RC25273 (WAT1203-073) March 28, 2011
Computer Science

IBM Research Report

Towards Cross-Lingual Sentence Similarity Comparison: A Generalized Non-Negative Matrix Factorization Approach

Juan M. Huerta
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Cambridge - Haifa - India - T. J. Watson - Tokyo - Zurich

Towards Cross-Lingual Sentence Similarity Comparison: A Generalized Non-Negative Matrix Factorization Approach

Juan M. Huerta
IBM T J Watson Research Center
1101 Kitchawan Road
Yorktown Heights, NY 10598
Huerta@us.ibm.com

Abstract

We describe a novel approach to cross-lingual sentence similarity comparison based on the joint factorization of language specific document-term matrices. Our approach takes two language specific document-term matrices representing a parallel corpus and obtains language specific projections intended to map sentences from their respective languages into a reduced rank common subspace. These projections are obtained using a joint multiplicative non-negative matrix factorization with sparsity constraints that minimizes the joint divergence. Our technique is demonstrated in a cross-lingual SMT/human translation classification task attaining a 73% accuracy which is significantly better than previously obtained using a long-span n-gram classifier.

1 Introduction and Related Work

The goal of cross lingual sentence similarity comparison is to evaluate the extent of the similarity between two sentences in different languages. The problem has been previously formulated in different ways such as: cross-lingual topic classification (Bei et al. (2003), Ni et al. (2011)), document retrieval and ad relevance (Yih et al (2011)), et cetera. Additional potential applications include parallel and comparable corpora creation, Web content analysis and organization, SMT output evaluation, among others. Many of the existing approaches to the problem are based multi-lingual extensions to topic models (Bei et al, (2003), etc). Most of the previous work on monolingual sentence similarity has typically focused on *semantic* similarity (e.g., Mihalcea et al. (2006)). A big segment of previous approaches to monolingual sentence similarity is based on LSA and SVD (Landauer et al. (1998), among others). Our approach can be seen as a multilingual generalization of LSA.

Methods like BLEU (and other SMT evaluation metrics) can be seen, to a certain extent, as mono-lingual sentence similarity measures. These compare sentences directly and when the sentences are in different languages, translation of one of the sentences is necessary. Our method avoids the use of Machine translation and instead compares two sentences directly through a pair of projection functions into a common subspace.

In this work we describe a new approach to broad cross lingual sentence similarity (e.g., addressing the question: Are these two sentences in different languages sufficiently similar?). Our technique is based on two projection matrices that are obtained from a training parallel corpus. The work of Yih et al. (2011) is similar to ours in the sense that they also obtain language dependent projections into a common subspace, however their focus is on discriminatively minimizing the classification rate. Our technique is based on different criteria. We now explain how to obtain these projections from the matrix representation of the parallel corpus.

2 Document-Term Matrix Representation

Let corpus C represent a very large set of sentence pairs:

$C = \{\{f_1, o_1\}, \{f_2, o_2\}, ..., \{f_n, o_n\}\}$

where f_i is the i^{th} sentence in foreign language, and o_i is the i^{th} sentence in the original language (or source language). We assume there are n sentences in the corpus.

We model corpus C with two separate matrices. Let matrix Y_1 be an n_1 x m matrix representing the sentences in the *foreign* language. Each of the m columns is associated with a specific foreign sentence. Rows are associated with language-specific linguistic feature counts (words, word n-grams, part-of-speech (POS) tags, POS n-grams, word dependency arcs, etc). Thus the element in the i^{th} row and j^{th} column is the integer representing the number of counts of feature i in sentence j (other related functions like TF-IDF etc can be used).

Similarly, let Y_2 be an n_2 x m matrix represent the sentences in the *source* language. Similarly, each of the m columns is associated with a specific source sentence. Rows are associated with language specific features. Because the linguistic features are non-overlapping (e.g., words and features are different across languages) both feature spaces are said to be disjoint.

Matrices Y_1 and Y_2 need to have the same number of columns because they represent a parallel corpus, but they do *not* need to have the same number of rows. In addition, Y_1 and Y_2 have non-negative entries, and are typically very sparse.

