二部グラフ上のランダムウォークによる言語横断関連語の抽出手法

Finding Cross-Lingual Related Words by Random Walk on Bipartite Graphs

Rudy Raymond[1]　坪井　祐太[1]　那須川　哲哉[1]　張 耀中[2]
[1] 日本IBM 東京基礎研究所　[2] 東京大学
{raymond, yutat, nasukawa}@jp.ibm.com
yaozhong.zhang@is.s.u-tokyo.ac.jp

1 Introduction

Many interactions in the real world can be expressed as bipartite graphs whose nodes can be partitioned into two parts (called *left* and *right* nodes) such that all edges of the graph link nodes from different parts. There are many natural examples of such graphs. For example, a subject-predicate bipartite graph (see Figure 1) whose nodes represent subjects or predicates, and whose edges link subject and predicate that appear on the same sentence. Proximity scores of nodes on such graphs have many important applications in recommendation, ranking similar nodes, link prediction, etc. In fact, the famous concept of the general framework for Distributional Similarity to measure the similarity of two nouns from their verb co-occurences as feature vectors can be regarded as measuring similarity of the corresponding nodes of the noun-verb bipartite graph.

Bipartite graphs have topological structures that can be exploited for designing efficient algorithms to compute proximity scores. One of the algorithms is the so-called Random Walk with Restart (or, RWR for short)[3], that calculates a proximity score of node i to node j from the steady-state probability of reaching j from i by a random walk. The principle of measuring similarity of node j from node i with RWR is to compute the similarity score from the steady state probability of reaching node j from a random walk (re)started from node i. The scores of RWR on bipartite graphs are easy compute, especially, when the number of left and right nodes is highly unbalanced. This has sparked widespread interest on measuring proximities with RWR, even for dynamic bipartite graphs [4].

In this paper, we propose using proximity scores of bipartite graphs for computing pair wise semantic

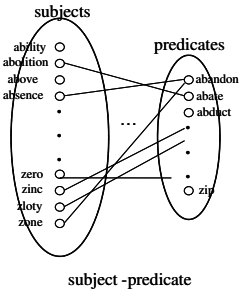

Fig. 1: A subject-predicate bipartite graph.

similarity between words in different languages. This application is motivated by [6] that developed a robust method to compare context similarity of words for identifying translation pairs. The idea can be summarized as in the left part of Figure 2. First, a feature vector of a target word in a language (say, "steering wheel") is constructed by the relation of the target word with all context words with respect to the corpus at hand. A similar procedure is performed to obtain a feature vector for each possible translation candidate in another language (say, "ハンドル" which is a Japanese word for steering wheel). Given a general dictionary for translating context words from English to Japanese (or vice versa), one can compare the vectors across languages. The main contribution of [6] is a new and robust statistical method to compare these vectors.

Notice that the relation of a target (or, a translation candidate word) with its context words is similar to the subject-predicate graph in Figure 1 and is essentially a bipartite graph. Moreover, by simply linking translation pairs across languages using a general dictionary one can connect the two bipartite graphs to obtain a bigger bipartite graph as shown in the right part of Figure 2. This way, proximity scores

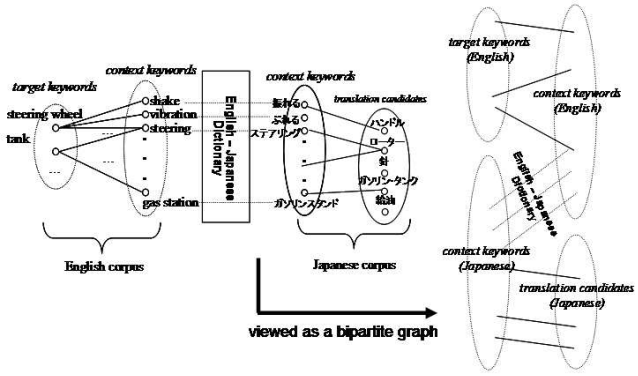

Fig. 2: (left) Identification of translation pairs of English and Japanese words by context similarity. (right) A bipartite representation of the context similarity method.

on bipartite graphs can be used to obtain proximity scores between a node corresponding to the target word with those of translation candidates which can then be used to identify translation pairs. Treating the relation of words in the framework of bipartite graph automatically gives some advantages over previous approaches such as incorporating user's feedback on the correct translation pairs (by dynamically adding links to corresponding node pairs) and the ability to refine the proximity scores by auxiliary information as proposed in [2]. We are not the first to propose random walk on graphs for Cross-Language Information Retrieval (CLIR) but we believe exploiting the structure of words as bipartite graphs deserve special attention.

