does not take into account higher-order co-occurrences since vectors are just direct co-occurrence vectors; (2) the unit of context is a sliding window of a few words which takes into account the lexical distance between words and (3) updating the semantic space with a new paragraph can be done easily.

Sparse Random Context Representation

SRCR (Sahlgren, 2001, 2002) is also based on the use of a sliding window applied to a large corpus. Words have an initial random vector representation (1,800 dimensions), which is updated with the vectors of the co-occurring words: they are all added to the current word, but with a multiplying factor which depends on their distance to the current word within the window. The way the initial

representation is computed is important: all 1,800 values are set to 0 except eight which are randomly selected and set to 1. This method is intrinsically incremental. It has better results than LSA on the famous TOEFL test. However, it does not take into account higher-order co-occurrences.

Word Association Space

WAS (Steyvers, Shiffrin & Nelson, in press) is not based on a corpus but on association norms providing associates for 5,000 words. The authors applied scaling methods to these data in order to assign a high-dimensional representation to each word. In particular, they relied on singular value decomposition, the mathematical procedure also used by LSA. The idea is similar to LSA: words that appear within similar contexts (i.e. words with similar associative relationships) are placed in similar regions in the space. WAS appeared to be a better predictor of memory performance than LSA.

Features

We will now discuss the six previous features in order to sketch out a model of construction and representation of the long-term semantic memory that would attempt to overcome existing limits.

Input

A corpus of free texts as input is cognitively more plausible than association norms or even a sublanguage of a few propositions (Frank et al. 2003). As humans, we do not obviously construct our semantic representations solely from written data (Glenberg & Robertson, 2000), but there is currently no formalism able to model all perceptual data such that they can be processed by a computational model. In addition, written data, although it is not perfect, seems to cover a large part of our semantic representations (Landauer, 2002).

Representation

Most models are based on a vector representation of word meaning. Dimensions of the semantic space can be the result of a statistical analysis which keeps hundreds of dimensions like in LSA or SCRC (they are therefore unlabelled), the most variant words as in HAL, the most frequent ones (Levy & Bullinaria, 2001) or even a predefined subset of words, either taken from a thesaurus (Prince & Lafourcade, 2002) or selected as being the most reliable across various sub-corpora (Lowe & McDonald, 2000).

One major interest of the vector representation is that it offers a simple way to measure the similarity between words. The angle between the corresponding vectors or its cosine are generally used.

One drawback of the vector representation however is the difficulty to determine the words that are similar to a given word or, say differently, the words that are activated in memory. It requires the scanning of all vectors in order to find the closest ones, which is both computationally and cognitively not satisfactory. A direct link between a word and its associates should exist in a plausible model of the semantic memory.

Another problem with the vector representation is that similarity is symmetrical: similarity(A,B)=similarity(B,A). This is not coherent with psycholinguistic findings showing that semantic similarity is not a symmetrical relation (Tversky, 1977). For instance, *bird* is a very close neighbor of *swallow*, but the opposite is not so obvious.

A network of words with simple numerical oriented links between nodes (what is called an oriented graph in graph theory) would be better for that purpose. Numerical links would represent semantic similarities. Such a basic network would offer a direct connection between a word and its neighbors and represent differently similarity(A,B) and similarity(B,A).

Memory updating

A model of the construction of the semantic memory should describe the way processing a new piece of written data affects the representation of the long-term memory. Some models like Latent Semantic Analysis are not incremental, which means that the whole process needs to be restarted in order to take into account a new context. Actually, a new paragraph can easily be represented by a vector in this model, by a simple linear combination of its words, but this operation does not affect at all the semantic space. Incremental models are much more cognitively plausible: processing new texts should modify, even slightly, the semantic memory.

Unit of context

The semantic relations between words are constructed from the occurrences of words within contexts. The size of such contexts plays an important role. Psychological experiments as well as computer simulations (Burgess, 1998) tend to consider that a context composed of a few words before and after the current word is reasonable. However, computational constraints have led some models to consider a whole paragraph as a unit of context, which is probably a too large unit. Latent Semantic Analysis is such a model. The use of a sliding window allows models like HAL or SRCR to take into account the distance between words within the window, whereas approaches based on paragraphs deal with bags of words.

Higher-order co-occurrences

It has been shown that higher-order co-occurrences play an important role (Kontostathis & Pottenger, 2002) in the latent structure of word usage. Two words should be considered associated although they never co-occur in context units, provided that they occur within similar contexts. A is said to be a second-order co-occurrence of B if it co-occurs with C which also co-occurs with B. If C were a second-order co-occurrence of B, A would be considered as a third-order co-occurrence of B, etc.

