HAL, COALS and GloVe - Building Global Semantic Word Vector Models

Ramaseshan Ramachandran

SUMMARY OF WORD2VEC

- Uses Skip-gram and CBOW Models to build distributed word vectors
- Optimizes computations using sub-sampling, negative sampling and Hierarchical Softmax at the output layer
- Uses co-occurrence words, but ignores the frequency of co-occurrence words
- Context is within the chosen window size (say 5 or 7)

SEMANTIC UNDERSTANDING USING CO-OCCURRENCE - EXAMPLE

The view from the top of the mountain was

- awesome
- breathtaking
- amazing
- stunning
- astounding
- astonishing
- awe-inspiring
- extraordinary
- incredible
- unbelievable
- magnificent
- wonderful
- spectacular
- remarkable

LEXICAL SEMANTIC MODEL

- LSA - Latent Semantic Analysis
- HAL - Hyperspace Analogue to Language[1]
- COALS - Correlated Occurrence Analogue to Lexical Semantic[2]
- GloVe - Global Vectors[3]

[1]Lund, K. & Burgess, C. Producing high-dimensional semantic spaces from lexical co-occurrence Behavior Research Methods, Instruments, & Computers, 1996, 28, 203-208

[2]Rhode et al, "An Improved Model of Semantic Similarity Based on Lexical Co-Occurrence", CACM, 2006, 8, 627-633

[3]Pennington et al, "Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, 1532-1543

CONSTRUCTING SEMANTIC MODELS

- Semantic spaces are constructed by selecting an axis
- Use human judgment to place words in each axis
- To place a set of desirable words, one must choose the axis and find a set of words that must be confined to the chosen axis
- In a *size* axis, placing *ant* and *mountain*
- Can we use lexical co-occurrence to construct semantic spaces?
- Is it possible to construct high-dimensional distributed semantic spaces automatically?

HYPERSPACE ANALOGUE TO LANGUAGE - HAL

- A Window size n representing a span of words is used
- Words within the window (or ramped window), are recorded
- The strength of co-occurrence is computed using a inverse relationship with respect to the word in question, w_i

  $C_s \propto \frac{1}{N}$, where C_s is the co-occurrence strength and N is the number of words separating them
- A Word w_j^1 immediately occurring next to w_i will have a higher value than the word, w_j^n separated by a distance of n from it.
- The co-occurring word strengths are distance and direction sensitive
- A term-term matrix is constructed with every cell representing summed co-occurrence counts for a single word pair
- If the words have similar values in the same dimensions, they will be closer together in the space,meaning they share similar contexts
- The word vectors closest to a given word are considered its neighbors.

EXAMPLE

Example Matrix for the sentence *The Horse Raced Past the Barn Fell.*[4]
(Computed for Window Width of Five Words)

	barn	fell	horse	past	raced	the
<period>	4	5	0	2	1	3
barn	0	0	2	4	3	6
fell	5	0	1	3	2	4
horse	0	0	0	0	0	5
past	0	0	4	0	5	3
raced	0	0	5	0	0	4
the	0	0	3	5	4	2

Rows - Count from right to left

Columns - Count from Left to right

- Pick up a text that contains conversational text, variety of topics to cover all type of co-occurrences

[4] All tables and figures mentioned in the presentation were taken from the respective papers as mentioned earlier

EXPERIMENT 1

- 160 million words from Usenet news groups
- Window size = 10
- A word appearing with a frequency of 50 or more is considered as a vocabulary item
- 20 target words selected at random from middle frequency words (Using Zipf's law) - to eliminate most common and rare words
- For each target word, a normalized Euclidean distance was computed from the word to each vocabulary item
- The neighbors with the smallest distances is shown in Table **??**
- These relationships appear to be both semantic and associative
- The high-dimensional neighborhood surrounding each word is similar to a semantic field

EXAMPLE FROM LUND AND BURGESS (1996)

Table 2
Five Nearest Neighbors for Target Words From Experiment 1 (*n*1 ... *n*5)

Target	*n*1	*n*2	*n*3	*n*4	*n*5
jugs	juice	butter	vinegar	bottles	cans
leningrad	rome	iran	dresden	azerbaijan	tibet
lipstick	lace	pink	cream	purple	soft
triumph	beauty	prime	grand	former	rolling
cardboard	plastic	rubber	glass	thin	tiny
monopoly	threat	huge	moral	gun	large