To summarize, our contributions in this paper are: (1)A novel approach of using the framework of bipartite graph for identifying translation or related cross-lingual word pairs. The task of identifying related cross-lingual word pairs in this paper is similar to that in [5] where documents in multiple languages are analyzed from a perspective of a single language. (2)Experimental results using proximity scores of the RWR for computing pair wise semantic similarity between words in a corpus. We show that RWR and its variants are sometimes better than known approaches for identifying cross-lingual related words.

2 Definitions

We first explain notation, and then give the definitions of the problems we consider in this paper.

A bipartite graph $G(V, E)$ consists of a set of nodes V and a set of edges E. Its nodes can be partitioned into two disjoint sets: the left-node set L and the right-node set R such that $V = L \cup R$, and any edge $e \in E$ links a node in L with a node in R. Without loss of generality, we assume that the number of nodes in R, denoted as r, is at most that of nodes in L, denoted as l.

Tab. 1: Symbols in the RWR

Symb.	Description
$\mathbf{G}$	$l \times r$ adjacency matrix of bipartite graph G
L	the set of left nodes of G (of size l)
R	the set of right nodes of of G (of size $r \ll l$)
$\mathbf{D}_L$	the $l \times l$ diagonal matrix whose element at row i and column i is $\sum_j G_{ij}$
$\mathbf{D}_R$	the $r \times r$ diagonal matrix whose element at row j and column j is $\sum_i G_{ij}$
$\mathbf{I}$	an identity matrix
$\mathbf{0}$	a zero matrix
c	fly-away probability (fap) of RWR
$\mathbf{Q}$	the proximity score matrix of dimension $(l+r) \times (l+r)$ from the original RWR, which is partitioned into 4 parts: $\mathbf{Q}_1, \mathbf{Q}_2, \mathbf{Q}_3$ and $\mathbf{Q}_4$, each of dimension, $l \times l$, $l \times r$, $r \times l$, and $r \times r$, respectively, such that $\mathbf{Q} = \begin{pmatrix} \mathbf{Q}_1 & \mathbf{Q}_2 \\ \mathbf{Q}_3 & \mathbf{Q}_4 \end{pmatrix}$.

Bipartite graphs are represented by their adjacency matrices, and we use $\mathbf{G}$ for denoting the adjacency matrix of G. A non-negative element of $\mathbf{G}$ at row i and column j is denoted by G_{ij} and represents the link weight between nodes $i \in L$ and $j \in R$, and therefore $\mathbf{G}$ is a $l \times r$ matrix. In this paper, matrices are always denoted by bold capital letters, such as, $\mathbf{G}$, and $\mathbf{G}^T$ as its transpose, where its i-th row is denoted by $\mathbf{G}(i,)$, and its j-th column by $\mathbf{G}(,j)$. Following [4], we list all math symbols related to the RWR in Table 1.

Given the above notation, the input and output to compute proximity scores in bipartite graphs in this paper is formalized as follows.

[Proximity Scores]
Input: A bipartite graph $G(L \cup R, E)$ and a list of pairs of nodes $(u_1, r_1), (u_2, r_2), ...$ such that $u_i \in L \cup R$ and $r_i \in R$ as queries.
Output: The proximity scores of node r_i from node u_i for each query.

