SBA-term: Sparse Bilingual Association for Terms

Xinyu Dai*, Jinzhu Jia†, Laurent El Ghaoui‡, Bin Yu§
*National Key Lab for Novel Software Technology, Nanjing University, Nanjing, China
Email: daixinyu@nju.edu.cn
†School of Mathematical Sciences, Beijing University, China, Email:jzjia@math.pku.edu.cn
‡Department of EECS, University of California Berkeley, CA 94760, Email:elghaoui@berkeley.edu
§Statistics Department, University of California Berkeley, CA 94760, Email:binyu@berkeley.edu

***Abstract*—Bilingual semantic term association is very useful in cross-language information retrieval, statistical machine translation, and many other applications in natural language processing. In this paper, we present a method, named SBA-term, which applies sparse linear regression (Lasso, Least Squares with l_1 penalty) and L^2 rescaling for design matrix to the task of bilingual term association. The approach hinges on formulating the task as a feature selection problem within a classification framework. Our experimental results indicate that our novel proposed method is more efficient than co-occurrence at extracting relevant bilingual terms semantic associations. In addition, our approach connects the vibrant area of sparse machine learning to an important problem of natural language processing.**

I. Introduction

Semantic word association, or more generally, *term association* plays an important role in many types of natural language processing tasks and applications. It can be useful for query expansion in information retrieval, as well as for candidate sentence selection in question answering and document summarization. Term association can also be used to drive semantic clustering, which is helpful for language models. A common approach for the term association task is to use co-occurrence or mutual information between terms within a large corpus; often additional knowledge sources such as Wordnet and Wikepedia are used to improve the association results.

Moving beyond the single-language case, term association between two languages, or *bilingual term association*, is also vital in many cross-language tasks and applications. In cross-language information retrieval [1], a non-Chinese speaking user may want to get some information from Chinese documents (say, news articles): bilingual term association can help translate or expand the English query terms into some Chinese terms. Bilingual term association can also be applied to word alignment in statistical machine translation [2]. The data used in bilingual term association usually includes translation dictionaries, and some parallel documents if available.

Unfortunately, translation dictionaries cannot provide all the relevant semantic associations between bilingual terms, other than those given by direct translation. For example, the English-Chinese dictionary can translate the English word “bush” into “布什” (person surname) or “灌木” (shrub). The dictionary fails to recognize the strong semantic association, which could be found with bilingual parallel documents, between the English word “bush” and the Chinese term “美国总统” (American President). Another deficiency of dictionary is poor coverage, a problem arises when new terms, such as ”Tea Party”, are introduced.

Clearly, if we have a large bilingual parallel data set, we may use co-occurrence [11] to associate the bilingual term pairs. Given a term, co-occurrence can generate the terms appearing the most often together with the given term, within a context window. Other frequency-based methods, like log-likelihood [10], can also be used for bilingual term association.

In this paper, we propose a novel approach that relies on a specific sparse regression method, the Lasso [3] that has been very popular in machine learning and statistics recently. We formulate the bilingual term association task as one of selecting a few features (terms in one language) that best predict the appearance of a given term in the other language in a bilingual parallel data set. To solve this feature selection task, we invoke the LASSO method, which is Least Squares with an l_1-norm penalty in order to encourage sparsity of the linear regression coefficients. The features (terms) selected correspond to the (few) non-zero values of the regression coefficients. To our knowledge, this is the first application of LASSO on the task of bilingual term association. Our focus in this paper will be on Chinese-English corpora. This work is part of the StatNews project[1], which aims at providing fast summarization of topics in multi-lingual news databases [4], [9].

Our paper is organized as follows. The SBA-term method is presented in section II. Some results and a case of bilingual association network graph is showed in section III. Concluding remarks are given section IV.

