Word Relatedness using Inter-context Similarity

No Author Given

No Institute Given

Abstract. Computing relatedness between terms in a given corpus is a key component of many tasks in data mining problems. Methods to automatically compute relatedness make use of the premise that co-occurrence in the same context and frequent appearances in multiple similar contexts is an indicator of similarity. We propose Cross-Context Similarity (CCS) that utilizes similarity among contexts (cross-occurrence) along with co-occurrence, and present a mathematical study of the relationship between corpus parameters and the computed similarity. Specifically, for a case where the dataset is a parallel text, we prove that our method favors large corpus size, large vocabulary, and small context size, for distinguishing related from unrelated words.

Keywords: text similarity, cross-lingual information retrieval, dictionary construction, evolving bipartite graphs

1 Introduction

A measure of relatedness between concepts is a key component of many tasks such as Information Search and Retrieval, Classification and Clustering. These measures are typically based on a notion of 'context', which may vary depending on the application. In particular, sentence context is central to tasks such as word sense disambiguation [13,7]. Several approaches to compute similarity using only the word distribution in a context have been proposed [12,8]. Explicit Semantic Analysis [4] and Temporal Semantic Analysis [14], that use Wikipedia and New York Times corpus respectively, have been shown to correlate well with human judgment. "Cross-domain similarity mining" has also been explored where the problem of schema and ontology matching [2], including cross-lingual ontology matching [15] is studied. An important application of cross-lingual ontology matching is cross-lingual information retrieval (CL-IR), which relies on an implicit or explicit relatedness among the words of different languages. This makes it useful in building a dictionary of multi-lingual word association [16,11].

In addition to co-occurrence in a context, similarity between contexts themselves can be incorporated into computing word similarity. For instance, Kandola et. al. [9] defined word similarity in terms of document similarity and vice-versa in an unsupervised way. The idea is in accordance with the Distributional Hypothesis which states that words that occur in the same/similar contexts tend to have similar meanings [6]. A central question in the use of context similarity to derive word similarity is the impact of corpus parameters (such as its size, and

number of entries) on the quality of derived word similarities. All the above-cited work only provide empirical evaluations on specific methods of computing word similarity.

We have developed *cross-context similarity* (CCS) to relate both co-occurrence and similarity between contexts to similarity between the individual terms [1] that comprise the context. CCS is a corpus-based measure, which is particularly useful for applications where related entities are to be computed from a domain-specific corpus with narrower scope, one that could express similarities differently from the methods that aim to find generic relatedness using large number of Wikipedia or New York Times articles. For instance, while *(hockey, soccer)* can be considered similar in typical situations, the pair should not be considered similar when the focus is on different sports. In such applications, a corpus-specific similarity measure would be more desirable than a method based on external knowledge. Our model is based on computing *cross-occurrence* which is defined as a measure of the weighted co-occurrence between terms in similar contexts. We present a detailed analysis when the corpus is restricted to an explicit one-to-one mapping of contexts and a latent one-to-one mapping of terms (referred to as *Substitution Translation Model*). We show that the cross-context similarity between two terms that are not a translation of each other is small. We quantify the effect of context size and vocabulary size on cross-context similarity of unrelated words. We also show that the maximum similarity value is reached only when the pair of terms perfectly co-occur in every context. We have also found that CCS performs well on multi-lingual word association task.

2 Cross Context Similarity

To find similarity between two terms, we make the assumption that a pair of words occurring in many pairs of similar contexts are similar to each other. Here a context could mean a phrase, a sentence or a document. The intuitive idea for calculating cross-context similarity of words w_i and w_j is to examine the similarity of all those pairs of contexts (D_m, D_n), such that $w_i \in D_m$ and $w_j \in D_n$. Next, we formally define cross-context similarity.

