How Well Can a Corpus-Derived Co-Occurrence Network Simulate Human Associative Behavior?

Gemma Bel Enguix Reinhard Rapp Michael Zock

Aix-Marseille Université, Laboratoire d'Informatique Fondamentale
UMR 7279, Case 901, 163 Avenue de Luminy, F-13288 Marseille

gemma.belenguix@gmail.com reinhardrapp@gmx.de zock@free.fr

Abstract

Free word associations are the words people spontaneously come up with in response to a stimulus word. Such information has been collected from test persons and stored in databases. A well known example is the Edinburgh Associative Thesaurus (EAT). We will show in this paper that this kind of knowledge can be acquired automatically from corpora, enabling the computer to produce similar associative responses as people do. While in the past test sets typically consisted of approximately 100 words, we will use here a large part of the EAT which, in total, comprises 8400 words. Apart from extending the test set, we consider different properties of words: saliency, frequency and part-of-speech. For each feature categorize our test set, and we compare the simulation results to those based on the EAT. It turns out that there are surprising similarities which supports our claim that a corpus-derived co-occurrence network can simulate human associative behavior, i.e. an important part of language acquisition and verbal behavior.

1 Introduction

Word associations in general and free word association in particular (Galton, 1879) have been used by psychologists of various schools[1] to understand the human mind (memory, cognition, language) and the hidden mechanisms driving peoples' thoughts, utterances, and actions. In the case of free word associations, a person typically hears or reads a word, and is asked to produce the first other word coming to mind. Kent & Rosanoff (1910) have used this method for comparisons, introducing to this end 100 emotionally neutral test words. Having conducted the first large scale study of word associations (1000 test persons) they reached the conclusion that there was a great uniformity concerning people's associations, that is, speakers of a language share stable, comparable associative networks (Istifci, 2010).

In this paper, we are mainly interested in the automatic acquisition of associations by computer. More precisely, we want to check whether a corpus-based method allows us to build automatically an associative network akin to the one in peoples' mind, that is, a network able to mimic human behavior. This means, given a stimulus word the system is supposed to produce the same responses as people do. We know since the old Greeks that thoughts and their expressions (words) are linked via associations. Yet, what we still do not know is the nature of these links. Also, links vary in terms of strength. Associationist learning theory (Schwartz & Reisberg, 1991) explains how these strengths (or weights) are acquired. The strength between two perceived events increases by a constant fraction of a maximally possible increment at each co-occurrence, and decreases in the opposite case.

Wettler et al. (2005) have shown that this mechanism can be replicated by looking at word co-occurrence frequencies in large text collections. But there had been earlier corpus-linguistic work: For example, Wettler & Rapp (1989) compared several association measures in order to find search terms to be used for queries in information retrieval. Church & Hanks (1990) suggested to use *mutual information*, an information theoretic measure, for computing association strength. Prior to this, a lot of work had been done without reliance of corpora. For example, Collins & Loftus (1975) used associative semantic networks to show the distance between words. Others (Rosenzweig, 1961:358; Ekpo-Ufot, 1978) tried to show the universal status of a large subset of associations. While all these findings are important, we will not consider them further

[1] For example, *cognitive psychology* (Collins and Loftus, 1975,), *psycholinguistics* (Clark, 1970) and *psychoanalysis* (Freud, 1901; Jung & Riklin, 1906).

Proc. of 5th Workshop on Cognitive Aspects of Computational Language Learning (CogACLL) @ EACL 2014, pages 43–48,
Gothenburg, Sweden, April 26 2014.

here. Rather we will focus on the claim that a -corpus-derived co-occurrence network is able to mimic human associative behavior.

