A Two-Stage Approach for Computing Associative Responses to a Set of Stimulus Words

Urmi Ghosh, Sambhav Jain and Soma Paul
Language Technologies Research Center
IIIT-Hyderabad, India
{urmi.ghosh, sambhav.jain}@research.iiit.ac.in,
soma@iiit.ac.in

Abstract

This paper describes the system submitted by the IIIT-H team for the CogALex-2014 shared task on multiword association. The task involves generating a ranked list of responses to a set of stimulus words. The two-stage approach combines the strength of neural network based word embeddings and frequency based association measures. The system achieves an accuracy of 34.9% over the test set.

1 Introduction

Research in psychology gives evidence that word associations reveal the respondents' perception, learning and verbal memories and thus determine language production. Hence, it is possible to simulate human derived word associations by analyzing the statistical distribution of words in a corpus. Church and Hanks (1990) and Wettler and Rapp (1989) were amongst the first to devise association measures by utilizing frequencies and co-occurrences from large corpora. Wettler and Rapp (1993) demonstrate that corpus-based computations of word associations are similar to association norms collected from human subjects.

The CogALex-2014 shared task on multi-word association involves generating a ranked list of response words for a given set of stimulus words. For example, the stimulus word *bank* can invoke associative responses such as *river*, *loan*, *finance* and *money*. Priming[1] *bank* with *bed* and *bridge*, results in strengthening association with the word *river* and it emerges as the best response amongst the aforementioned response choices. This task is motivated by the tip-of-the-tongue problem, where associated concepts from the memory can help recall the target word. Other practical applications include query expansion for information retrieval and natural language generation where missing words can be predicted from their context.

The participating systems are distinguished into two categories - *Unrestricted* systems that allows usage of any kind of data and *Restricted* systems that can only make use of the *ukWaC* (Baroni et al., 2009) corpus, consisting of two billion tokens. Our proposed system falls in the *restricted* track since we only used *ukWaC* for extracting information on word associations. It follows a two-staged approach: *Candidate Response Generation*, which involves selection of words that are semantically similar to the primes and *Re-ranking by Association Measure*, that re-ranks the responses using a proposed weighted Pointwise Mutual Information ($wPMI$) measure. Our system was evaluated on test-datasets derived from the Edinburgh Associative Thesaurus (Kiss et al., 1972) and it achieved an accuracy of 34.9%. When ignoring the inflectional variations of the response word, an accuracy of 39.55% was achieved.

2 Observations on Training Data

The training set consists of 2000 sets of five words (multiword stimuli or primes) and the word that is most closely associated to all of them (associative response). For example, a set of primes such as *wheel*, *driver*, *bus*, *drive* and *lorry* are given along with the expected associative response - *car*.

In this section, our initial observations on the given training data are enlisted.

[1]The phenomenon of providing multiple stimulus words is called *priming*.

Zock/Rapp/Huang (eds.): Proceedings of the 4th Workshop on Cognitive Aspects of the Lexicon, pages 15–21,
Dublin, Ireland, August 23, 2014.

2.1 Relation between the Associative Response and the Prime Words

It is observed that a response largely exhibits two kind of relations with a priming word.

Primes	Associative Response
presents, Christmas, birthday, shops, present	gifts
butterfly, light, ball, fly, insect	moth
mouse, cat, catcher, race, tail	rat

Table 1: Some examples of primes and their associative responses from the training set

Type A relation depicts a synonymous/antonymous behavior or "of the same kind" nature. Word pairs with paradigmatic relation are highly semantically related and belong to the same part of speech. And hence, they tend to show a substitutive nature amongst themselves without affecting the grammar of the sentence. From Table - 1, we observe that *present/presents* , *butterfly/insect* and *mouse/cat* can be substituted in place of *gifts*, *moth* and *rat* respectively. *Type B* relation depicts contextual co-occurrence, where the words tend to occur together or form a collocation. This kind of relationship can be demonstrated by taking examples from Table - 1, such as *Christmas gifts*, *gift shops*, *birthday gifts*, *moth ball*, *rat catcher*, *rat race* and *rat tail*. In theory, the above have been formally categorized as paradigmatic (*Type A*) and syntagmatic (*Type B*) relations by De Saussure et al. (1916) and we will be referring to them accordingly in rest of the paper.

Type C relation, depicting associations based on the phonological component of the words was also observed. According to McCarthy (1990), responses can be affected by phonological shapes and orthographic patterns especially when instantaneous paradigmatic or syntagmatic association is difficult. Examples from the training data set include *ajar-Ajax*, *hypothalamus-hippopotamus* and *cravat-caravan*. Such examples were very few and hence, have not been dealt with in this paper.