II. SBA-term: Sparse Bilingual Association for Terms

A. Task Description

We are given a large parallel Chinese-English dataset with thousands of sentences pairs. For a given Chinese term, chosen in a set of n terms $J = \{C_1, C_2, ..., C_n\}$, we would like to provide a few English terms within the set of m terms $I = \{E_1, E_2, ..., E_m\}$ that have strong semantic association with the original Chinese term. Correspondingly, if given an English term, we are interested in its Chinese associated terms.

[1] `http://www.eecs.berkeley.edu/~elghaoui/StatNews/`

We will set up the bilingual term association task as a feature selection problem, within a classification framework involving the so-called LASSO model. Our next sub-section describes this model in more detail.

B. The LASSO method

Linear regression is a classical approach to modeling the relationship between a response variable Y and one or more predictors denoted X. Gaussian linear regression is formulated as a least-squares optimization problem, which can be efficiently solved. LASSO is a relatively recent variant on least-squares, which includes an l_1-norm penalty. Precisely, LASSO takes the form

$$\hat{\beta}^{\lambda} = \arg\min_{\beta} \|Y - X\beta\|^2 + \lambda \sum_{i=1}^{n} |\beta_i|, \qquad (1)$$

where $X = (X_1, \ldots, X_p)$ is the $n \times p$ *design matrix* whose columns consist of the n-dimensional fixed predictor variables X_k, $k = 1, \ldots, p$. The vector Y contains the n-dimensional set of real-valued observations of the response variable. The λ in (1) controls the amount of shrinkage that is applied to the estimates of β, and the penalty term (sum of absolute values of the β_i's) encourages many zeroes in the solution. Effectively, a large value of λ results in a very sparse β vector, which in terms allows to identify those features (in our case, terms) that have good predictive value. More details on the interpretation of LASSO, and related recent algorithms, can be found in [7]. In this paper, we use a fast algorithm for LASSO when data are large but sparse [9]. This algorithm is a modification of the BBR [8] from the sparse logistic model to linear regression model or LASSO.

C. Algorithm

The next step is to apply the LASSO model to our bilingual term association task. Recall that our data set consists in l Chinese-English sentence pairs. Each sentence pair means a Chinese sentence and its corresponding English sentence. An example is given in Figure 1. We consider each sentence pair as one document.

我 喜欢 踢 足球 ， 但是 她 喜欢 打 排球 。
I like to play football, but she like to play volleyball.

Fig. 1. Example of Chinese English sentence pair

Based on the m English indexed terms, and n Chinese indexed terms, we construct two document-term matrices, M_E and M_C, of size $l \times m$ and $l \times n$ respectively. The element M_{ij} in M_E or M_C is the times of term E_j or C_j appearing in the ith document. Now consider the task of finding terms associated with a given term in Chinese, say C_j. Let Z to be the column in M_C corresponding to that term: that is, Z is the observed appearance times of C_j across all documents. Further, set Y to be a vector containing the *signs* of Z. Now choose the design matrix X to be the full English-language M_E. As commonly used together with LASSO, L^2-rescaling is used to reweight the design matrix to reduce the impact of the larger variance and weight of higher frequency features. And according to our previous work [9], L^2-rescaling is really better than $tf * idf$ rescaling when the document is short like our sentence pairs. By applying LASSO, we are effectively trying to model the "response" (the appearance or not of Chinese term C_j) as a linear combination of the L^2-rescaling scores of English terms. Because LASSO encourages a sparse result, only a few English terms are selected as predictors by LASSO if a large λ is used. Those terms that receive a non-zero value of the regression coefficient are precisely those which we select as associated terms.

A summary of the SBA-term algorithm is given below.

- Preprocess the parallel data. Tokenize English sentences and segment Chinese sentences into tokens. Index all the Chinese and English terms.
- Construct the document-term matrix for each language.
- Given a Chinese or English term, generate the response Y and design matrix X as described above.
- Run LASSO to select a fixed number of predictors, and generate the associated terms in another language by finding the terms with the non-zero coefficients.