Such a network consists of nodes, which in our case correspond to words (or lemmas), and of weights connecting the nodes. The strengths of these weights are computed on the basis of word co-occurrence data, and by optionally applying an association measure. But there are many association measures. Given their number and diversity some researchers (Evert & Krenn, 2001) felt that there was a need to define some criteria and methods in order to allow for quantitative comparisons via task-based evaluations. Pursuing a similar goal, Pecina & Schlesinger (2006) compared 82 different association measures for collocation extraction, while Hoang et al. (2009) classified them. Michelbacher et al. (2011) investigated the potential of asymmetric association measures, i.e. "associations whose associational strength is significantly greater in one direction (e.g., from *Pyrrhic* to *victory*) than in the other (e.g., from *victory* to *Pyrrhic*)". Washtell & Markert (2009) tried to determine whether word associations should be computed via window-based co-occurrence counts or rather via a windowless approach measuring the distances between words.

Our work is related to previous studies comparing human word associations with those derived from corpus statistics (e.g. Wettler et al., 2005; Tamir, 2005, Seidensticker, 2006). The main differences are that we categorize our stimulus words and present results for each class, and that we have a stronger focus on the graph aspect of our network.

2 Resources and processing

In order to simulate human associative behavior via corpora, we need them to encode knowledge that people typically have, that is, encyclopedic or universally shared knowledge (e.g. Paris capital of France) and episodic knowledge (i.e. knowledge momentarily true: Nadal winner of the French Open). To meet these goals we decided to use the *British National Corpus* (BNC, Burnard & Aston, 1998) as it is well balanced and relatively large (about 100 million words of contemporary British English).

To lemmatize the corpus we used the NLTK (Bird et al., 2009) which for this purpose utilizes information from WordNet. Hence, inflected forms (e.g. *wheels* or *bigger*) were replaced by their base forms (e.g. *wheel* or *big*). This reduces noise and data sparsity while improving speed and accuracy during evaluation. Since this latter is based on exact string matching, our system would consider *wheels*, produced in response to *car*, as a mistake as the primary associative response of the test persons is *wheel*, the singular form. Lemmatization solves this problem. Since we were interested here only in content words (nouns, verbs, and adjectives) we removed all other words from the BNC.

To evaluate the performance of our system we compared its results with the associations collected by Kiss et al. (1973), the *Edinburgh Associative Thesaurus*. The association norms of the EAT were produced by presenting each stimulus word to 100 subjects, and by collecting their responses. The subjects were 17 to 22 year old British students. Table 1 shows the associations produced by at least five participants in response to the stimulus words *bath* and *cold* together with the number of participants producing them.

bath		*cold*	
observed response	number of subjects	observed response	number of subjects
water	20	hot	34
tub	8	ice	10
clean	5	warm	7
hot	5	water	5

Table 1: Extracts from the EAT for the stimulus words *bath* and *cold*.

The EAT lists the associations to 8400 stimulus words. Since we were only interested in nouns, verbs, and adjectives, we eliminated all other words and also multiword units (e.g. *a lot*). After having lemmatized the data with the NLTK we obtained a list of 5910 test items which is considerably more than the usual 100 used in many previous studies (e.g. Wettler et al., 2005).

3 A graph-based approach for computing word associations

Unlike previous work (Wettler et al. 2005; Church & Hanks, 1990) which is described in the terminology of the well known vector space model, in the construction of the current system we had a graph-based approach in mind so we describe the system in such terms. We built up a graph on the basis of the nouns, verbs, and adjectives occurring in the corpus, these tokens being the nodes of the graph.[2] The links (also called

[2] As preliminary experiments have shown, including function words in the graph can create noise in the retrieval of

weights, connections, or edges) between these nodes are zero at the beginning, and are incremented by one whenever the two connected words co-occur in the corpus as direct neighbors.[3] Put differently, the weight of each link represents the number of times two words (nodes) co-occur in the corpus.

The associations to a given stimulus word are calculated by searching the nodes which are direct neighbors of this stimulus word, and by ranking them according to the weights of the connections. Given a graph G=V,E with V={i,j,…,n} as its set of vertices and E as its set of edges linking pairs of nodes over V, we express by N(i) the neighborhood of a node $i \in V$, where N(i) is defined as every $j \in V \mid e_{i,j} \in E$.