2.2 Context Window Size

Words exhibiting syntagmatic associations often occur in close proximity in the corpus. We tested this phenomenon on 500 randomly chosen sets of primes by calculating the distance of each prime from the associative responses in the corpus. Figure - 1 testifies that a majority of primes occur within a context window size of ± 2 from the associative response.

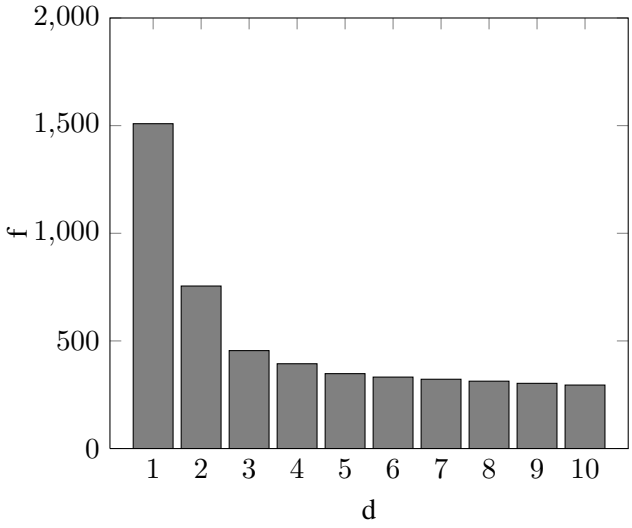

Figure 1: Co-occurrence frequency *f* of an association at distance *d* from the response, averaged over the 2500 stimulus word and response word pairs from randomly chosen 500 training datasets

Next, a mechanism to interpret the above associations in a quantitative manner is required.

3 Word Representation

In order to have a quantitative comparison of association, first we need a representation for words in a context. Traditionally co-occurrence vectors serve as a simple mechanism for such a representation. However, such vectors are unable to effectively capture deeper semantics of words and also tend to suffer from sparsity due to high dimensional space (equal to the vocabulary size). Several efforts have been made to represent word vectors in a lower dimensional space. Largely, these can be categorized into:

1. **Clustering**: Clustering algorithms like Brown et al. (1992), are used to form clusters and derive a vector based representation for each cluster, where semantically similar clusters are closer in distance.

2. **Topic Modeling**: In this approach a word (or a document) is represented as a distribution of topics. Latent Semantic Analysis (LSA) (Deerwester et al., 1990; Landauer and Dutnais, 1997) , which falls in this category, utilizes SVD (Singular Value Decomposition) to produce a low rank representation of a word. Latent Dirichlet Allocation (Blei et al., 2003) is an improvement with dirichlet priors over the probabilistic version of LSA (Hofmann, 1999).

3. **Neural Network based Word Embeddings**: Here, a neural network is trained to output a vector corresponding to a word which effectively signifies its position in the semantic space. There has been different suggestions on the nature of the neural-net and how the context needs to be fed to the neural-net. Some notable works include Collobert and Weston (2008), Mnih and Hinton (2008), Turian et al. (2010) and Mikolov et al. (2013a).

4 Methodology

Our system follows a two-staged approach, where we first generate response candidates which are semantically similar to prime words, followed by a re-ranking step where we give weightage to the responses likely to occur in proximity.

4.1 Candidate Response Generation

The complete vocabulary (of *ukWaC* Corpus) is represented in a semantic space by generating word embeddings induced by the algorithm described in Mikolov et al. (2013a). Our choice is motivated by the fact that this approach models semantic similarity and outperforms other approaches in terms of accuracy as well as computational efficiency(Mikolov et al., 2013a; Mikolov et al., 2013c).

The *word2vec*[2] utility is used to learn this model and thereby create 300-dimensional word embeddings. *word2vec* implements two classification networks - the Skip-gram architecture and the Continuous Bag-of-words (CBOW) architecture. We applied CBOW architecture as it works better on large corpora and is significantly faster than Skip-gram(Mikolov et al., 2013b). The CBOW architecture predicts the current word based on its context. The architecture employs a feed forward neural network, which consists of:

1. An *input layer*, where the context words are fed to the network.
2. A *projection layer*, which projects words onto continuous space and reduces number of parameters that are needed to be estimated.
3. An *output layer*.

This log-linear classifier learns to predict words based on its neighbors in a window of ± 5. We also applied a minimum word count of 25 so that infrequent words are filtered out.

With the vector representation available, a response r to a set of primes S, is searched in the vocabulary by measuring its cosine similarity with each prime x_i in S. The overall similarity of the response r, with the prime word set S, is defined as the average of these similarities.

[2] *word2vec* : https://code.google.com/p/word2vec/

$$sim(r, S) = \frac{1}{|S|} \times \sum_{i=1}^{|S|} \frac{x_i.r}{|x_i|.|r|}$$

Using the best similarity score as the selection criterion for response, the approach resulted in an accuracy of 20.8% over the test set. Error analysis revealed that the above approach is biased towards finding a paradigmatic candidate. However, it is further observed that much of the correct answers ($> 80\%$) exist in a k-best(k=500) list but with a relatively lower similarity score. This confirmed that our broader selection is correct but a better re-ranking approach is required.