3 Generalized Non-Negative Factorization

Motivated by the Generalized Singular Value Decomposition (Golub and Van Loan (1996)) we propose a joint factorization of Y_1 and Y_2 in order to obtain representation of elements from disjoint subspaces in a common low-rank subspace. This type of factorization has proven useful in applications from text and language analysis (Landauer et al. (1998)), to Recommender Systems (Tikk et al. (2008)) among others. Given the non-negative nature of our features (i.e., counts) we will focus on non-negative factorizations. NMF was introduced and described by Lee and Seung (2001). Further motivations for non-negativity are provided in Wild et al. (2003) and Xu et al. (2003). We observe that the GSVD theorem states that given matrices A and B there exists a U, a V and an invertible X such that $U^T AX = C$ and $V^T BX = S$ where both C and S are non-negative and diagonal. In this factorization, matrices A and B share a common matrix X. Thus, in our particular case, we want to factorize Y_1 and Y_2 as follows:

$$Y_1 = A_1 X + V_1, A_1 \geq 0, X \geq 0$$
$$Y_2 = A_2 X + V_2, A_2 \geq 0, X \geq 0$$

Where V_1 and V_2 are noise or approximation distortion terms. In the above factorization, matrices A_1 and A_2 have a language specific term-to-concept interpretation similar to LSA. The common matrix X, in turn, has a sentence-to-concept interpretation, and thus models a common subspace.

These factorizations are useful when evaluating the similarity of an unpaired couple of vectors by mapping these into the common subspace where the actual comparison takes place. The common space is equivalent to what LSA (Landauer et al. (1998)) describes as the *concept space*.

Specifically, after obtaining the above factorizations, we can map feature vector y_1 and y_2 represented in the original feature spaces into the common space as follows:

$$Y_1 = A_1 X + V_1, A_1 \geq 0, X \geq 0$$
$$\hat{x}_1 = (A_1^T A_1)^{-1} A_1^T \hat{y}_1$$
$$Y_2 = A_2 X + V_2, A_2 \geq 0, X \geq 0$$
$$\hat{x}_2 = (A_2^T A_2)^{-1} A_2^T \hat{y}_2$$

In practice the factorization described above is a reduced rank representation of order k, where k is typically much smaller than n. The above joint factorization can be seen as a multilingual generalization of LSA.

4 Iterative Non-Negative Solution

Multiplicative approaches for non-negative matrix factorization (NMF) were introduced in (Lee and Seung (2001)). Dhillon and Sra (2006) focused on

NMF using a divergence type of loss function. Cichoki et al. described sparse NMF under different loss functions. Other authors have focused on improved optimization methods to iteratively find the NMF solutions (e.g., Lin (2007), and Mairal et al. (2010)).

Our approach is based on the divergence criterion. The single matrix divergence between Y and AX is defined as:

$$D(Y \| AX) = \sum_{ik} \left(y_{ik} \ln \frac{y_{ik}}{[AX]_{ik}} + [AX]_{ik} - y_{ik} \right)$$

In the case of our particular joint factorization the objective function, i.e., the joint divergence, corresponds to the sum of the individual divergences:

$$D(Y_1 \| A_1, A_2, X) = D(Y_1 \| A_1, X) + D(Y_2 \| A_2, X)$$

We want to obtain non-negative, sparse X, A_1 and A_2 matrices that minimize the divergence. To obtain a sparse solution we add l-1 sparsity constraints.

In a way similar to equation (11) in Cichocki et al. (2006), we compute the gradient with respect to elements of matrices and choose the suitable learning rates we obtain a final multiplicative set of iterative update rules (the derivation of these results is omitted here for space purposes):

$$x_{jk} \leftarrow x_{jk} \frac{\left[\sum_{i=1}^{m} a^1_{ij} (y_{ik} / [A^1 X]^{1-\beta}_{ik}) + a^2_{ij} (y_{ik} / [A^2 X]^{1-\beta}_{ik} - \alpha_X \right]}{\sum a^1_{ij} [A^1 X]^{\beta}_{ik} + a^2_{ij} [A^2 X]^{\beta}_{ik}}$$