In above, we restrict the query to contain a node from the right set, with $r_i \in R$ as the second element of the query because we will mainly use the proximity scores for adjusting the proximity scores between the right nodes, and use the scores for predicting relations between left and right nodes,

3 The RWR on Bipartite Graphs

For any graph (not limited to bipartite), whose adjacency matrix is $\mathbf{M}$, the proximity score of node i to node j by the RWR is defined as the steady-state probability of being in node j when performing RWR (re)started at node i [3]. The value of the steady-state probability of RWR from node i to node j can be computed recursively from the equation

$$Q_{ij} = c \sum_k Q_{ik} \frac{M_{kj}}{\sum_l M_{kl}} + (1-c)\delta_{ij},$$

which is the sum of probabilities reaching nodes connected to j multiplied by the probabilities of moving from those nodes to node j. The second term of the right-hand side is due to the restart process. We can rewrite this equation to obtain the linear matrix equality

$$\mathbf{Q} = c\mathbf{Q}\mathbf{M} + (1-c)\mathbf{I},$$

where $\mathbf{M}$ is the row-normalized adjacency matrix, which is anti-diagonal with normalized $\mathbf{G}$ and $\mathbf{G}^T$ as its submatrices for a bipartite graph.

Since we are only interested in the ranking, the $1-c$ factor can be omitted. Thus, the proximity score matrix is $\mathbf{Q} = (\mathbf{I} - c\mathbf{M})^{-1}$, where $\mathbf{Q}_1, \mathbf{Q}_2$, and $\mathbf{Q}_3$ are linear in $\mathbf{Q}_4$, while the $r \times r$ matrix $\mathbf{Q}_4$ is obtained from the equation

$$\mathbf{Q}_4 = \left(I - c^2 \mathbf{D}_R^{-1} \mathbf{G}^T \mathbf{D}_L^{-1} \mathbf{G}\right)^{-1}, \qquad (1)$$

which is relatively easy to compute when r is small. The fact that all of the proximity scores of RWR on bipartite graphs are computable from $\mathbf{Q}_4$ results in computational advantages. We can instead compute the inverse of the smaller $r \times r$ matrix in Eq. (1) to obtain the inverse of the larger $\mathbf{I} - c\mathbf{M}$ matrix . We refer the readers to [3, 2] for details in computing proximity scores.

4 Experiment Results

We compare three cross-lingual IR approaches. (1)*SimpleRWR* is a random walk on a bipartite graph that consists of a single target English word, its English context words and their corresponding Japanese ones obtained from a dictionary, and a list of Japanese words as translation candidates. It is closely related to [1]. (2) *RWR* is a random walk on a bipartite graph that consists of 100 target English words, their context words and translation candidates determined similarly as the SimpleRWR. (3) *Baseline* is a simple implementation of [6] that has been shown to outperform other systems.

In all of the approaches, we manually picked 100 target English words (with their translations as gold-standard) each of which has at most 10 English highly associated context words. The context word is then paired with (possibly more than one) Japanese context words according to a dictionary. The translation candidates are the union of sets that each consists of 10 words that are highly associated to a Japanese context word. The association level between English target and context words is determined by the Pointwise Mutual Information (PMI) as described in [6], while that between Japanese context and candidate words is by the frequency for RWRs, and by the PMI for Baseline. This particular choice is made to optimize the accuracy for each method.

We use datasets in complaints about cars: the English corpus is provided by the USA National Highway Traffic Safety Administration (NHTSA)[1] that consists of 618,437 documents, while the Japanese corpus is by the Japanese Ministry of Land, Infrastructure, Transport and Tourism (MLIT)[2] that consists of 24,458 documents. The size of the resulting bipartite graph is 2,375 × 3,239 with 10,146 edges. All of the approaches return a ranked list of translation candidates for each target word, and therefore the accuracy is measured by the existence of a correct translation word in the list. Tab. 2 shows examples of ranked-list of translation candidates for the English word “gas pedal”, “gas tank”, “heater”, “sensor” and “speedometer”.

Figure 3 shows the averaged accuracy for each of the aforementioned approaches with regards to the length of the candidate list. We can see that despite its simplicity, SimpleRWR already outperforms Baseline, while RWR in overall achieves the best accuracy performance.