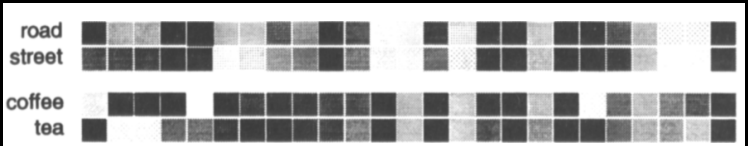
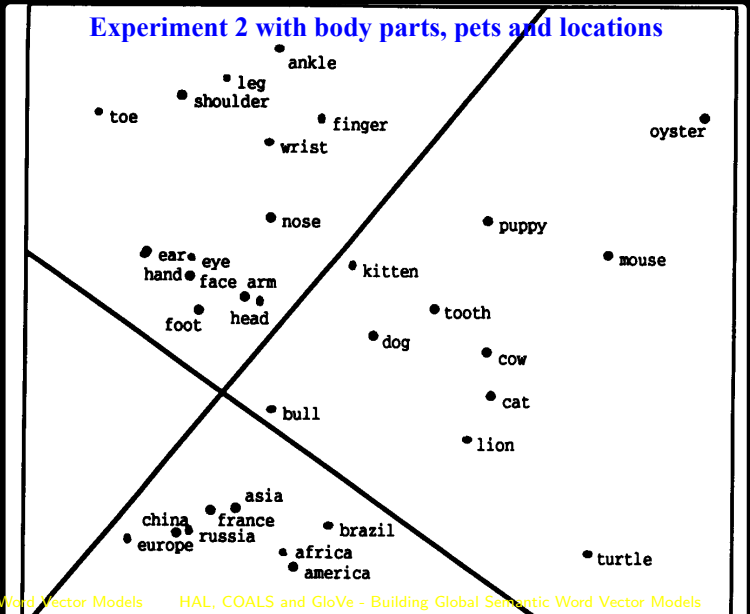


Figure 1. Gray-scaled 25-element co-occurrence vectors.

EXAMPLE FROM LUND BURGESS (1996)



SUMMARY

- HAL captures information about word meanings through the unsupervised analysis of text
- This produces word vectors that are more semantic (similar words) than associative in nature
- HAL acquires word meanings as a function of keeping track of how words are used in context
- The term-term co-occurrence matrix carries the history of the contextual experience by using a moving window and weighting of co-occurring words based on the distance
- HAL exploits the regularities of language such that conceptual generalizations can be captured in a data matrix

CORRELATED OCCURRENCE ANALOGUE TO LEXICAL SEMANTIC - COALS

1. Gather co-occurrence counts, typically ignoring closed-class neighbors and using a ramped, size 4 window
2. Discard all but the m (14,000, in this case) columns reflecting the most common open-class words.
3. Convert counts to word pair correlations - Instead of using the raw frequency score, correlation score is used to analyze the relationship between pair of words
4. Set negative values to 0, and take square roots of positive ones.
5. The semantic similarity between two words is given by the correlation of their vectors. The correlation coefficient values with this normalization will be in the range of [-1,1]
6. The matrix constructed using this correlation would be semantic space

In HAL, high frequency neighbors have undue influence on the scores. COALS method employs a normalization strategy that largely factors out lexical frequency.
Columns representing low-frequency words are removed

COALS-STEP 1

Consider the corpus

How much wood would a woodchuck chuck,
if a woodchuck could chuck wood?
As much wood as a woodchuck would,
if a woodchuck could chuck wood.

Table 5
Step 1 of the COALS method: The initial co-occurrence table with a ramped, 4-word window.

	a	as	chuck	could	how	if	much	wood	woodch.	would	,	.	?
a	0	5	9	6	1	10	4	8	18	9	10	0	0
as	5	4	2	1	0	0	7	10	3	2	1	0	5
chuck	9	2	0	8	0	5	1	9	11	2	4	3	3
could	6	1	8	0	0	4	0	6	8	0	2	2	2
how	1	0	0	0	0	0	4	3	0	2	0	0	0
if	10	0	5	4	0	0	0	0	10	3	8	0	0
much	4	7	1	0	4	0	0	10	2	3	0	0	3
wood	8	10	9	6	3	0	10	2	8	5	0	4	6
woodch.	18	3	11	8	0	10	2	8	0	8	10	1	1
would	9	2	2	0	2	3	3	5	8	0	5	0	0
,	10	1	4	2	0	8	0	0	10	5	0	0	0
.	0	0	3	2	0	0	0	4	1	0	0	0	0
?	0	5	3	2	0	0	3	6	1	0	0	0	0

