Word Clustering using PLSA enhanced with Long Distance Bigrams

Nikoletta Bassiou and Constantine Kotropoulos
Department of Informatics, Aristotle University of Thessaloniki
Box 451, Thessaloniki 541 24, GREECE
{*nbassiou, costas*}*@aiia.csd.auth.gr*

Abstract

Probabilistic latent semantic analysis is enhanced with long distance bigram models in order to improve word clustering. The long distance bigram probabilities and the interpolated long distance bigram probabilities at varying distances within a context capture different aspects of contextual information. In addition, the baseline bigram, which incorporates trigger-pairs for various histories, is tested in the same framework. The experimental results collected on publicly available corpora (CISI, Cranfield, Medline, and NPL) demonstrate the superiority of the long distance bigrams over the baseline bigrams as well as the superiority of the interpolated long distance bigrams against the long distance bigrams and the baseline bigram with trigger-pairs in yielding more compact clusters containing less outliers.

1 Introduction

Word clustering is one of the most challenging tasks in natural language processing [5]. In this paper, word clustering based on the Probabilistic Latent Semantic Analysis (PLSA) [3] is proposed that takes into consideration long distance bigram probabilities at varying distances within a context as well as their interpolated variants and the probabilities of the baseline bigram with trigger-pairs for varying histories. The partition entropy coefficient of the derived clusterings reveals the superiority of the interpolated long distance bigrams against the long distance bigrams and the bigrams with trigger-pairs in producing more crisp clusters. In addition, the intra-cluster dispersion demonstrates that the use of interpolated long distance bigrams generates meaningful clusters, similar to those formed when the bigram model is interpolated with trigger word pairs for various histories, eliminating the cluster outliers, which are observed when long distance bigrams are used. However, clustering with trigger pairs assigns similar words into more than one clusters, and needs appropriate trigger pair selection, which is not an easy task.

2 Language Modeling and the PLSA

The n-gram model estimates the probability of a word given only the most recent $n-1$ preceding words [2]. Frequently, the bigram or the trigram models are employed only. For long distance bigrams [4], a word w_i is predicted by the d-th preceding word w_{i-d}. It is obvious that for $d=1$, the long distance bigram degenerates to the baseline bigram. The efficiency of the long distance bigram model can be further enhanced by estimating the probability of long distance bigrams in H different distances [7].

The PLSA performs a probabilistic mixture decomposition by defining a generative latent data model, the so called *aspect model*, which associates an unobserved class variable $z_k \in Z = \{z_1, z_2, \ldots, z_R\}$ with each observation. Here, the observation is simply the occurrence of a word $w_j \in V = \{w_1, w_2, \ldots, w_Q\}$ in a text/document $t_i \in T = \{t_1, t_2, \ldots, t_M\}$, while the unobserved class variable z_k models the topic a text was generated from. Summing over all possible realizations of z_k, the joint distribution of the observed data is obtained

$$P(t_i, w_j) = P(t_i) \underbrace{\sum_{k=1}^{R} P(z_k|t_i) P(w_j|z_k)}_{P(w_j|t_i)} . \quad (1)$$

As can be seen in (1), the text-specific word distributions $P(w_j|t_i)$ are obtained by a convex combination of the R aspects/factors $P(w_j|z_k)$. Representing each text t_i as a sequence of words $< v_1\ v_2\ \ldots\ v_{Q_i} >$, where Q_i is the number of words in text t_i, $P(t_i, w_j)$ can be decomposed as follows:

$$P(t_i, w_j) = P(v_{Q_i}|v_{Q_{i-1}}\ \ldots\ v_1, w_j) \cdot \\ \cdot P(v_{Q_{i-1}}|v_{Q_{i-2}}\ \ldots\ v_1, w_j)\ \ldots\ P(v_1|w_j)\ P(w_j). \quad (2)$$

Taking into consideration the long distance bigram model, (2) can be expressed as (Method I)

$$P(t_i, w_j) \simeq P(w_j) \prod_{w_l \in t_i} P_d(w_l|w_j). \quad (3)$$

where $P_d(w_l|w_j) = P(w_l|w_j, j = l - d)$. Motivated by (3) and following similar lines to the PLSA derivations, $P_d(w_l|w_j)$ can be obtained by summing over all possible realizations of z_k, i.e.

$$P_d(w_l|w_j) = \sum_{k=1}^{R} P_d(z_k|w_j)\, P_d(w_l|z_k). \quad (4)$$

By formulating the problem as maximization of the log-likelihood function with respect to the entailed probabilities, the Expectation Maximization (EM) algorithm [6] can be used, which alternates between the 1) Expectation (E)-step, where the posterior probabilities are computed for the latent variables based on the current estimates of the parameters

$$\hat{P}_d(z_k|w_j, w_l) = \frac{P_d(w_l|z_k)P_d(z_k|w_j)}{\sum_{k'=1}^{R} P_d(w_l|z_{k'})P_d(z_{k'}|w_j)} \quad (5)$$

and the 2) Maximization (M)-step, which maximizes the expected log-likelihood, computed in the previous E-step, with respect to $P_d(w_l|z_k)$ and $P_d(z_k|w_j)$ yielding the following update equations [3]:

$$P_d(w_l|z_k) = \frac{\sum_{j=1}^{Q} N_d(w_j\ w_l)\hat{P}_d(z_k|w_j, w_l)}{\sum_{l'=1}^{Q}\sum_{j=1}^{Q} N_d(w_j\ w_{l'})\hat{P}_d(z_k|w_j, w_{l'})} \quad (6)$$

$$P_d(z_k|w_j) = \frac{\sum_{l=1}^{Q} N_d(w_j\ w_l)\hat{P}_d(z_k|w_j, w_l)}{\sum_{k'=1}^{R}\sum_{l=1}^{Q} N_d(w_j\ w_l)\hat{P}_d(z_{k'}|w_j, w_l)}. \quad (7)$$

By alternating (5) with (6)-(7), a procedure that converges to a local maximum of the log-likelihood results. Each word w_j is assigned to one only cluster C_{s_j} such that $s_j = \arg\max_k P_d(z_k|w_j)\ ,\ \ j = 1, 2, \ldots, Q$.