[1] http://www-odi.nhtsa.dot.gov/downloads/index.cfm
[2] http://www.mlit.go.jp/jidosha/carinf/rcl/defects.html

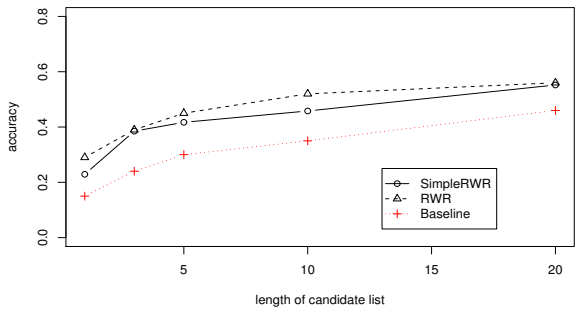

Fig. 3: Comparing RWR methods with the baseline using the context similarity

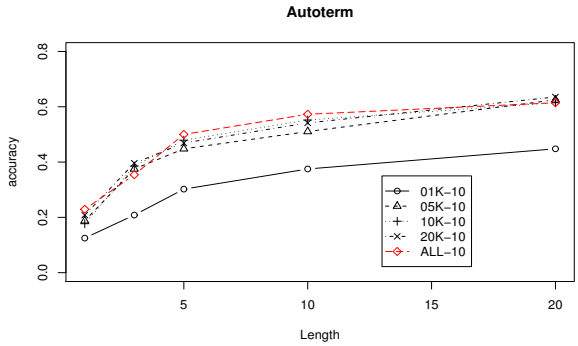

Fig. 4: Accuracy vs. varying the size of NHTSA documents by 1K, 5K, 10K, 20K

Figure 4 shows the effect of the size of the corpus against the accuracy in the RWR method. We randomly chose 1K, 5K, 10K and 20K NHTSA documents (while the MLIT data is kept fixed) and applied the RWR method on each of the resulting bipartite graph. We can see the tendency that the larger the English corpus, the better the accuracy is. However, the accuracy of the RWR with only 5K English documents is already close that with all documents. (Although not directly comparable, we should note that the previous work [6] using Baseline reported that approximately 20K English documents are required to achieve similar stability in accuracy.)

5 Concluding Remarks

We have presented a novel framework to use proximity scores between nodes of bipartite graphs for identifying related words between documents in different languages. For English-to-Japanese translation pairs, the accuracies are quite good and the experiment on the reverse direction is left for future work. Meanwhile, output examples in Tab. 2 also showed that the translation candidates are useful for cross-lingual text mining task as in [5]. For example, we obtain that “ブレーキペダル”(brake pedal) and “フロアマット”(floor mat) are among keywords that are identified close to “gas pedal”, which are suspiciously reflected the recent brake pedal problem in the US.

target word	top-10 ordered translation candidates
gas pedal	ブレーキペダル, アクセルペダル, アクセル, 回り, ペダル, ブレーキ, 再発, フロアマット, 固定, 足
gas tank	燃料タンク, 燃料, 給油, ガソリン, ガソリンタンク, 臭, 噴射, 満, タン, オイル
heater	ヒーター, 冷却, 中心部, ハンドル, 冬季 エアコン, ファンベルト, 車両, ディーラー, 曇り
sensor	酸素センサー, カムシャフト, プーリー, センサー, オイル, ボルト, オイルシール, 調整, プラグ, 運転
speedometer	スピードメーター, メーター, デジタル表示, ドット, 切り替え, 速度計, 指針, 針, 燃料, タコメーター

Tab. 2: Examples of ordered Japanese translation candidates of English auto keywords by the RWR method. Underlined Japanese keywords are the correct ones.

参考文献

[1] Cao, Gao, Nie, and Bai. Extending query translation to cross-language query expansion with markov chain models. In *CIKM'07*, 2007.

[2] Raymond, Tsuboi, Nasukawa, Sato, and Kashima. Finding related words by random walk utilizing auxiliary information with application in cross-lingual text mining. In *Submitted*, 2011.

[3] Tong, Faloutsos, and Pan. Random walk with restart: Fast solutions and applications. *Knowledge and Information Systems: An International Journal (KAIS)*, 2008.

[4] Tong, Papadimitriou, Yu, and Faloutsos. Proximity tracking on time-evolving bipartite graphs. In *SDM '08*, pp. 704–715. SIAM, 2008.

[5] 海野, 那須川. 言語横断テキストマイニング. In *JSAI2010*. 第２４回人工知能学会全国大会, 2010.

[6] 那須川, Andrade, 海野, 村松, 山本. 言語横断テキストマイニングのための翻訳対抽出. In *NLP2009*, pp. 108–111. 言語処理学会, 2009.