4 Results

Given the way this network is built, one could expect the system to retrieve only syntagmatically related words, i.e. words often occurring in close proximity (e.g. *blue* → *sky*). Yet, to our surprise, the system also retrieves many paradigmatic associations, that is, words which can substitute each other (e.g. *blue* → *red*).

Table 2 shows some results. While not all computed primary responses are identical to the ones produced by humans (in the EAT), the responses seem perfectly plausible. This raises the question whether the answers are within the bandwidth of variation of human associative behavior.

We measured the quality of our results by counting (for all 5910 items) the number of times the subjects participating in the creation of the EAT had given the same answer as our system. This number is 6.2 on average. In comparison, the number of other subjects giving the same answer as an average test person is 5.8. If the two numbers were identical, our system would be perfectly within the range of variation of the human associative responses, i.e. our system's answers could hardly be distinguished from the ones given by a human. This is actually the case. The answers of our system are, on average, even slightly closer to the ones given by the test persons than the answers of a randomly selected test person.

associations. Hence we preferred to keep only these three categories.

[3] Note that this refers to the pre-processed corpus where all stopwords have been removed.

Stimulus Word	Human Primary Response	Computed Primary Response
afraid	fear	person
anger	hate	frustration
baby	boy	mother
bath	water	shower
beautiful	ugly	woman
bed	sleep	hospital
bible	book	God
bitter	sweet	taste
black	white	white
blossom	flower	white

Table 2: Comparison between human and computed associations for the 10 alphabetically first words of the Kent/Rosanoff (1910) list.

In the following subsections we split our set of 5910 test items into three categories to check how well each one of them matches our intuition that a corpus-derived co-occurrence network can indeed simulate human associative behavior.

4.1 Word saliency

Our goal is twofold: find out to what extend the saliency of a stimulus word has an effect on the homogeneity of human responses, and whether these findings can also be replicated in our computer simulation.

To this end we divided our 5910 EAT stimulus words into six categories, i.e. saliency classes (SC). *Saliency* is defined here as the proportion of subjects producing the *Primary Associative Response* (PAR), this latter being the response produced by the largest number of subjects.

SC 1: less than 10% producing the PAR (10.7%)
SC 2: 10 to 20% producing the PAR (36.0%)
SC 3: 20 to 30% producing the PAR (24.3%)
SC 4: 30 to 40% producing the PAR (13.3%)
SC 5: 40 to 50% producing the PAR (8.0%)
SC 6: more than 50% producing the PAR (7.6%)

The percentages at the end of each line denote the proportion of words belonging to the respective saliency class. All classes are reasonably well covered. Here are some representative words for each class:

SC 1: leader, professor, yellow
SC 2: horse, mountain, semaphore
SC 3: chief, jungle, kiss
SC 4: driver, monarchy, tornado
SC 5: aid, cell, gasoline
SC 6: black, aunt, woman

As can be seen from these examples, our intuitions do not easily allow us to make predictions concerning the saliency classifications of words.

Figure 1 (blue curve) shows how well our system performs for each class. For the words in each class we counted the average number of times a human subject had come up with the same associative response as the system. It appears that the system's performance is best for very salient words, performing less well in the opposite case. Note that this correlates perfectly well with the observed human associative behavior: Our system tends to produce the same answers as people for stimulus words yielding homogeneous human responses. Likewise, the system's answers tend to differ in cases where peoples' answers are heterogeneous.

The red curve in Figure 1 shows for each saliency class the number of persons giving the same associative answer as an average test person. As can be seen this line is almost identical to the one representing the system's performance, which means that the system's behavior is very similar to human behavior with respect to saliency.