2.1 Similarity Formula

Consider a corpus as a set of contexts $\mathbf{D} = \{D_1, D_2, \ldots, D_{|\mathbf{D}|}\}$, where every context is a set of words. Let $\mathbf{W}$ be the set of all words covered by $\mathbf{D}$, i.e., $\mathbf{W} = \bigcup_m D_m$. Assume that the similarity of every pair of contexts is known. Let $e_{m,n} \geq 0$ denote the similarity between D_m and D_n. We represent the corpus as a two-layer network (Figure 1). The bottom layer is a directed graph representation of the contexts, where each node is a context and a weighted directed edge between D_m and D_n with weight $e_{m,n}$ represents context similarity. Similarly, the top layer is a directed graph, with words as nodes and directed edges among

[1] *term* and *word* has been used interchangeably in this paper.

pairs of words with unknown weights. There is an undirected link with a weight $t_{i,m}$ between D_m in the bottom level and w_i in the top level, representing the membership of w_i in D_m. The values of $t_{i,m}$ could be frequency counts, tf-idf values or simply binary values representing presence/absence. We wish to assign weights to the edges between word pairs representing a relatedness score. For this, first we define *Cross-Occurrence (CO_λ)* as a score between two words w_i and w_j that measures a weighted co-occurrence of these words in similar contexts -

$$CO_\lambda(w_i, w_j) = \sum_m \sum_{n \neq m} t_{i,m} e_{m,n} t_{j,n} + \lambda M \sum_m t_{i,m} t_{j,n} \tag{1}$$

where $t_{i,m} \geq 0$ is the weight associated with w_i in D_m, and M is the maximum context similarity that can be attained. $\lambda \in [0, 1]$ is a parameter that represents the weight we wish to assign to the similarity arising from co-occurrence in same context. Based on the *Cross-Occurrence* scores, we define *cross-context similarity (CCS_λ)* between two words w_i and w_j as

$$CCS_\lambda(w_i, w_j) = \frac{CO_\lambda(w_i, w_j)}{\max\{CO_1(w_i, w_i), CO_1(w_j, w_j)\}}. \tag{2}$$

This can be rewritten in the form of matrix operations as

$$\begin{aligned} \mathbf{C}_\lambda &= (D^T E D) + \lambda M (D^T D), \\ \text{And, } \mathbf{S} &= \mathbf{C}_\lambda \oslash (F_{max}), \end{aligned} \tag{3}$$

where $[\mathbf{C}]_{i,j} = CO(w_i, w_j)$, $[\mathbf{S}]_{i,j} = CCS(w_i, w_j)$, $[D]_{i,j} = t_{j,i}$, $[F_{max}]_{i,j} = \max\{[\mathbf{C_1}]_{i,i}, [\mathbf{C_1}]_{j,j}\}$, and '$\oslash$' represents matrix element-wise division. E is the weighted adjacency matrix, such that $[E]_{m,n} = e_{m,n}$ for $m \neq n$ and 0 otherwise. If E follows certain properties, then we can show that CCS is bounded.

Theorem 1. *Let* $\mathbf{I}$ *be the unit matrix of size* $|\mathbf{D}| \times |\mathbf{D}|$. *If* $E + M\mathbf{I}$ *is positive semi-definite, then* $0 \leq CCS_\lambda(w_i, w_j) \leq 1 \forall w_i, w_j$.

3 Translation Model for Multi-lingual Word Association Mining

In the previous section, we modeled a single set of contexts $\mathbf{D}$ and a set of words W, and we used CCS to find the similarities between all pairs (w_i, w_j). By refining this formulation, it is possible to represent more complex relationships in a corpus. For instance, consider a corpus of scientific articles that contains some scientific terms and some English words. High similarity between a scientific term w_i and an English word w_j would represent how well w_j describes w_i. Here, the top level is a bipartite graph, i.e, $W = W^1 \cup W^2, W^1 \cap W^2 = \emptyset$. There are two types of words present in the corpus and we compute $CCS(w_i, w_j)$ such that $w_i \in W^1$ and $w_j \in W^2$ (Figure 2(a)). As a more complex example, suppose in our corpus we have translations of sentences from one language to the other,