By means of the singular value decomposition procedure, LSA semantic similarity indeed involves higher-order co-occurrences (Lemaire & Denhière, submitted). Other approaches such as SCRC or HAL do not.

Table 1: Features of different models

	Input	Representation	Memory updating	Unit of context	higher-order co-occurrences	Compositionality
LSA	corpus	vectors	not incremental	paragraph	yes	easy
HAL	corpus	vectors	incremental	sliding window	no	easy
SCRC	corpus	vectors	incremental	sliding window	no	easy
WAIS	association norms	vectors	not incremental	N/A	no	easy
ICAN	*corpus*	*network*	*incremental*	*sliding window*	*yes*	*hard*

Compositionality

Compositionality is the ability of a representation to go from words to texts. The vector representation is very convenient for that purpose because the linear combination of vectors still produces a vector, which means that the same representation is used for both words and texts. This might be a reason why vector representations are so popular. On the contrary, symbolic representations of word meaning like semantic networks do not offer such a feature: it is not straightforward to build the representation of a group of words from the individual representations of words, especially if the representation is rich, for instance with labelled links.

Summary

Table 1 describes some of the existing models along the previous six features. We present ICAN, our proposal, at then end of the next section.

ICAN

Basic mechanisms

Like others, this model takes as input a corpus of free texts and produces a computational representation of word meanings. This model is based on a network representation, which we believe is more accurate in modeling the process of semantic activation in memory. The idea is to associate to each word a set of neighbors as well as their association weights in [0..1], exactly as in rough semantic network. The model is incremental which means that the set of connected words for each word evolves while processing new texts. In particular, new words can be added according to the co-occurrence information and other words can be ruled out if their association strengths with the current word become too low.

Links between words are updated by taking into account the results of a previous simulation on 13,637 paragraphs of a corpus (Lemaire & Denhière, submitted), which showed that:

- co-occurrence of W_1 and W_2 tends to strongly increase the W_1-W_2 similarity;
- occurrence of W_1 without W_2 or W_2 without W_1 tends to decrease the W_1-W_2 similarity;
- second and third-order co-occurrence of W_1 and W_2 tends to slightly increase the W_1-W_2 similarity.

In our model, a sliding window is used as a unit of context. Therefore, each word of the corpus is considered with respect to its preceding and following contexts. The size of the window can be modified. For the sake of simplicity, we will not use the third-order co-occurrence effect. The algorithm is the following:

For each word W, its preceding context $C_1..C_k$ and its following context $C_{k+1}..C_{2k}$ (the sliding window therefore being $[C_1\ C_2\ ...\ C_k\ \mathbf{W}\ C_{k+1}\ C_{k+2}...\ C_{2k}]$):

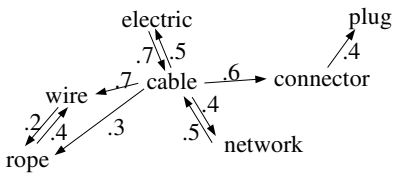
Direct co-occurrence effect: reinforce the link W-C_i (if this link does not exist, create it with a weight of 0.5, otherwise increase the weight p by setting it to $p+(1-p)/2$;

Second-order co-occurrence effect: let p be the weight of the W-C_i link. For each M linked to C_i with weight m, reinforce the link W-M (if such a link does not exist, create it with a weigth of $p.m$, otherwise, increases the weight q by setting it to $q+A(1-q)(p.m)$, A being a parameter;

Occurrence without co-occurrence effect: reduce the links between W and its other neighbors (if the weights were p, set them to a fraction of p, e.g., $0.9p$). If some of them fall under a threshold (e.g., .1), then remove these links.

Example

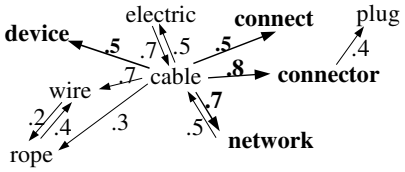
As an example, consider the following association network, which is the result of processing several texts:



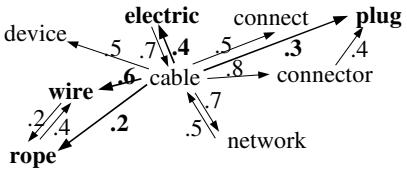
The new text being analyzed is:

... if you have such a device, connect the cable to the network connector then switch...