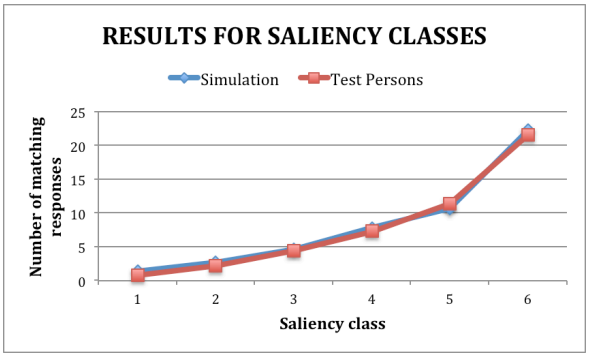

Fig. 1: Quality of our system's (blue curve) and an average test person's (red curve) performance (measured as the number of matching responses found in the EAT) with respect to saliency.

4.2 Word frequency

Encouraged by the findings for saliency, we conducted a similar experiment for word frequency. In this case the EAT stimulus words were split into frequency classes according to their corpus frequencies in the BNC.

Since a logarithmic scale seems to be appropriate for word frequencies (Rapp, 2005; van Heuven et al., in press), we used the following six frequency classes (FC):

FC1: 1 occurrence BNC (0.5%)
FC2: from 1 to 10 occurrences BNC (9.2%)
FC3: form 10 to 100 occurrences BNC (30.2%)
FC4: from 100 to 1000 occurrences BNC (42.6%)
FC5: from 1000 to 10000 occurrences BNC (17.3%)
FC6: from 10000 to 100000 occurrences BNC (0.1%)

As can be seen from the percentages at the end of each line, extremes, i.e. very high and very low frequencies are covered only marginally.

In the first group we find words like *cornucopia*, *jewelry*[4] and *quaff*, each appearing only once in the corpus, while the frequency class 6 contains only high frequency words such as the (auxiliary) verbs *be*, *do*, *have*, and *make*.

The results obtained for the frequency classes are shown in Figure 2. As can be seen, the general tendency is that the results improve with decreasing frequency. Our explanation for this is that frequent words tend to be more polysemous, and that increased ambiguity tends to yield more heterogeneous responses. For example, the ambiguous stimulus word *palm* is likely to evoke not only responses related to its *tree* sense, but also to its *hand* sense.

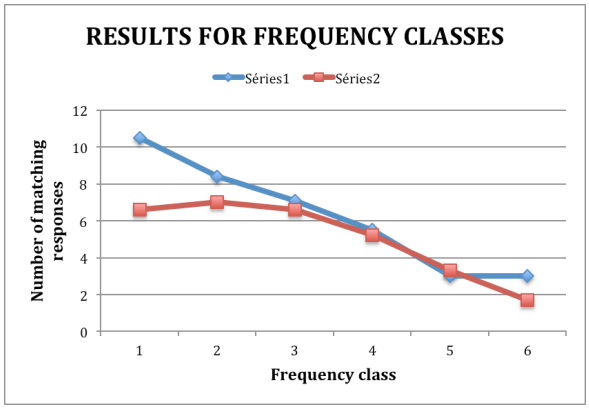

Fig. 2: Quality of our system's (blue curve) and an average test person's (red curve) performance with respect to frequency.

Whereas for mid frequency words the results for the test persons and in the simulation show a high agreement, this is not the case for high frequency and for low frequency words. For high frequency words (FC 6) a plausible explanation might be the sampling error due to the low sample size of only 0.1% of the stimulus words in the EAT test set. However, for low frequency words the sample sizes are larger and the discrepancy is clearly systematic. Our explanation is that in this case we might have a systematic sampling error concerning the observed frequencies. The simulation has an advantage because the frequency classes were set up according to

[4] Note that this is the American spelling which is rare in the BNC. The British spelling is *jewellery*.

the BNC frequencies rather than according to the subjective frequencies (= word familiarities) of the test persons. For example, the words of FC 1 are guaranteed to occur in the BNC, while it is not certain at all that the test persons ever encountered them. This leads to a systematic bias in favor of the simulation results.

4.3 Part of speech

In a last experiment we considered the results for the three parts of speech used in our system, namely nouns, verbs, and adjectives. We assigned to each word in the EAT test set its part of speech. Syntactically ambiguous words (which can belong to several parts of speech) were assigned to their most frequently occurring part of speech. Of the 5910 EAT items, 89.2% were classified as nouns, 2.4% as verbs, and 8.4% as adjectives.