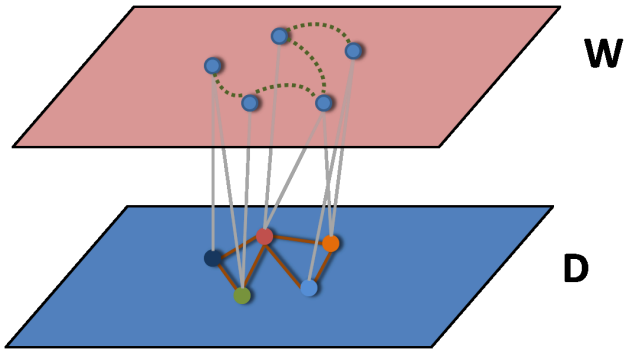
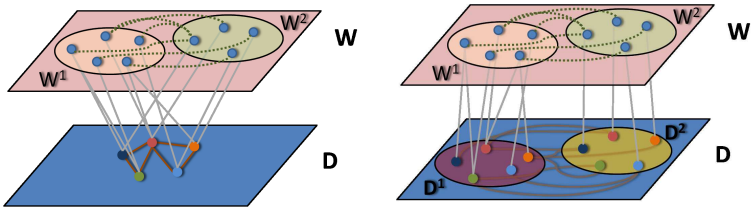


Fig. 1. Representing a corpus as a two-layer network.



(a) Top level bipartite (b) Both levels bipartite

Fig. 2. Capturing more refined relationships with cross-context similarity.

with scores representing how good a pair of translation is. We wish to construct a word-level dictionary between these two languages. Here, both the layers are bipartite graphs, i.e., $W = W^1 \cup W^2$ and $\mathbf{D} = \mathbf{D^1} \cup \mathbf{D^2}, \mathbf{D^1} \cap \mathbf{D^2} = \emptyset$. There are directed edges only from contexts in $\mathbf{D^1}$ to contexts in $\mathbf{D^2}$ (Fig 2(b)). Again, we compute $CCS(w_i, w_j)$, such that $w_i \in W^1$ and $w_j \in W^2$. Next we analyze the behavior of CCS for dictionary construction.

Consider two hypothetical languages $L1$ and $L2$. Assume that we have parallel texts (sentences) of these two languages. There are two types of contexts, one type for each of the two languages. The set of words can also be partitioned into two sets - words belonging to $L1$ and $L2$ respectively. Each context of one language can be mapped to one context in the other language, which is its translation. Our objective is to learn multi-lingual word association dictionary, i.e., the "translation" of individual words. A score associated with pair of words across the two languages is to be mined, based on sentence-level translations $\mathbf{D^1}$ and $\mathbf{D^2}$ consists of the contexts in $L1$ and $L2$ respectively. Similarly, W^1 and W^2 consist of words from $L1$ and $L2$ respectively. There is an edge of unit weight between contexts D_m and D_n if they are translations of each other. Therefore, $e_{m,n} \in \{0, 1\}$. Also, the maximum context similarity that can be attained is 1, i.e., $M = 1$. The score $t_{i,m}$ of a word w_i in a context D_m is the indicator function

$I_{i,m} \in \{0,1\}$, representing the presence of the word in the context. Then, for $w_i \in W^1$ and $w_j \in W^2$,

$$CO_\lambda(w_i, w_j) = \sum_m \sum_{n \neq m} I_{i,m} e_{m,n} I_{j,n} + \lambda \sum_m I_{i,m} I_{j,m}. \tag{4}$$

Since, both $w_i \in W^1$ and $w_j \in W^2$ cannot occur in the same context (because they are from different languages), $I_{i,m} I_{j,m} = 0, \forall m$. So, Equation 4 becomes

$$CO_\lambda(w_i, w_j) = \sum_m \sum_{n \neq m} I_{i,m} e_{m,n} I_{j,n} \tag{5}$$

And,

$$CO_1(w_i, w_i) = \sum_m \sum_{n \neq m} I_{i,m} e_{m,n} I_{i,n} + \sum_m I_{i,m} I_{i,m}.$$

Since, w_i cannot occur in both D_m and its translation D_n, $I_{i,m} e_{m,n} I_{j,n} = 0, \forall m \neq n$, and so,

$$\begin{aligned} CO_1(w_i, w_i) &= \sum_m I_{i,m} I_{i,m} \\ &= \sum_m I_{i,m} = f_i \,, \end{aligned} \tag{6}$$

where f_i is the number of contexts containing w_i. We refer to this model as a *Translation Model*, when $t_{i,m} = I_{i,m}$, and $e_{m,n} = 1$ only if $m = n$ or D_m is the correct translation of D_n. Thus, for *Translation Model*,

$$CCS(w_i, w_j) = \frac{\sum_m \sum_n I_{i,m} e_{m,n} I_{j,n}}{\max\{f_i, f_j\}}. \tag{7}$$

Note that we have dropped the subscript λ because CCS in Translation Model becomes independent of λ (Equation 5).