4.2 Re-ranking by Association Measures

To give due weightage to responses with high syntagmatic associativity, we utilize word co-occurrences from the corpus. Since we are dealing with semantically related candidates, applying even a basic lexical association measure like Pointwise Mutual Information (PMI) (Church and Hanks, 1990) tend to improve the results.

PMI

For each prime word, we calculate co-occurrence frequency information for its neighbors within a window of ± 2 as mentioned in Section 2. Also, a threshold of 3 is set to the observed frequency measures as PMI tends to give very high association score to infrequent words.

For each candidate response r, we calculate its PMI_i with each of the primes (x_i) in the set S. The total association score $Score_{PMI}$ for a candidate is defined as the average of the individual measures.

$$PMI_i = \frac{p(x_i r)}{p(x_i)p(r)} \qquad\qquad Score_{PMI} = \frac{1}{|S|} \times \sum_{i \in S} PMI_i$$

Ranking the candidates based on PMI improved the results to 30.45%

Weighted PMI

It should be duly noted that only some primes exhibit a syntagmatic relation with the response, while the rest exhibit a paradigmatic relation. For example, the expected response for primes *Avenue*, *column*, *dimension*, *sixth*, *fourth* is *fifth*. The first three words share a syntagmatic relation with the response while the last two words share a paradigmatic relation with the response. As PMI deals with word co-occurrences, ideally, only primes exhibiting syntagmatic associations should be considered for re-ranking. However, a clear distinction between the two categories of primes is a difficult task as the target response is unknown.

In order to take effective contribution of each prime, we propose a weighed extension of PMI which gives more weightage to syntagmatic primes as to the paradigmatic ones. Since, primes sharing a paradigmatic relation with the response word are highly semantically related, they are expected to be closer in the semantic space too. On the other hand, the primes showcasing syntagmatic relations are expected to be distant.

Using the vector representation described in Section 4.1, we calculate an average vector of the five primes, p_{avg}, and compute its cosine distance from individual primes. The cosine distance thus obtained is used as the weight w for the PMI associativity of a prime. In a nutshell, larger the distance of a prime from p_{avg}, the greater is its contribution in the PMI based re-ranking score. This ranking schema assumes that the prime set consists of at least two words demonstrating paradigmatic relation with the target response. Table - 2 displays the primes along with their distance from p_{avg}.

$$Score_{wPMI} = \frac{1}{|S|} \times \sum_{i \in S} w_i PMI_i$$

Next, a ranked list of candidate responses for each set is generated by sorting the previously ranked list according to the new score. The new ranking scheme based on weighted PMI ($wPMI$) improves the results to 34.9%. Table -3 displays some sets which show improvement upon implementing the $wPMI$

Primes	Cosine Distance
Avenue	0.612
column	0.422
dimension	0.390
sixth	0.270
fourth	0.212

Table 2: An example demonstrating Cosine Distance between the primes and the p_{avg} of the prime set

ranking scheme. Taking a case from Table - 3, we observe that the correct response *skeleton* is generated for primes *cupboard*, *body*, *skull*, *bone* and *bones* when ranked according to the $wPMI$ scheme. This is due to larger weights being assigned to primes *cupboard* and *body* which have a closer proximity to the word *skeleton* than the word *vertebral* which is generated by the simple PMI ranking scheme.

Primes(with weights)	cupboard	0.615	pit	0.553	boat	0.499
	body	0.410	band	0.549	sailing	0.476
	skull	0.248	hand	0.426	drab	0.338
	bone	0.244	limb	0.0.340	dark	0.318
	bones	0.172	leg	0.270	dull	0.307
PMI	vertebral		amputated		drizzly	
$wPMI$	skeleton		arm		dingy	
Expected Response	skeleton		arm		dingy	

Table 3: Comparison between results from PMI and wPMI re-ranking approaches

5 Results and Evaluation

The system was evaluated on the test set derived from the Edinburgh Associative Thesaurus (EAT) which lists the associations to thousands of English stimulus words as collected from native speakers. For example, for the stimulus word *visual* the top associations are *aid*, *eyes*, *aids*, *see*, *eye*, *seen* and *sight*. For the shared task, top five associations for 2000 randomly selected stimulus words were provided as prime sets and the system was evaluated based on its ability to predict the corresponding stimulus word for each set. Table - 4 displays the top ten responses generated by our system for some prime sets and their corresponding stimulus word.