The algorithm relies on a choice of the regularization parameter λ. Before running Lasso, it should be pointed out that we should select a value for λ. As we have mentioned in section II-B, a larger λ results in fewer features (terms selected). In our application, it will be more natural for us to choose how many features (associated terms) we want to keep. Thus if we need to get k selected features, we need to search over λ until the appropriate number selected. Note that several very efficient algorithms exist for solving LASSO, including ones that generate the entire path of solutions as λ changes. Here, k (or in fact, λ) can be interpreted as a tuning parameter that allows to control the number of the associated terms.

III. RESULTS

Our bilingual dataset is from LDC data[2]. There are $466,991$ Chinese-English sentences pairs in our dataset. After removing all words appearing fewer than five times in the dataset, we obtain a dictionary with $39,734$ Chinese words and $30,093$ English words to be indexed.

We first do an experiment with unigrams. Table I shows some Chinese/English words and their associated English/Chinese words. From Table I, we first compare the co-occurrence approach and our SBA-term method. Co-occurrence seems to generate more stop-words in the list of associated words, such as some function words and punctuations. SBA-term is more robust: although we did not remove stop words as a pre-processing step, SBA-term does not select them. And SBA-term gives us more semantically associated terms like "countryside" "villages" in the third column of Table I.

In our results, the associated words are sorted by the β-coefficient values from the LASSO solution (eq. 1). In the

[2]Including LDC2003E14, LDC2003E07, LDC2005T10, LDC2006E34, LDC2006E85, and LDC2006E92 http://www.ldc.upnn.edu

TABLE I
CHINESE/ENGLISH TERMS AND THEIR ASSOCIATED ENGLISH/CHINESE TERMS

Chinese/English Term	Co-Occurrence (top 10)	SBA-term (k =10)	SBA-term (k=20)
农村	rural, areas, peasants, and, in, agricultural, agriculture, of, the, reform	rural, peasants, countryside, agriculture, villages, agricultural, areas, urban	rural, peasants, countryside, villages, agriculture, areas, agricultural, peasant, farming, feetotax, anhui, farmers, burdens, secondround, dezhan, cities, township, urban
人权	rights, human, china, on, united, and, states, the, of, in	human, rights, humanrights, antichina, intellectual, beings, property, legitimate	human, rights, humanrights, antichina, democracy, robinson, united, falungong, issues, states, cshrs, creditor, obligations, resources, genome, intellectual, property, beings, legitimate
capital	资本, 首都, 外资, 资金, 、, 投资, , , 市场, 和, 企业	资本, 首都, 外资, 资金, 首府, 资本金, 基本建设, 台资	资本, 首都, 外资, 资金, 首府, 资本金, 基本建设, 台资, 基建, 都城, 省会, 投资, 古都, 省城, 本钱, 生产资料, 游资, 兑换, 迁都
struggle	斗争, 、, 和, 反, 奋斗, ” , “, 腐败, 的, 人民	斗争, 奋斗, 艰苦奋斗, 阶级斗争, 争斗, 挣扎, 负隅顽抗, 垂死挣扎	斗争, 奋斗, 艰苦奋斗, 阶级斗争, 争斗, 挣扎, 争, 负隅顽抗, 垂死挣扎, 搏斗, 明争暗斗, 氛, 恶斗, 白景富, 角力, 拼搏, 争权, 决斗, 反霸, 悉尼

SBA-term lists, the first one or two associated words are almost the exact translations of the corresponding words. Most of the other terms in the list are very semantically related words. In Table I, we also show results with different values for the number k of selected features, $k = 10$ and $k = 20$. It seems that the larger k can present us more meaningful associated words, such as ”人权” (“human rights”), which is associated to ”falungong”. However, a larger value of k may also result in more noisy term lists.

Our approach can be used to detect “pair-wise” bilingual associations. In the method described above, the association between two terms in different languages are unidirectional. Given a Chinese word C_i, SBA-term yields associated English words, which we symbolize as $C_i \rightarrow E_j$. If in turn, E_j is also associated with $C_i (E_j \rightarrow C_i)$, then we can declare the association to be pair-wise, or bi-directional. Bi-directional associations are stronger, and we can be more confident about such associations. From Table II, we observe that indeed, the bilingual terms have the closest relationship when they are in pairwise association.