COALS-STEP 2

Table 6
Step 2 of the COALS method: Raw counts are converted to correlations.

	a	as	chuck	could	how	if	much	wood	woodch.	would	,	.	?
a	-0.167	-0.014	0.014	0.009	-0.017	0.085	-0.018	-0.033	0.096	0.069	0.085	-0.055	-0.079
as	-0.014	0.031	-0.048	-0.049	-0.037	-0.077	0.133	0.103	-0.054	-0.021	-0.050	-0.037	0.133
chuck	0.014	-0.048	-0.113	0.094	-0.045	0.021	-0.061	0.031	0.048	-0.046	-0.002	0.088	0.031
could	0.009	-0.049	0.094	-0.075	-0.037	0.033	-0.070	0.022	0.049	-0.075	-0.021	0.069	0.023
how	-0.017	-0.037	-0.045	-0.037	-0.018	-0.037	0.192	0.070	-0.055	0.069	-0.037	-0.018	-0.026
if	0.085	-0.077	0.021	0.033	-0.037	-0.077	-0.071	-0.106	0.085	0.006	0.138	-0.037	-0.053
much	-0.018	0.133	-0.061	-0.070	0.192	-0.071	-0.065	0.128	-0.061	0.019	-0.071	-0.034	0.072
wood	-0.033	0.103	0.031	0.022	0.070	-0.106	0.128	-0.113	-0.033	0.001	-0.106	0.111	0.100
woodch.	0.096	-0.054	0.048	0.049	-0.055	0.085	-0.061	-0.033	-0.167	0.049	0.085	-0.017	-0.051
would	0.069	-0.021	-0.046	-0.075	0.069	0.006	0.019	0.001	0.049	-0.075	0.060	-0.037	-0.053
,	0.085	-0.050	-0.002	-0.021	-0.037	0.138	-0.071	-0.106	0.085	0.060	-0.077	-0.037	-0.053
.	-0.055	-0.037	0.088	0.069	-0.018	-0.037	-0.034	0.111	-0.017	-0.037	-0.037	-0.018	-0.026
?	-0.079	0.133	0.031	0.023	-0.026	-0.053	0.072	0.100	-0.051	-0.053	-0.053	-0.026	-0.037

Pearson's correlation coefficient

$$r = \frac{T w_{a,b} - \sum_j w_{a,j} \sum_i w_{b,i}}{\sqrt{\sum_j w_{a,j} (T - \sum_j w_{a,j}) \sum_i w_{b,i} (T - \sum_i w_{b,i})}} \qquad (1)$$

where,

$$T = \sum_i \sum_j w_{i,j}$$

COALS-STEP 3

Table 7
Step 3 of the COALS method: Negative values discarded and the positive values square rooted.

	a	**as**	**chuck**	**could**	**how**	**if**	**much**	**wood**	**woodch.**	**would**	**,**	**.**	**?**
a	0	0	0.120	0.093	0	0.291	0	0	0.310	0.262	0.291	0	0
as	0	0.175	0	0	0	0	0.364	0.320	0	0	0	0	0.365
chuck	0.120	0	0	0.306	0	0.146	0	0.177	0.220	0	0	0.297	0.175
could	0.093	0	0.306	0	0	0.182	0	0.149	0.221	0	0	0.263	0.151
how	0	0	0	0	0	0	0.438	0.265	0	0.263	0	0	0
if	0.291	0	0.146	0.182	0	0	0	0	0.291	0.076	0.372	0	0
much	0	0.364	0	0	0.438	0	0	0.358	0	0.136	0	0	0.268
wood	0	0.320	0.177	0.149	0.265	0	0.358	0	0	0.034	0	0.333	0.317
woodch.	0.310	0	0.220	0.221	0	0.291	0	0	0	0.221	0.291	0	0
would	0.262	0	0	0	0.263	0.076	0.136	0.034	0.221	0	0.246	0	0
,	0.291	0	0	0	0	0.372	0	0	0.291	0.246	0	0	0
.	0	0	0.297	0.263	0	0	0	0.333	0	0	0	0	0
?	0	0.365	0.175	0.151	0	0	0.268	0.317	0	0	0	0	0