Primes	*knight, plate, soldier, protection, sword*	*ants, flies, fly, bees, bite*	*babies, baby, rash, wet, washing*	*butterfly, moth, caterpillar, cocoon, insect*
Top 10 Responses	armor armour helmet shield guard bulletproof guards warrior enemy gallant	mosquitoes wasps beetles insects spiders sting moths butterflies arachnids bedbugs	nappy shaving nappies clothes skin bathing dry eczema bedding dirty	larva larvae pupa species pests beetle silkworm wings pupate pollinated
Target	*armour*	*insects*	*nappies*	*chrysalis*

Table 4: Top ten responses for some prime sets and their corresponding target response

As we have considered exact string match(ignoring capitalization), the evaluation does not account for spelling variations. For example, the response output *armor* instead of the expected response *armour* results in counting it as incorrect.

We achieved an accuracy of 34.9% by considering the top response for each list of ranked responses. However, it was observed that the correct response was present within the top ten responses in 59.8% of the cases. For example, the primes *ants*, *flies*, *fly*, *bees*, *bite* generate the response output *mosquitoes*. The expected output *insects* ranks 4^{th} in our list of responses.

For primes *babies*, *baby*, *rash*, *wet*, *washing*, our system outputs *nappy* while the expected response is *nappies*. Such inflected forms of the responses are challenging to predict and hence, another evaluation is presented which ignores the inflectional variation of the response word. Under this evaluation, we achieved an accuracy of 39.55% for the best response and 63.15% if the expected response occurs in the top ten responses. Table - 5 displays accuracy of our system when the target response lies within the top-n responses for both evaluation methods.

	Exact Match	**Ignoring Inflections**
n=**1**	34.9	39.55
n=**3**	48.15	49.65
n=**5**	53.2	55.45
n=**10**	59.8	63.15

Table 5: Evaluation results in %

6 Conclusion

There exist some word associations that are *asymmetric* in nature. Rapp (2013) observed that the primary response of a given stimulus word may have stronger association with another word and need not generate the stimulus word back. For example, the strongest association to *bitter* is *sweet* but the strongest association to *sweet* is *sour*. Therefore, the EAT data set chosen for evaluation, may not be the best judge for certain cases. Taking a case from our test data, for primes *butterfly*, *moth*, *caterpillar*, *cocoon*, *insect*, our system outputs *larva* instead of the original stimulus word *chrysalis* which does not feature even in the top ten responses (Refer Table - 4).

In this work, we proposed a system to generate a ranked list of responses for multiple stimulus words. Candidate responses were generated by computing its semantic similarity with the stimulus words and then re-ranked using a lexical association measure, PMI. This system scored 34.9% when the top ranked response was considered and 59.8% when the top ten responses were taken into account. When ignoring inflectional variations, the accuracy improved to 39.55% and 63.15% for the two evaluation methods respectively.

In future, a more sophisticated re-ranking approach in place of PMI measure can be used such as product-of-rank algorithm (Rapp, 2008). Since, the re-ranking methodologies discussed by far, take into account word co-occurrences, it is biased towards syntagmatic responses. A better trade-off can be worked out to give due weightage to paradigmatic responses too.

References

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3):209–226.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Peter F. Brown, Peter V. Desouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. 1992. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Comput. Linguist.*, 16(1):22–29, March.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM.

Ferdinand De Saussure, Charles Bally, Albert Sechehaye, and Albert Riedlinger. 1916. *Cours de linguistique générale: Publié par Charles Bally et Albert Sechehaye avec la collaboration de Albert Riedlinger*. Libraire Payot & Cie.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE*, 41(6):391–407.

Thomas Hofmann. 1999. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 289–296. Morgan Kaufmann Publishers Inc.

George R. Kiss, Christine A. Armstrong, and Robert Milroy. 1972. *An associative thesaurus of English*. Medical Research Council, Speech and Communication Unit, University of Edinburgh, Scotland.

Thomas K. Landauer and Susan T. Dutnais. 1997. A solution to platos problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, pages 211–240.

Michael McCarthy. 1990. *Vocabulary*. Oxford University Press.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013b. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013c. Linguistic regularities in continuous space word representations. *Proceedings of NAACL-HLT*, pages 746–751.

Andriy Mnih and Geoffrey E. Hinton. 2008. A scalable hierarchical distributed language model. In *NIPS*, pages 1081–1088.

Reinhard Rapp. 2008. The computation of associative responses to multiword stimuli. In *Proceedings of the workshop on Cognitive Aspects of the Lexicon*, pages 102–109. Association for Computational Linguistics.

Reinhard Rapp. 2013. From stimulus to associations and back. *Natural Language Processing and Cognitive Science*, page 78.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 384–394. Association for Computational Linguistics.

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

Manfred Wettler and Reinhard Rapp. 1993. Computation of word associations based on the co-occurrences of words in large corpora.