Theorem 2. *In Translation Model,* $0 \leq CCS(w_i, w_j) \leq 1, \forall i, j$

It can be shown that CCS correctly assigns a value of 1 to true word translations and low values to other pairs for a refinement of the Translation Model.

3.1 Substitution Translation Model

We proceed with a simplified Translation Model and build a generative model for its analysis.

Definition 1. *A Translation Model is Substitution Translation Model (STM), if every word in* w_{i1} *in* $L1$ *has a unique translation* w_{i2} *in* $L2$ *and vice-versa.*

Theorem 3. *In STM, if* w_{i1} *and* w_{i2} *are translations of each other, then* $CCS(w_{i1}, w_{i2}) = 1$

The converse of this theorem is not necessarily true. For instance, suppose w_{i1} and w_{j1} always co-occur, then their corresponding translations w_{i2} and w_{j2} always co-occur too. This leads to $CCS(w_{i1}, w_{j2}) = CCS(w_{i2}, w_{j1}) = 1$. This can be seen from Theorem 4. However, if w_{i1} and w_{j1} can occur independently in the language, in a large enough corpus, we should observe that w_{i1} and w_{j1} do not always co-occur.

Theorem 4. *In STM, if $w_{i1}, w_{j1} \in L1$ do not always co-occur in the same context, then $CCS(w_{i1}, w_{j2}) < 1$, where w_{j2} is the translation of w_{j1}.*

Proof. The numerator of $CCS(w_{i1}, w_{j2})$ in Substitution Translation Model counts the number of times w_{i1} occurs in a context D_m whenever w_{j2} occurs in its translation D_n. Since, w_{j2} must have its translation w_{j1} in D_m, the numerator is equivalent to counting the number of co-occurrence of w_{i1} and w_{j1}. Therefore,

$$\begin{aligned} CCS(w_{i1}, w_{j2}) &= \frac{\sum_m \sum_n I_{i1,m} e_{m,n} I_{j2,n}}{\max\{f_{i1} f_{j1}\}} \\ &= \frac{\sum_m I_{i1,m} I_{j1,m}}{\max\{f_{i1} f_{j1}\}} \end{aligned} \tag{8}$$

We define the following probabilities based on frequency counts -

$$P(w_i) = \frac{f_i}{|\mathbf{D}|}, \text{ and } P(w_i, w_j) = \frac{\sum_m I_{i,m} I_{j,m}}{|\mathbf{D}|}.$$

Without loss of generality, we may assume $P(w_{j1}|w_{i1}) \leq P(w_{i1}|w_{j1})$. Notice that $P(w_{j1}|w_{i1}) < 1$, otherwise, $P(w_{j1}|w_{i1}) = 1 \implies P(w_{j1}|w_{i1}) = 1 \implies P(w_{i1}) = P(w_{j1}) = P(w_{i1}, w_{j1})$, i.e., w_{i1} and w_{j1} always co-occur, which is a contradiction. Then, Equation 8 can be rewritten as

$$\begin{aligned} CCS(w_{i1}, w_{j2}) &= \frac{P(w_{i1}, w_{j1})}{\max\{P(w_{i1}), P(w_{j1})\}} \\ &= P(w_{j1}|w_{i1}) \frac{P(w_{i1})}{\max\{P(w_{i1}), P(w_{j1})\}} \\ &\leq P(w_{j1}|w_{i1}) < 1 \end{aligned} \tag{9}$$

Hence, $CCS(w_{i1}, w_{j2}) < 1$. More precisely $CCS(w_{i1}, w_{j2}) \leq \min\{P(w_{j1}|w_{i1}), P(w_{i1}|w_{j1})\}$.