COALS - RESULTS

Table 10

The 10 nearest neighbors and their percent correlation similarities for a set of nouns, under the COALS-14K model.

	gun	point	mind	monopoly	cardboard	lipstick	leningrad	feet
1)	46.4 handgun	32.4 points	33.5 minds	39.9 monopolies	47.4 plastic	42.9 shimmery	24.0 moscow	59.5 inches
2)	41.1 f rearms	29.2 argument	24.9 consciousness	27.8 monopolistic	37.2 foam	40.8 eyeliner	22.7 sevastopol	57.7 foot
3)	41.0 f rearm	25.4 question	23.2 thoughts	26.5 corporations	36.7 plywood	38.8 clinique	22.7 petersburg	52.0 metres
4)	35.3 handguns	22.3 arguments	22.4 senses	25.0 government	35.6 paper	38.4 mascara	20.7 novosibirsk	45.7 legs
5)	35.0 guns	21.5 idea	22.2 subconscious	23.2 ownership	34.8 corrugated	37.2 revlon	20.3 russia	45.4 centimeters
6)	32.7 pistol	20.1 assertion	20.8 thinking	22.2 property	32.3 boxes	35.4 lipsticks	19.6 oblast	44.4 meters
7)	26.3 weapon	19.5 premise	20.6 perception	22.2 capitalism	31.3 wooden	35.3 gloss	19.5 minsk	40.2 inch
8)	24.4 rif es	19.3 moot	20.4 emotions	21.8 capitalist	31.0 glass	34.1 shimmer	19.2 stalingrad	38.4 shoulders
9)	24.2 shotgun	18.9 distinction	20.1 brain	21.6 authority	30.7 fabric	33.6 blush	19.1 ussr	37.8 knees
10)	23.6 weapons	18.7 statement	19.9 psyche	21.3 subsidies	30.5 aluminum	33.5 nars	19.0 soviet	36.9 toes

Table 11

The 10 nearest neighbors for a set of verbs, according to the COALS-14K model.

	need	buy	play	change	send	understand	explain	create
1)	50.4 want	53.5 buying	63.5 playing	56.9 changing	55.0 sending	56.3 comprehend	53.0 understand	58.2 creating
2)	50.2 needed	52.5 sell	55.5 played	55.3 changes	42.0 email	53.0 explain	46.3 describe	50.6 creates
3)	42.1 needing	49.1 bought	47.6 plays	48.9 changed	40.2 e-mail	49.5 understood	40.0 explaining	45.1 develop
4)	41.2 needs	41.8 purchase	37.2 players	32.2 adjust	39.8 unsubscribe	44.8 realize	39.8 comprehend	43.3 created
5)	41.1 can	40.3 purchased	35.4 player	30.2 affect	37.3 mail	40.9 grasp	39.7 explained	42.6 generate
6)	39.5 able	39.7 selling	33.8 game	29.5 modify	35.7 please	39.1 know	39.0 prove	37.8 build
7)	36.3 try	38.2 sells	32.3 games	28.3 different	33.3 subscribe	38.8 believe	38.2 clarify	36.4 maintain
8)	35.4 should	36.3 buys	29.0 listen	27.1 alter	33.1 receive	38.5 recognize	37.1 argue	36.4 produce
9)	35.3 do	34.0 sale	26.8 playable	25.6 shift	32.7 submit	38.0 misunderstand	37.0 refute	35.4 integrate
10)	34.7 necessary	31.5 cheap	25.0 beat	25.1 altering	31.5 address	37.9 understands	35.9 tell	35.2 implement

Table 12

The 10 nearest neighbors for a set of adjectives, according to the COALS-14K model.