$$a^1_{ij} \leftarrow a^1_{ij} \frac{\left[\sum_{k=1}^{N} (y_{ik} / [A^1 X]^{1-\beta}_{ik}) y_{jk} - \alpha_X - \alpha_c \sum_k a^2_{kj} \right]}{\sum a^1_{ij} [A^1 X]^{\beta}_{ik} + a^2_{ij} [A^2 X]^{\beta}_{ik}}$$

$$a^2_{ij} \leftarrow a^2_{ij} \frac{\left[\sum_{k=1}^{N} (y_{ik} / [A^2 X]^{1-\beta}_{ik}) y_{jk} - \alpha_X - \alpha_c \sum_k a^1_{kj} \right]}{\sum a^2_{ij} [A^2 X]^{\beta}_{ik} + a^1_{ij} [A^1 X]^{\beta}_{ik}}$$

Where the α_x and α_c are empirically chosen constants introduced to control the sparsity of the solution. β is chosen to match the underlying distribution of the data (Cichocki et al. (2006)).

5 Experiments

We conducted two types of experiments intended to evaluate the empirical qualities of our approach. In the first experiments we focused on assessing the convergence and modeling characteristics of our solution. For this purpose we created simulated matrices Y_1 and Y_2 each with 100 rows and 500 columns using randomly generated positive numbers.

Figure 1 top panel, shows the logarithm of the frobenius norm of the residual (error) matrix of the estimate as a function of number of iterations for a benchmark additive algorithm in solid line (interior point NMF-GSVD corresponding to eq. 16 in Cichocki et al. (2006)), as well as for our approach, shown in dash-dot line. We can see how our approach provides much faster convergence than the baseline additive method.

The second panel shows the error rate as a function of iterations under several values of k (the first iteration's error value (resulting from random initializations) is not shown to provide better detail in the figure). We can see in this plot how higher k provides better modeling (i.e., less error).

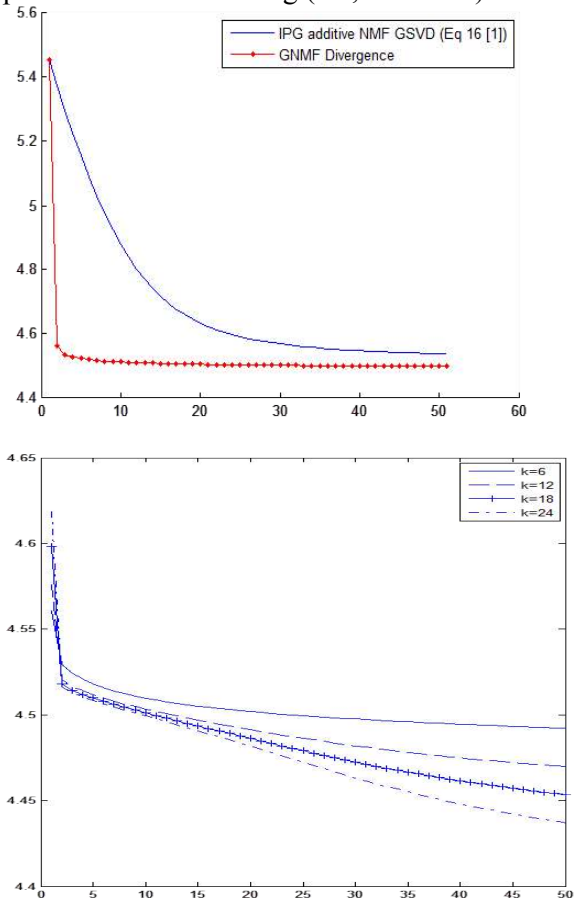

Figure 1: Estimation error using our algorithm and a benchmark method (left panel) and under various values of k (right panel) as a function of iteration step

In the second set of experiments we focused on cross lingual sentence analysis through a cross-lingual SMT vs. human translation classification task. These experiments are based on the factorization of the matrices of a corpus of 112

thousand *human* translated sentence pairs (English and Spanish) plus an additional 1 thousand held-out sentences for evaluation purposes. We generated language specific feature vectors for each sentence. These consisted exclusively of 1-gram, 2-gram and 3 gram part of speech (POS) tag sequences. For English the dimensionality of the vectors is 17 thousand (i.e., we allowed for 17 thousand different POS 3-grams in English). In the case of Spanish the dimensionality of the vectors is about 1400 because the set of tags we used is smaller.