The choice of k, as said before, is to set a parameter in our method. From Table I, we can see that a smaller value of k generates semantically closer associations than a larger k. The choice of k depends on the specific application. In cross-language information retrieval, we can focus on pairwise association or choose a smaller k. In some summarization applications [4], a larger k may give us more meaningful association words, and further help summarization.

Based on term association, lexical chain and term association network is also an attractive research topic [5]. We can also construct a graph based on our term association results like Figure 2. To make our graph more legible, we first selected the 100 most frequent Chinese and English terms, and just kept the pairwise association in the LASSO-generated graph. The nodes represent Chinese or English words, which are distinguished by different colors. Here, the edges between nodes represented their association, as given by LASSO (with $k = 10$). The thickness of edges represents the association degree between two nodes, which depends on the magnitude of β value from LASSO. Figure 2 illustrates how our proposed method is also a possible and efficient techniques to construct bilingual term association network. Note that LASSO is at the heart of rigorous methods allowing to construct sparse graphs (graphical models) based on data, see [6].

TABLE II
CHINESE/ENGLISH PAIRWISE ASSOCIATION

Chinese Words	English Words
农村	rural
农村	agriculture
国际	world
国际	international

From Table I, we can see the Chinese words ”人权” has associated English words like ”humans” and ”rights”. This leads to consider if bigram terms can help us get more interesting associations between two languages. Based on the same dataset and also removing all unigram or bigrams appearing fewer than five times, we get $267,030$ Chinese tokens of unigrams and bigrams, and likewise $221,354$ English tokens. The result in Table III shows the English words ”capital” almost get the same associated Chinese words as when we use unigrams only. But the Chinese words ”人权” and ”美国” get their more exact associated terms, such as ”human rights” and ”united states”. And some Chinese named entity like ”江泽民” and ”国务院” also get better associated terms like ”jiang zeming” and ”state council”.

IV. CONCLUSION AND FUTURE WORK

Our work in this paper focuses on bilingual semantic terms association in bilingual documents context. Other than co-occurrence and other purely frequence-based methods, our SBA-term apply l_1 norm penalized linear regression method to generate sparse features which corresponding association

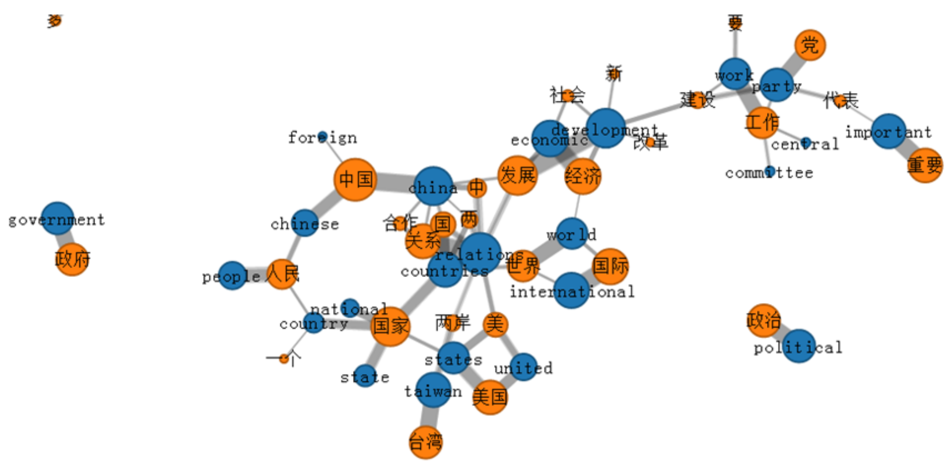

Fig. 2. Bilingual terms association graph

TABLE III
CHINESE/ENGLISH TERMS AND THEIR ASSOCIATED ENGLISH/CHINESE BI-GRAM TERMS

Chinese/English Term	Lasso (k =10)
capital	资本, 首都, 外资, 资金, 首府, 投资
美国	united states, the us, us, american, the united, states, bush
人权	human rights, rights, rights and, and human, the united, human, falungong, china
江泽民	jiang zemin, zemin, comrade jiang, jiang, party, central
国务院	state council, state department, the state, premier, state, committee, central