We now describe how this text will modify the association network, according to the previous rules. Suppose a window of size 5 (2 preceding words, 1 current word, 2 following words). Since functional words are not taken into account, the current window is then [device, connect, **cable**, network, connector], *cable* being the current word. The direct co-occurrence effect leads to reinforce the links between *cable* and the four co-occurring words. Two of them are new links, while others are existing links whose weights are simply increased. The network becomes:



The second-order co-occurrence effect reinforces the links between *cable* and all words connected to one of its four co-occurring word. In this small example, this is only the case for the word *plug*. Finally, the occurrence without co-occurrence effect leads to a decrease of the links between cable and its other neighbors. The network is then:



The next current word is *network*, the window is [connect, cable, **network**, connector, switch] and the process repeats again.

Measure of similarity

Similarity between words W_1 and W_2 is the combination (i.e. the product) of the links of the shortest path between W_1 and W_2. If W_2 is connected to W_1, it is just the weight of the link; if W_1 is connected to Z which is connected to W_2, it is the combination of the two weights. If W_2 does not belong to the neighbors of W_1's neighbors it is probably sufficient to set the semantic similarity to 0. Since the graph is oriented (the link weight between A and B might be different from the link weight between B and A), this way of measuring the similarity mimics the asymmetrical property of the human judgment of similarity better than the cosine between vectors.

Tests

Comparison to association norms

In order to test this model, we compared the association links it provides to human association norms. The corpus we relied on is a 3.2 million word French child corpus composed of texts that are supposed to reproduce the kind of texts children are exposed to: stories and tales for children (~1,6 million words), children productions (~800,000 words), reading textbooks (~400,000 words) and children encyclopedia (~400,000 words). All functional words were ruled out. Words whose frequency was less than 3 were not taken into account. The program is written in C, it is available on demand. Processing the whole corpus takes a few hours on a standard computer, depending on the window size.

Once the association network was built, we measured the similarity between 200 words and 6 of their associates (the first three and the last three), as provided by the de la Haye (2003) norms for 9 year-old children. The association value in these norms is the percentage of subjects who provided the associate. For instance, the six associates to *abeille(bee)* are:

- miel(honey): 19%
- insecte(insect): 14%
- ruche(hive): 9%
- animal(animal): 1%
- oiseau(bird): 1%
- vole(fly): 1%

Actually, 16 words were not part of the corpus. Only 1184 pairs of words were therefore used.

We then compared these values to the similarity values provided by the model. We had two hypotheses. First, the model should distinguish betwen the three first associates and the last ones and there should be a gradient of similarity from the first one to the last ones. Second, there should be a good correlation between human data and model data.

Several parameters have to be set in the model. The best correlation with the human data was obtained with the following parameters (see the algorithm presented earlier):

- window size = 11 (5 preceding and 5 following words);
- co-occurrence effect: $p \rightarrow p+(1-p)/2$;
- 2nd-order co-occurrence effect : $p \rightarrow q+.02(1-q)(p.m)$;
- occurrence without co-occurrence effect : $p \rightarrow .9p$.

Using these parameters, the average similarity values between stem words and associates, as well as the children data, are the following:

	1st associates	2nd associates	3rd associates	Last associates
ICAN	.415	.269	.236	.098
Norms	30.5	13.5	8.2	1

All model values are highly significantly different, except for the 2nd and 3rd associates which differ only at the 10% level. Our model reproduces quite well the human gradient of association.

We also calculated the coefficient of correlation between human data and model data. We found an interesting significant correlation: $r(1184)=.50$.

The exact same test from the same corpus was also applied to Latent Semantic Analysis. Results are the following:

	1st associates	2nd associates	3rd associates	Last associates
LSA	.26	.23	.19	.11

Similarities between the stem word and the first associates appear stronger in the ICAN model. LSA' scorrelation with human data is $r(1184)=.39$, which is worse than our correlation.

Similarity as direct co-occurrence

One can wonder whether the similarity could be mainly due to the direct co-occurrence effect. Similarity between words is indeed often operationalized in psycholinguistic researches by their frequency of co-occurrence in huge corpus. Experiments have indeed revealed the correlation between both factors (Spence & Owens, 1990). However, this shortcut is questionable. In particular, there are words that are strongly associated although they never co-occur. Burgess & Lund (1998) mentioned the two words *road* and *street* that almost never cooccur in their huge corpus although they are almost synonyms. In a 24-million words French corpus from the daily newspaper *Le Monde* in 1999, we found 131 occurrences of *internet*, 94 occurrences of *web*, but no co-occurrences at all. However, both words are strongly associated. Edmonds (1997) showed that selecting the best typical synonym requires that at least second-order co-occurrence is taken into account. There is clearly a debate: is the frequency of co-occurrence a good model of word similarity?