We have shown that for a sufficiently large corpus, CCS between two words which are not the translation of each other is less than one. In practice, we would like this value to be small. To investigate the effect of size of the contexts (number of words in a context) and vocabulary size, we model the generation of contexts as an evolving bipartite graph. In [5], evolving bipartite graphs have been studied by constructing an analogy with Pólya's urn scheme to predict the degree distribution after infinite time. Proceeding with similar modeling, we assume that the vocabulary W^1 is fixed and the set of contexts $\mathbf{D^1}$ is growing. At each time step t, a new vertex (context) D_t^1 is introduced in $\mathbf{D^1}$. This new node

produces k edges and sequentially attaches with words in W^1. The words are selected preferentially [1] based on their degrees. Sequential attachment means that the degree of the words are updated after each edge attachment. So we use a more granular time τ, which increments after each edge attachment.

The following events take place at every time-step t.

- A new vertex D_t^1 is introduced in $\mathbf{D^1}$.
- A number k is selected with probability p_k from a probability distribution with first moment μ and second moment μ'.
- For each $r = 1, 2, \ldots, k$ an edge is introduced from D_t^1 to a word selected from W^1 according to the preferential rule based on its degree -

$$P(deg_\tau(w_i) = deg_{\tau-1}(w_i) + 1) \propto deg_{\tau-1}(w_i) + \delta$$

where δ is a constant to introduce some randomness in preferential attachment. We have dropped the subscript 1 from w_{i1} for ease of notation. Henceforth, unless mentioned otherwise, w_i refers to w_{i1}.

We assume that the process continues for infinite time. It can be shown [5] that the sequence of indicator functions representing if a word was selected at time τ is exchangeable. Therefore, by de Finetti's theorem [3] there exists a probability distribution function $f(\theta_i)$, such that the selection of a word w_i according to the above mentioned rules is equivalent to a collection of i.i.d. Bernoulli processes each with parameter θ_i, where θ_i is drawn from the distribution $f(\theta_i)$. Since $t = |\mathbf{D^1}|$,

$$\begin{aligned} P(w_i) &= \lim_{t\to\infty} \frac{1}{t} \sum_m I_{i,m} = \mathbb{E}(I_{i,m}), \forall m \\ &= \sum_k p_k (1 - (1-\theta_i)^k) \\ &= \sum_k p_k (k\theta_i + k(k-1)\theta_i^2/2 + \ldots) \end{aligned} \tag{10}$$

Assuming $k\theta \ll 1$ and ignoring terms with cubic and higher powers, and after some algebraic manipulation (omitted for brevity), we get

$$P(w_i) \cong (\mu' - \mu)\theta_i^2 \left(\frac{\mu}{(\mu' - \mu)\theta_i} - \frac{1}{2} \right). \tag{11}$$

Similarly we proceed to find the probability of co-occurrence $P(w_i, w_j) = \mathbb{E}(I_{i,m} I_{j,m})$. Suppose for a randomly selected context of size k, the number of edges attached to w_i is k_1, and the number of edges attached to w_j is k_2. Then

$$P(k_1 = l_1, k_2 = l_2 | k) = \binom{k}{l_1, l_2} \theta_i^{l_1} \theta_j^{l_2} (1 - \theta_i - \theta_j)^{k - l_1 - l_2}. \tag{12}$$

Now, co-occurrence probability is given by

$$\begin{aligned} P(w_i, w_j) &= \mathbb{E}(I_{i,m} I_{j,m}) \\ &= \sum_k p_k \sum_{l_1>0, l_2>0} P(k_1 = l_1, k_2 = l_2 | k) \\ &= \sum_k p_k (1 - P(k_1 > 0, k_2 = 0 | k) \\ &- P(k_1 = 0, k_2 > 0 | k) + P(k_1 = 0, k_2 = 0 | k)) \\ &= \sum_k p_k (1 - (1-\theta_i)^k - (1-\theta_j)^k + (1 - \theta_i - \theta_j)^k) \end{aligned} \tag{13}$$