The typical sentence in the corpus is of length 10, therefore the typical vector has a few dozen non-zero entries and the rest are zeros. We assembled 2 matrices: one using the English feature vectors (17k x 112k) and the other matrix using the Spanish (1.4k x 112k) feature vectors. We then performed the joint factorization of these matrices following our approach with k= 16.

For the test set, we kept a held out corpus, obtained both human translations (error free) and machine translation sentences (noisy observations) and generated the corresponding feature vectors for both languages. We then carried out a sentence classification experiment. For each English sentence vector in the test set we compared it versus the Spanish vector counterparts (human translation and machine translation) using the projections into the reduced rank common subspace described in section 2. Our goal is to determine which of the two vectors is closer to the reduced rank subspace representation of the corresponding original English vector using the obtained projections. Intuitively, the human translations should be closer to the English counterparts in the common subspace. The comparison is carried out as a vector classification task in the common reduced rank subspace.

Several basic classification methods in the common space were explored (single best dimension selection, LVQ classification, perceptron). The best classification was achieved using a simple perceptron rule. The accuracy using this approach is 73% which is significantly better than the results obtained using a long-span language model (cfr., Anonymized Reference). The results for the single best dimension approach and LVQ are 67% and 68.7% respectively.

6 Discussion

We have introduced a method to perform sentence similarity evaluation when such sentences are in different languages. In the experiments we conducted the features are exclusively based on n-gram sequences of POS labels. Therefore, we believe that with these experiments we have demonstrated that our technique is reasonable enough to distinguish the subtle differences existing between human and SMT output using without using word features. We expect that extending these features to include not only POS information but also other syntactic and word base information will result in further performance enhancements.

References

N. Bei, C.H.A.Koster and M.Villegas. “Cross-lingual text categorization”. In Proc. of ECDL-03, pages 126-139, 2003.

A. Cichocki, S. Amari, R. Zdunek, R. Kompass, G. Hori, Z. He, “Extended SMART Algorithms for Non-negative Matrix Factorization”, in Artificial Intelligence and Soft Computing, LNCS, Springer Verlag, 2006.

I. Dhillon, S. Sra, “Generalized nonnegative matrix approximations with Bregman divergence”, Advances in neural information processing systems, 2006 .

G. H. Golub, C. E. Van Loan, “Matrix Computations”, the Johns Hopkins University Press; 3rd edition, 1996.

T. Landauer, P. W. Foltz, D. Laham, “Introduction to Latent Semantic Analysis”, Discourse Processes 25: 259–284, 1998.

D. D. Lee, H. S. Seung, “Algorithms for non-negative matrix factorization”, in Advances in Neural Information Processing Systems 13, 2001.

C. J. Lin, “Projected Gradient Methods for Nonnegative Matrix Factorization”, Neural Computation vol. 19, 2007.

J. Mairal, F. Bach, J. Ponce, G. Sapiro, “Online Learning for Matrix Factorization and Sparse Coding”, Journal of Machine Learning Research, 2010.

R. Mihalcea., C. Corley, and C. Strapparava, C. 2006. “Corpus-based and knowledge-based measures of text semantic similarity”. In Proceedings of AAAI 2006.

X. Ni, J.-T. Sun, J. Hu, and Z. Chen. 2011. “Cross lingual text classification by mining multilingual

topics from wikipedia". In WSDM '11. ACM, New York, NY, USA,, 2011.

D. Tikk, G. Takacs, I. Pilaszy, B. Nemeth. "Investigation of Various Matrix Factorization Methods for Large Recommender Systems", 2nd WLRS 13th ACM SIGKDD. 2008.

S. Wild, J. Curry, A. Dougherty, "Motivating nonnegative matrix factorizations", SIAM Linear Algebra Meeting, 2003.

W. Xu, X. Liu, Y. Gong, "Document clustering based on nonnegative matrix factorization", In SIGIR'03, 2003.

W. Yih, K. Toutanova, J. Platt, and C. Meek, 2011. "Learning discriminative projections for text similarity measures". In CoNLL 2011.