In order to test that hypothesis, we modified our model so that only direct co-occurrences are taken into account: the 2nd order co-occurrence effect as well as the occurrence without co-occurrence effect were inhibited. Results are the following:

	1st assoc.	2nd assoc.	3rd assoc.	Last assoc.
ICAN (only direct co-occurrences)	.903	.781	.731	.439

The gradient of similarity is still there but the correlation with human data is worse ($r(1184)=.39$). This is in accordance with our previous findings (Lemaire & Denhière, submitted) which show that the frequency of co-occurrence tends to overestimate semantic similarity.

Effect of second-order co-occurrence

Another test consisted in measuring the effect of second-order co-occurrences. This time, we only inhibited this effect in order to see whether the loss would be significant. Results are presented in the next table:

	1st assoc.	2nd assoc.	3rd assoc.	Last assoc.
ICAN (no 2nd-order co-occurrences)	.371	.225	.191	.056

Correlation with human data was not significantly different from the full model. It only decreased from .50 to .48. This means that second-order co-occurrences do not seem to have much effect in this simulation. One reason might be due to the mathematical formula we used to model higher-order co-occurrences. It might not be the right one. Another reason could be that we only implemented second-order co-occurrence effects. Third and higher-order co-occurrence effects might play a much more significant role than could be expected. A final reason could be that higher-order co-occurrence does not play any role. But, how then could we explain the high similarity between words that almost never co-occur? More experiments and simulations need to be carried out to investigate this issue.

Window size

We also modified the model in order to shed light on the role of the window size. Results are as follows:

Window size	Correlation with human data
3 (1+1+1)	.34
5 (2+1+2)	.38
7 (3+1+3)	.44
9 (4+1+4)	.48
11 (5+1+5)	.50
13 (6+1+6)	.49
15 (7+1+7)	.47

We found that the best window size is 11 (5 preceding words and 5 following words). This is in agreement with the literature: Burgess (1998) as well as Lowe and McDonald (2000) use a window of size 10, Levy and Bullinaria (1998) found best performance for a window size from 8 to 14, according to the similarity measure they relied on.

Conclusion

This model could be improved in many ways. However, preliminary results are encouraging: the model produces better results than the outstanding Latent Semantic Analysis model on a word association test. In addition, it adresses two major LSA drawbacks. The first one has to do with the representation itself: the fact that LSA's associations are symmetrical is not satisfactory. A network representation seems better for that purpose than a vector representation. The second limitation of LSA concerns the way the semantic space is built. LSA is not incremental: adding a new piece of text requires that the whole process is run again. Like HAL or SCRC, ICAN has the advantage of being incremental.

ICAN's main limitation is related to compositionality. The construction of a text's representation is not straightforward, given the representation of its words. Representing every text as a simple function of its words, and in the same formalism, as in the vector representation, is very convenient since text comparisons are then easy to perform. Compared to other approaches, LSA is for instance very good at simulating the human judgment of text comparisons (Foltz, 1996). However, the cognitive plausibility of such a representation can be questioned. Do we really need the exact same representation for words and texts? Is it cognitively reasonable to go directly without any effort from words to texts? Why having two ways of processing texts: one which would be computationally costly (singular

value decomposition in LSA) and another one very quick (adding its words)?

A solution could be to process each new text by the mechanism described in this paper: a text would then be represented by a subgraph, that is by a small subset of the huge semantic network, composed of the text words, their neighbors and their links. An information reduction mechanism like the integration step of the construction-integration model (Kintsch, 1998) could then be used to condense this subgraph in order to retain the main information. This smaller subgraph would constitute the text representation. This way, there would be a single mechanism used to process a text, construct its representation and update the long-term semantic memory. However, much work remains to be done in that direction.

Once a corpus was processed, it would be interesting to study the resulting network structure. In particular, this structure could be compared to existing semantic networks, in terms of connectivity or average path-lengths between words, much like Steyvers & Tenenbaum (submitted) did recently.

Acknowledgements

We would like to thank Maryse Bianco, Philippe Dessus and Sonia Mandin for their valuable comments on this model, as well as Emmanuelle Billier, Valérie Dupont and Graham Rickson for the proofreading .