Expanding to quadratic terms and some algebraic manipulations lead to

$$P(w_i, w_j) \approxeq (\mu' - \mu)\theta_i \theta_j \tag{14}$$

Finally, replacing the expressions for $P(w_i, w_j)$, $P(w_i)$ and $P(w_j)$ in Equation 9, obtained from the evolving graph model, we get

$$CCS(w_{i1}, w_{j2}) \approxeq \frac{\theta_i \theta_j}{\max\{\theta_i^2 \left(\frac{\mu}{(\mu'-\mu)\theta_i} - \frac{1}{2}\right), \theta_j^2 \left(\frac{\mu}{(\mu'-\mu)\theta_j} - \frac{1}{2}\right)\}} \tag{15}$$

Insight 1: Effect of context size and vocabulary size Since we expect two independent words to have low similarity, we get from Equation 15 that $\frac{\mu}{(\mu'-\mu)\theta_i} \gg 1/2$. This means that both θ_i and μ'/μ should be small. For an alternate interpretation, consider the case where the contexts have fixed size K. Then $\mu = K$ and $\mu' = K^2$, and Equation 15 becomes

$$CCS(w_{i1}, w_{j2}) \approxeq \frac{\theta_i \theta_j}{\max\{\theta_i^2 \left(\frac{1}{(K-1)\theta_i} - \frac{1}{2}\right), \theta_j^2 \left(\frac{1}{(K-1)\theta_j} - \frac{1}{2}\right)\}}. \tag{16}$$

Assuming without loss of generality that $\theta_i \geq \theta_j$, it can be shown that

$$CCS(w_{i1}, w_{j2}) \approxeq \frac{\theta_i \theta_j}{\theta_i^2 \left(\frac{1}{(K-1)\theta_i} - \frac{1}{2}\right)} \tag{17}$$

From this equation, it follows that $(K-1)\theta_i \ll 2$, i.e., the size of contexts should be small. Also, θ_i should be small, which is likely to happen if the vocabulary is large. Further, we have also made the assumption that the number of contexts $|\mathbf{D}^1|$ is very large.

Insight 2: The independence assumption While deriving Equation 15 using evolving bipartite graph model, we made an assumption that picking of a word in a context is independent of what words have already been picked in the same context. In practice, this may not be true, and so we may get some pairs which are not the translations of each other, and yet have high similarity. This, however, indicates a certain relatedness among these two words, and this could be useful because it may convey some more information, like synonyms of a word.

3.2 Model Evaluation

We experimentally verify the inferences drawn from the evolving bipartite model by applying the similarity measure on a cipher that fits Substitution Translation Model. since one to one mapping of words is not frequent among natural languages, it is improper to use them to verify the theory. However such mappings are observed in cryptanalysis tasks, where plaintext and ciphertext both are available. The objective of this experiment is not to solve a cipher but to gain more insight of CCS. We consider a simplified form of Substitution-Permutation cipher [10], as shown in Figure 3. The hypothetical cipher is applied by the following steps:

- Apply a one-to-one map of letter substitution to the plain-text.
- Choose a block size k and partition the string into blocks of size k (k characters).
- Permute the letters within each block.

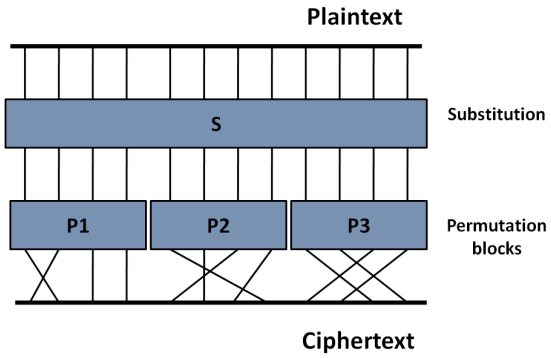
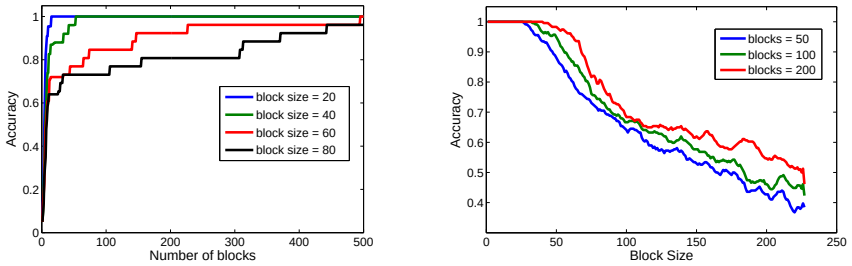


Fig. 3. A hypothetical substitution permutation: First a substitution is applied. Then the text is split into blocks of size k (here k=4). A random permutation is applied within each block to get the final ciphertext.