	high	frightened	red	correct	similar	fast	evil	christian
1)	57.5 low	45.6 scared	53.7 blue	59.0 incorrect	44.9 similiar	43.1 faster	24.3 sinful	48.5 catholic
2)	51.9 higher	37.2 terrif ed	47.8 yellow	37.7 accurate	43.2 different	41.2 slow	23.4 wicked	48.1 protestant
3)	43.4 lower	33.7 confused	45.1 purple	37.5 proper	40.8 same	37.8 slower	23.2 vile	47.9 christians
4)	43.2 highest	33.3 frustrated	44.9 green	36.3 wrong	40.6 such	28.2 rapidly	22.5 demons	47.2 orthodox
5)	35.9 lowest	32.6 worried	43.2 white	34.1 precise	37.7 specif c	27.3 quicker	22.3 satan	47.1 religious
6)	31.5 increases	32.4 embarrassed	42.8 black	32.9 exact	35.6 identical	26.8 quick	22.3 god	46.4 christianity
7)	30.7 increase	32.3 angry	36.8 colored	30.7 erroneous	34.6 these	25.9 speeds	22.3 sinister	43.8 fundamentalist
8)	29.2 increasing	31.6 afraid	35.6 orange	30.6 valid	34.4 unusual	25.8 quickly	22.0 immoral	43.5 jewish
9)	28.7 increased	30.4 upset	33.5 grey	30.6 inaccurate	34.1 certain	25.5 speed	21.5 hateful	43.2 evangelical
10)	26.5 lowering	30.3 annoyed	32.4 reddish	29.8 acceptable	32.7 various	24.3 easy	21.3 sadistic	41.2 mormon

COALS - SOME INSIGHTS

- The majority of the correlations are negative [5]
- Word with negative correlations do not contribute well to finding similarity than the ones with positive correlation
- Closed-class words (147) convey syntactic information than semantic - could be removed from the correlation table

  punctuation marks, she, he, where, after, ...

[5] Based on a Usenet corpus of size 1.2 billion words

MAJOR DIFFERENCES - SUPERVISED VS UNSUPERVISED

Skip-gram model - Supervised training model, scales with corpus size, does not use the count of co-occurrences, good model for similarity estimation, frequent phrase identification, analogy

COALS, HAL, LSI - unsupervised models, uses the co-occurrence statistics, frequently co-occurring words have undue advantage, captures only word similarity

THE GLOVE MODEL

- The Global Vector[6] (GloVe) models the word vectors with the computed statistics of the co-occurrence count
- The authors of this model introduce the idea that the co-occurrence ratio between two words in a context are strongly connected to the meaning

Let X represent the counts of co-occurrence matrix. Every element of X_{ij} represent the number of times the word j occurs in the context of word i.
Let $X_i = \sum_k X_{ik}$ be the number of times any word appears in the context of word i

Let $P_{ij} = P(j|i) = \frac{X_{ij}}{X_i}$ be the probability that word j appear in the context of word i
The skip-gram model captures the co-occurrences patterns one window at a time while the Glove captures it using the statistics of the co-occurrences or how often the patterns occur together

[6] J.Pennington, R.Socher, C.D Manning, "GloVe: Global Vectors for Word Representation," Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1532-1543, October 25-29, 2014

GLOVE - EXAMPLE

Probability and Ratio	k = solid	k = gas	k = water	k = fashion
P(k|ice)	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
P(k|steam)	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}
P(k|ice)/P(k|steam)	8.9	8.5×10^{-2}	1.36	0.96

Figure: Co-occurrence probabilities for target words ice and steam with selected context words from a 6 billion token corpus

GLOVE

Since the ratio $\frac{P_{ik}}{P_{jk}}$ depends i, j, k, it can be modeled by $F(w_i, w_j, \tilde{w}_k)$. There could be several possible ways to encode the ratio. We would like to estimate the parameters of this model given the ratio

Using the factoring approach similar to LSA, the new weighted least square regression model is proposed that minimizes the cost function

$$J(\theta) = \sum_{i,j=1}^{|V|} f(X_{ij})(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij})^2 \quad (2)$$

where $|V|$is the size of the vocabulry and (3)

$$f(x) = \begin{cases} \left(\frac{x}{x_{max}}\right)^{\alpha} & \text{if } x < x_{max} \\ 1, & otherwise \end{cases} \quad (4)$$

The cutoff $x_{max} = 100$ and $\alpha = 0.75$ (5)