References

Burgess, C. (1998). From simple associations to the building blocks of language: modeling meaning in memory with the HAL model. *Behavior Research Methods, Instruments, & Computers*, 30, 188-198.

Burgess, C., Livesay, K & Lund, K. (1998). Explorations in context space: words, sentences, discourse. *Discourse Processes*, 25, 211-257.

de la Haye, F. (2003). Normes d' associations verbales chez des enfants de 9, 10 et 11 ans et des adultes. *L' Année Psychologique*, 103, 109-130.

Edmonds, P. (1997). Choosing the word most typical in context using a lexical co-occurrence network. *Meeting of the Association for Computational Linguistics*, 507-509.

Foltz, P. W. (1996). Latent Semantic Analysis for text-based research. *Behavior Research Methods, Instruments and Computers*, 28-2, 197-202.

Frank, S.L., Koppen, M., Noordman, L.G.M. & Vonk, W. (2003). Modeling knowledge-based inferences in story comprehension. *Cognitive Science* 27(6), 875-910.

Glenberg, A. M. & Robertson, D. A., (2000). Grounding symbols and computing meaning: a supplement to Glenberg & Robertson. *Journal of Memory and Language,* 43, 379-401.

Kintsch, W. (1998). *Comprehension: A Paradigm for Cognition*. Cambridge University Press.

Kontostathis, A. & Pottenger, W.M. (2002). Detecting patterns in the LSI term-term matrix. *Workshop on the Foundation of Data Mining and Discovery, IEEE International Conference on Data Mining.*

Landauer T.K. (2002). On the computational basis of learning and cognition: Arguments from LSA. In N. Ross (Ed.), *The psychology of Learning and Motivation, 41*, 43-84.

Landauer, T. K. & Dumais, S. T. (1997). A solution to Plato's problem : the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104, 211-240.

Lemaire, B. & Denhière, G. (submitted). Effects of higher-order co-occurrences on semantic similarity of words.

Levy, J.P., Bullinaria, J.A. & Patel, M. (1998). Explorations in the derivation of semantic representations from word co-occurrence statistics. *South Pacific Journal of Psychology,* 10, 99-111.

Levy, J.P., Bullinaria, J.A. (2001). Learning lexical properties from word usage patterns: which context words should be used? In R. French & J.P. Sougne (Eds) *Connectionist Models of Learning, Development and Evolution: Proceedings* of *the Sixth Neural Computation and Psychology Workshop*, 273-282. London: Springer.

Lowe, W. & McDonald, S. (2000). The direct route: mediated priming in semantic space. In Gernsbacher, M. A. and Derry, S. D., editors, *Proceedings of the 22nd Annual Meeting of the Cognitive Science Society*, 675-680, New Jersey. Lawrence Erlbaum Associates.

Prince, V. & Lafourcade, M. (2003). Mixing semantic networks and conceptual vectors: the case of hyperonymy. In *Proc. of ICCI-2003 (2nd IEEE International Conference on Cognitive Informatics),* South Bank University, London, UK, August 18 - 20, 121-128.

Sahlgren, M. (2001). Vector-based semantic analysis: representing word meaning based on random labels. *Semantic Knowledge Acquisition and Categorisation Workshop at ESSLLI '01*, Helsinki, Finland.

Sahlgren, M. (2002). Towards a flexible model of word meaning. *AAAI Spring Symposium 2002*. March 25-27, Stanford University, Palo Alto.

Spence, D.P. & Owens K.C. (1990). Lexical co-occurrence and association strength. *Journal of Psycholinguistic Research* 19, 317-330.

Steyvers, M., Shiffrin R.M., & Nelson, D.L. (in press). Word Association Spaces for predicting semantic similarity effects in episodic memory. In A. Healy (Ed.), *Cognitive Psychology and its Applications: Festschrift in Honor of Lyle Bourne, Walter Kintsch, and Thomas Landauer.* Washington DC: American Psychological Association.

Steyvers, M., & Tenenbaum, J. (submitted). Graph theoretic analyses of semantic networks: small worlds in semantic networks.

Tversky, A. (1977). Features of similarity. *Psychological Review*, 84(4), 327-352.

Wolfe, M. B. W., Schreiner, M. E., Rehder, B., Laham, D., Foltz, P. W., Kintsch & W., Landauer, T. K. (1998). Learning from text: Matching readers and texts by Latent Semantic Analysis. *Discourse Processes, 25, 309-336.*