This fits the substitution translation model as for each letter w_{i1} in plaintext, there is a unique letter substitution w_{i2} in ciphertext. Presence of permutation within the blocks makes it difficult to recover the substitutions. However, by applying CCS and looking at the similarity values, we may recover the substitutions.

We take a long document in English [2] which has all the punctuations and spaces removed. The entire document consists of letters A to Z. We applied a random substitution from $\{A, B, \ldots, Z\}$ to $\{a, b, \ldots, z\}$, and split the resultant text into blocks of equal size. Then we applied permutation within each

[2] The document is publicly accessible at `https://www.dropbox.com/s/fwbiwb9s3nd3w5j/english.data`



(a) Accuracy vs number of blocks (b) Accuracy vs block size

Fig. 4. Experiments with deciphering: Effect of (a) number of blocks and (b) block size on accuracy. Both figures demonstrate that a high accuracy is achieved when block size is small and number of blocks is large.

block. We attempt to recover the substitutions by finding $CCS(w_{i1}, w_{j2}) \forall w_{i1} \in \{A, B, \ldots, Z\}$ and $w_{j2} \in \{a, b, \ldots, z\}$. Finally, substitution for w_{i1} is predicted using -

$$w_{i2} = \arg\max_{j2} CCS(w_{i1}, w_{j2}). \tag{18}$$

Note that the maximum value of $CCS(w_{i1}, w_{j2})$ in Equation 18 is always 1. Let $L = \{w_{j2} : CCS(w_{i1}, w_{j2}) = 1\}$. Theorem 3 guarantees that the original substitution will receive a CCS value of 1. However, since the converse is not true, more than one letter may receive a CCS value of 1 with w_{i1}, i.e., $|L| \geq 1$. If $|L| = 1$, then we select the one element as w_{i2}, otherwise we return a randomly selected element from $|L|$.

Figure 4 shows the accuracies obtained in the experiments. First, we fixed the block size and varied number of blocks (Figure 4(a)). As the number of blocks increases the accuracy of the method increases. Also, a higher accuracy is reached more quickly when the block size is small. If the block size increases, more number of blocks are required to achieve the same accuracy. Second, we fixed the number of blocks and varied the block size (Figure 4(b)). A decreasing trend in accuracy is clear with increasing block size. It can again be noted that for a given block size, larger number of blocks produces better accuracy. These results are consistent with our inferences from the evolving bipartite model, that small context (block) size and large number of contexts (blocks) are favorable for accurately learning the substitution.

Multilingual Word Association We used CCS to find translations of words among English and Spanish words using a corpus of parallel text[3]. CCS outperformed conditional probability, point-wise mutual information and cross-lingual LSI (See Figure 5). The details of the experiment have been omitted for brevity.

[3] The dataset is publicly available at `http://www.statmt.org/europarl/v7/es-en.tgz`

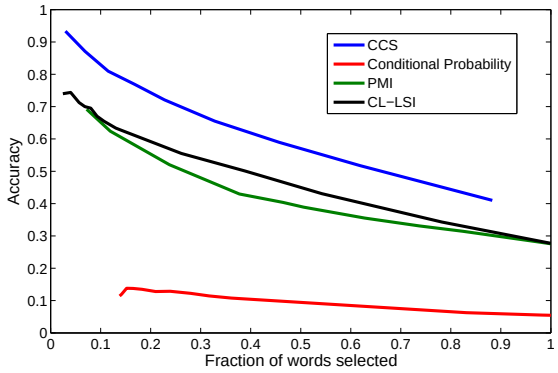


Fig. 5. Comparison of accuracy of *CCS* with baselines for English-Spanish word translation.