terms in our task scenario. SBA-term can expand our horizon and make the sparse model applicable to more natural language-related tasks. After L_2-rescaling with LASSO, the SBA-term provides us more semantically association terms than co-occurrence and removes stop-words automatically. Following the statnews framework, we will perform human evaluations to exam our approach against human understanding. To bring confidence to the selection of features that is lacking in using a fixed number of features(k in our method), we would like to test statistical significance on the features. We will run the Lasso on several resampled data sets. Features passing tests at a 5% level are kept and the rest not. And the selected terms from SBA-term will be used for real applications, such as query expansion in cross-language information retrieval and word alignment in machine translation to further validate our proposed approach.

ACKNOWLEDGMENT

We would like to thank Brian Gawalt and Luke Miratrix for their useful suggestion for this work. Xinyu Dai thanks the departments of EECS and Statistics at UC Berkeley for hosting his visit sponsored by China Scholarship Council, during which this work was done. Jinzhu Jia's work was done when he was a postdoc in the Statistics Department at UC Berkeley. Support for Xinyu Dai from NSFC (No.61003112), 973 Program (No.2010CB327903), is gratefully acknowledged. The work of Jinzhu Jia, Laurent El Ghaoui, and Bin Yu is supported in part by National Science Foundation grants SES-0835531, CMMI-0969923, and CCF-0939370.

REFERENCES

[1] Jian-Yun Nie. *Cross-Language Information Retrieval*. Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers, 2010.
[2] Robert C. Moore, *Association-based bilingual word alignment*, In Proceedings of the ACL Workshop on Building and Using Parallel Texts, 2005.
[3] Robert Tibshirani, *Regression Shrinkage and Selection Via the Lasso*, Journal of the Royal Statistical Society, Series B, 1996.
[4] Brian Gawalt, Jinzhu Jia, Luke Miratrix, Laurent El Ghaoui, Bin Yu and Sophie Clavier, *Discovering Word Associations in News Media via Feature Selection and Sparse Classification*. Proc. ACM International Conference on Multimedia Information Retrieval (MIR2010), Philadelphia, PA, Mar. 2010.
[5] Thad Hughes and Daniel Ramage. *Lexical Semantic Relatedness with Random Graph Walks*. In Proceedings of EMNLP-CoNLL2007.
[6] O.Banerjee, L. El Ghaoui and A. d'Aspremont, *Model Selection Through Sparse Maximum Likelihood Estimation for Multivariate Gaussian or Binary Data*. Journal of Machine Learning Research, 485-516. 2008.
[7] Friedman, J. and Hastie, T. and Tibshirani, R. *Regularization paths for generalized linear models via coordinate descent*. Journal of Statistical Software:33-1,2010
[8] Genkin, A. and Lewis, D.D. and Madigan, D. *Large-scale Bayesian logistic regression for text categorization*. Technometrics:49-3.2007
[9] Luke Miratrix, Jinzhu Jia,Brian Gawalt, Bin Yu, and Laurent El Ghaoui.*Summarizing large-scale, multiple-document news data: sparse methods and human validation*. Technical Report 801, Statistics Department, UC Berkeley.
[10] Ha, L. A., Fernandez, G., Mitkov, R. and Corpas, G. Mutual,*bilingual terminology extraction*. In proceedings of LREC 2008, Marrakesh, Morocco
[11] Gaussier, E.. *Flow Network Models for Word Alignment and Terminology Extraction from Bilingual Corpora*. In Proceedings of ACL1998, San Francisco, California.