GLOVE - TRAINING

- Corpus size
  - 2010 Wikipedia dump - 1 billion tokens
  - 2014 Wikipedia dump - 1.6 billion tokens
  - 2014 Wikipedia dump + Gigaword5+ Common crawl of web - 42 billion tokens
- Input matrix size $X \in R^{|V| \times |V|}$
- Vocabulary - 400K frequent words
- Initial learning rate - 0.05
- Context words to the left $= 10$
- Context words to the right $= 10$
- Generates two words vectors W and $\tilde{W}$
- The final word vector $= W + \tilde{W}$

GLOVE - RESULTS

$SVDL = \log(1 + X_{ij})$
$SVDS = \sqrt{X_{ij}}$
HPCA: PMI version of LSA (PCA)
vLBL, ivLBL: log-bilinear model

Model	Dim.	Size	Sem.	Syn.	Tot.
ivLBL	100	1.5B	55.9	50.1	53.2
HPCA	100	1.6B	4.2	16.4	10.8
GloVe	100	1.6B	67.5	54.3	60.3
SG	300	1B	61	61	61
CBOW	300	1.6B	16.1	52.6	36.1
vLBL	300	1.5B	54.2	64.8	60.0
ivLBL	300	1.5B	65.2	63.0	64.0
GloVe	300	1.6B	80.8	61.5	70.3
SVD	300	6B	6.3	8.1	7.3
SVD-S	300	6B	36.7	46.6	42.1
SVD-L	300	6B	56.6	63.0	60.1
CBOW†	300	6B	63.6	67.4	65.7
SG†	300	6B	73.0	66.0	69.1
GloVe	300	6B	77.4	67.0	71.7
CBOW	1000	6B	57.3	68.9	63.7
SG	1000	6B	66.1	65.1	65.6
SVD-L	300	42B	38.4	58.2	49.2
GloVe	300	42B	81.9	69.3	75.0

EVALUATION OF WORD EMBEDDINGS

There are two type of evaluations

1. Intrinsic Evaluation - word embeddings are compared with human judgments on words relations[7]
2. Extrinsic Evaluation - traditionally judged by its utility in downstream NLP tasks. The performance of the word embedding is measured indirectly by the performance of these downstream applications

[7] Tsvetkov, Y., Faruqui, M., Ling, W., Lample, G. & Dyer, C., "Evaluation of Word Vector Representations by Subspace Alignmen," Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2015, 2049-2054

EVALUATION OF WORD EMBEDDINGS - EXTRINSIC EVALUATION

- Sentiment Analysis
- Named Entity Recognition
- Machine Translation
- Language Model
- Conversation Modeling
- Text Classification

EVALUATION OF WORD EMBEDDINGS - INTRINSIC EVALUATION

- Evaluate word vector representation quality by judging the similarity of representations assigned to similar words by humans.
- The most popular evaluation sets at present consist of word pairs with similarity ratings produced by human annotators

  Use a correlation method to compare word vectors and linguistic vectors using common words
- If the correlation score is higher, then the word vector quality is good

LINGUISTIC SCORES BY HUMAN ANNOTATORS

word1	word2	score	r1	r2	r3	r4	r5	r6	r7	r8	r9	r10
choose	pick	9.63	6	6	6	5	6	6	5	6	6	6
	predict	4.32	4	0	0	4	2	3	6	2	0	5
	want	2.99	1	4	1	3	4	0	0	2	3	0
	elect	8.475	5	5	5	4	5	6	4	5	6	6
	determine	7.47	1	6	5	2	6	4	6	5	5	5

The score $\{0,6\}$ is mapped to $\{0,10\}$

EVALUATION

- Let N be the number of common words in the word embedding.
- Let $X \in R^{D \times N}$ be the word vector matrix and let $x_j \in R^{1 \times N}$ be the word vector. D denotes the word vector dimension
- Let $S \in R^{P \times N}$ be the linguistic property matrix. Let $s_j \in R^{1 \times N}$ be the linguistic property vector for a word. P denotes linguistic properties obtained from a manually annotated linguistic resource.

Let $A \in \{0,1\}^{D \times P}$ be a matrix of alignments such that $a_{ij} = 1$ iff x_i is aligned to s_j, otherwise $a_{ij} = 0$. If $r(x_i, s_j)$ is the Pearson's correlation between vectors x_i and s_j, then our quality of word vector is defined as:

$$Q = \max_{A | \sum_j a_{ij} \leq 1} \sum_{l=1}^{D} \sum_{j=1}^{P} r(x_i, s_j) \times a_{ij}$$