4 Conclusion

We introduced the concept of cross-context similarity, which is useful when context sizes are small and hence co-occurrence information is not enough to learn word relatedness. The cross-context similarity model can be used to mathematically analyze the relationship between vocabulary and context size: for cross-context similarity to be small for unrelated words, the size of the vocabulary should be large and the size of each context should be small. We demonstrated the correctness of this theory through substitution-permutation deciphering; deciphering is accurate when the permutation block size is small. As the block size increases, a larger number of blocks is required to achieve the same accuracy. We have also seen that *CCS* outperforms PMI and CL-LSI in the task of finding the right translation of a word among two languages.

In subsequent work, we will further expand on the multilingual word association task. We will also extend cross-context similarity to take into account hierarchical contexts, for instance, similarity of two paragraphs affects the similarity of the sentences, which in turn may affect similarity of terms. The cross-context similarity formulation is not limited to text datasets. We will explore similarity calculation in non-text applications such as ontology matching.

5 Acknowledgments

This work is supported by Chevron U.S.A. Inc. under the joint project, Center for Interactive Smart Oilfield Technologies (CiSoft), at the University of Southern California.

References

1. Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.
2. Guozhu Dong. Cross domain similarity mining: Research issues and potential applications including supporting research by analogy. *SIGKDD Explor. Newsl.*, 14(1):43–47, December 2012.
3. W. Feller. *An introduction to probability theory and its applications*, volume 2. Wiley, 3 edition, 1971.
4. Evgeniy Gabrilovich and Shaul Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJCAI*, volume 7, pages 1606–1611, 2007.
5. S. Ghosh, S. Saha, A. Srivastava, T. Krueger, N. Ganguly, and A. Mukherjee. Understanding evolution of inter-group relationships using bipartite networks. *Selected Areas in Communications, IEEE Journal on*, 31(9):584–594, 2013.
6. Zellig S. Harris. Distributional structure. In J. J. Katz, editor, *The Philosophy of Linguistics*, pages 26–47. Oxford University Press, 1985.
7. Eric H Huang, Richard Socher, Christopher D Manning, and Andrew Y Ng. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 873–882. Association for Computational Linguistics, 2012.
8. Aminul Islam, Evangelos Milios, and Vlado Kešelj. Text similarity using google tri-grams. In *Advances in Artificial Intelligence*, pages 312–317. Springer, 2012.
9. Jaz Kandola, Nello Cristianini, and John S Shawe-taylor. Learning semantic similarity. In *Advances in neural information processing systems*, pages 657–664, 2002.
10. Jonathan Katz and Yehuda Lindell. *Introduction to modern cryptography*. CRC Press, 2008.
11. Philippe Muller and Philippe Langlais. Comparing distributional and mirror translation similarities for extracting synonyms. In *Advances in Artificial Intelligence*, pages 323–334. Springer, 2011.
12. Patrick Pantel and Dekang Lin. Discovering word senses from text. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, pages 613–619, New York, NY, USA, 2002. ACM.
13. Siddharth Patwardhan, Satanjeev Banerjee, and Ted Pedersen. Using measures of semantic relatedness for word sense disambiguation. In *Proceedings of the 4th International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing'03, pages 241–257, Berlin, Heidelberg, 2003. Springer-Verlag.
14. Kira Radinsky, Eugene Agichtein, Evgeniy Gabrilovich, and Shaul Markovitch. A word at a time: computing word relatedness using temporal semantic analysis. In *Proceedings of the 20th international conference on World wide web*, pages 337–346. ACM, 2011.
15. Dennis Spohr, Laura Hollink, and Philipp Cimiano. A machine learning approach to multilingual and cross-lingual ontology matching. In *Proceedings of the 10th International Conference on The Semantic Web - Volume Part I*, ISWC'11, pages 665–680, Berlin, Heidelberg, 2011. Springer-Verlag.
16. Zheng Ye, Xiangji Huang, and Hongfei Lin. A graph-based approach to mining multilingual word associations from wikipedia. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 690–691. ACM, 2009.