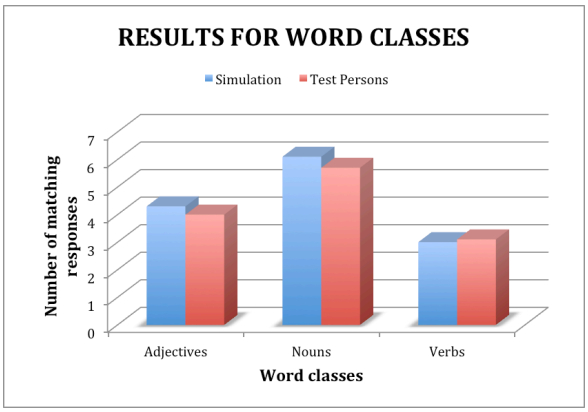

Fig. 3: Quality of our system's (blue curve) and an average test person's (red curve) performance with respect to parts of the speech.

For the three categories we obtained the results shown in Figure 3. The results are best for nouns and worst for verbs. Our explanation for this is once again average word ambiguity which is higher for verbs than it is for nouns. As with the saliency classes, we have again a high correlation between the results produced by humans and the ones produced by machine.

5 Discussion and conclusion

We have presented a novel graph-based algorithm for the computation of word associations. The goal was to check whether and to what extent an automatically built association network based on a large text corpus would yield similar results to the ones produced by humans. The results were evaluated with a test set comprising all nouns, verbs, and adjectives of the EAT stimulus words. This test set is considerably larger than the ones used in most previous computational association studies.

Contrary to what could be expected our system predicts not only syntagmatic but also paradigmatic relations. For instance, the pairs *black* → *white*, *bread* → *butter* and *boy* → *girl* are correctly computed. This shows that texts contain not only word pairs encoding syntagmatic relations but also pairs encoding paradigmatic relations. The results also show that statistical co-occurrence-based methods are suitable for tasks that traditionally were supposed to require more sophisticated symbolic approaches.

In sum, our approach allows not only to correctly predict thousands of associations, it also matches human performance in other respects: For the first time it was shown that the predictions for salient words are much better than for non-salient ones. Similarly, concerning word frequency and part of speech the simulated results also closely mimic the behavior as found in the human data.

Altogether, our results provide evidence that human associative behavior as observed in the classical association experiments can be modeled by exploiting the co-occurrences of words in large text corpora. There seems to be a circularity: (a) the word co-occurrences found in text and speech[5] appear to be externalized forms of the associations stored in the human brain, and (b) the associations stored in the brain appear to be internalized forms of the co-occurrences as found in text and speech. This contradiction disappears as soon as we realize that time has elapsed between these two events. Hence, one network may be fed by the other, and this may go on.

Note that our corpus-based approach has further virtues: (a) it allows to generate associations from corpora covering particular time spans; (b) it can produce associations based on corpora covering specific topics; (c) it accounts for the fact that languages, hence associations, change over time. Think of the ideas associated with Dominique Strauss-Kahn, one of the top candidates before the last presidential campaign in France. While the associations prior to May 18, 2011 were probably IMF, politics or election, the ones after the Sofitel event were probably quite different, shifting towards a much more delicate topic.

[5] Note that the BNC also contains transcribed speech.

Acknowledgments

This research was supported by the Marie Curie Intra European Fellowships DynNetLAc and AutoWordNet within the 7th European Community Framework Programme.

References

Bird, S.; Klein, E. and Loper, E. (2009). *Natural Language Processing with Python*. O'Reilley Media.

Burnard, L. and Aston, G. (1998). *The BNC Handbook: Exploring the British National Corpus*. Edinburgh: Edinburgh University Press.

Church, K.W. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics* 16 (1), 22–29.

Clark, H. H. (1970). Word associations and linguistic theory. In J. Lyons (Ed.), New horizons in linguistics (pp. 271-286). Baltimore: Penguin.

Collins, A. M. and Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review* 8. Vol. 82, No. 6, 407-428.

Ekpo-Ufot, A. (1978). Word associations: a comparative study among college students in Nigeria and the United States. *Journal of Cross-Cultural Psychology*, Vol. 9(4), 455-468.

Evert, S. and Krenn, B. (2001). Methods for qualitative evaluation of lexical association measures. In *Proceedings of the 39th Annual Meeting of the Association of Computational Linguistics,* Toulouse, France, 188-915.

Freud, S. (1901/1975). The psychopathology of everyday life. Harmondsworth: Penguin. http://psychclassics.yorku.ca/Freud/Psycho/chap5.htm

Galton, F. (1879). Psychometric experiments. *Brain (2)*, 149-162.

Van Heuven, W.J.B., Mandera, P., Keuleers, E., & Brysbaert, M. (in press). Subtlex-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology*.

Hoang, H.H, Kim, S. N. and Kan, M.Y. (2009). A re-examination of lexical association measures. *Proceedings of the Workshop on Multiword Expressions*, ACL-IJCNLP 2009, Suntec, Singapore, 31-39.

Istifci, I. (2010). Playing with words: a study on word association responses. *The Journal of International Social Research,* 3(10), 360–368

Jung, C. and F. Riklin. 1906. Experimentelle Untersuchungen über Assoziationen Gesunder. In Jung, C. G., editor, Diagnostische Assoziationsstudien, 7–145. Barth, Leipzig.

Kent, G.H. and Rosanoff, A.J. (1910). A study of association in insanity. *American Journal of Insanity*, 67, 37–96, 317–390.

Kiss, G.R., Armstrong, C., Milroy, R., and Piper, J. (1973). An associative thesaurus of English and its computer analysis. In: A. Aitken, R. Beiley, N. Hamilton-Smith (eds.): *The Computer and Literary Studies*. Edinburgh University Press.

Michelbacher, L., Evert, S. and Schütze, H. (2011). Asymmetry in corpus-derived and human associations. Corpus Linguistics and Linguistic Theory, Vo. 7, No. 2, 245–276.

Pecina, P., and Schlesinger, P. (2006). Combining association measures for collocation extraction. *Proceedings of the 21th International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL 2006)*, Sydney, Australia, 651-658.

Rapp, R. (2005). On the relationship between word frequency and word familiarity. In: B. Fisseni; H.-C. Schmitz; B. Schröder; P. Wagner (Hg.): *Sprachtechnologie, mobile Kommunikation und linguistische Ressourcen. Beiträge zur GLDV-Tagung 2005 in Bonn.* Frankfurt: Peter Lang. 249–263.

Rosenzweig, M. R. (1961). Comparisons among word-assocation responses in English, French, German, and Italian. *The American Journal of Psychology*, Vol. 74, No. 3, 347-360.

Schwartz, B. and Reisberg, D. (1991). *Learning and Memory*. New York: Norton.

Seidensticker, P. (2006). *Simulation von Wortassoziationen mit Hilfe von mathematischen Lernmodellen in der Psychologie*. Dissertation an der Universität Paderborn.

Tamir, R. (2005). A Random Walk through Human Associations. *Proceedings of ICDM 2005*: 442-449.

Washtell, J.; Markert, K. (2009). A comparison of windowless and window-based computational association measures as predictors of syntagmatic human associations. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP '09)*, Volume 2, 628-637

Wettler, M. and Rapp, R. (1989). A connectionist system to simulate lexical decisions in information retrieval. In: R. Pfeifer, Z. Schreter, F. Fogelman, L. Steels (eds.): *Connectionism in Perspective*. Amsterdam: Elsevier, 463–469.

Wettler, M., Rapp, R. and Sedlmeier, P. (2005). Free word associations correspond to contiguities between words in texts. *Journal of Quantitative Linguistics* 12(2), 111–122.