When the PLSA employs the interpolated long distance bigrams (3) is rewritten as (Method II)

$$P(t_i, w_j) \simeq P(w_j) \prod_{w_l \in t_i} P^{(H)}(w_l|w_j). \quad (8)$$

where

$$P^{(H)}(w_l|w_j) = \sum_{k=1}^{R} \left[\sum_{d=1}^{H} \lambda_d P_d(w_l|z_k) \right] P(z_k|w_j) \quad (9)$$

and λ_d are weights for each component probability estimated on held out data by means of the EM algorithm [6]. $P_d(w_l|z_k)$ and $P(z_k|w_j)$ can be obtained by an EM algorithm which alternates between the E-step

$$\hat{P}_d(z_k|w_j, w_l) = \frac{P_d(w_l|z_k)P(z_k|w_j)}{\sum_{k'=1}^{R} P_d(w_l|z_{k'})P(z_{k'}|w_j)} \quad (10)$$

and the M-step

$$P_d(w_l|z_k) = \frac{\sum_{j=1}^{Q} N(w_j\ w_l)\hat{P}_d(z_k|w_j, w_l)}{\sum_{l'=1}^{Q}\sum_{j=1}^{Q} N(w_j\ w_{l'})\hat{P}_d(z_k|w_j, w_{l'})} \quad (11)$$

$$P(z_k|w_j) = \frac{\sum_{l=1}^{Q}\sum_{d=1}^{H} \lambda_d N(w_j\ w_l)\hat{P}_d(z_k|w_j, w_l)}{\sum_{k'=1}^{R}\sum_{l=1}^{Q}\sum_{d=1}^{H} \lambda_d N(w_j\ w_l)\hat{P}_d(z_{k'}|w_j, w_l)}. \quad (12)$$

Each word is assigned to a single cluster C_{s_j} using $s_j = \arg\max_k P(z_k|w_j)$, $j = 1, 2, \ldots, Q$.

3 Experimental Results

The word clustering algorithms that enhance PLSA with long distance bigrams (Method I) and interpolated long distance bigrams (Method II) are implemented and tested for bigram at distance $d = 1, 2, 3, 4$ or $d = 6$ and $H = 2, 3, 4$, respectively. In addition, language models combining the baseline bigram with trigger pairs [9] at various histories ($d = 2, 3, 4$ and $d = 6$) were also incorporated in the PLSA-based clustering algorithm for comparison purposes.

The experiments were conducted on four publicly available document collections[1], namely the CISI, the Cranfield, the Medline, and the NPL that include 1460, 1400, 1033, and 11429 documents, respectively. The texts have been pre-processed in order to remove any tags, non-English words, numbers or symbols with no meaning. The words were also stemmed using the Porter stemmer [8]. A vocabulary cut-off was also applied by discarding words with a frequency of appearance less than 5 for the CISI corpus and 3 for the Medline and NPL, while no vocabulary cut-off was applied to the Cranfield corpus. To derive the conditional probabilities needed for Methods I and II, first the frequencies of the distance bigrams at distances $d = 1, 2, 3, 4$ and $d = 6$ were estimated. Furthermore, the weights $\lambda_d, d = 1, 2, \ldots, H$ needed for the interpolated long distance bigrams at $H = 2, 3, 4$ were estimated by a two-way cross validation on held-out data using the EM algorithm. The predefined number R of the resulting classes was set to 250, 320, 290, and 300 for the CISI, the Cranfield, the Medline, and the NPL corpus,

[1] http://ir.dcs.gla.ac.uk/resources/test_collections/

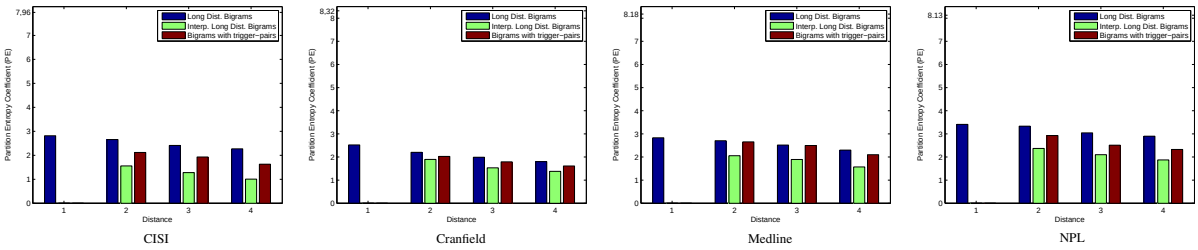

Figure 1. Partition Entropy Coefficient of the clusterings derived when the long distance bigrams, the interpolated long distance bigrams, and the bigrams with trigger-pairs are employed.

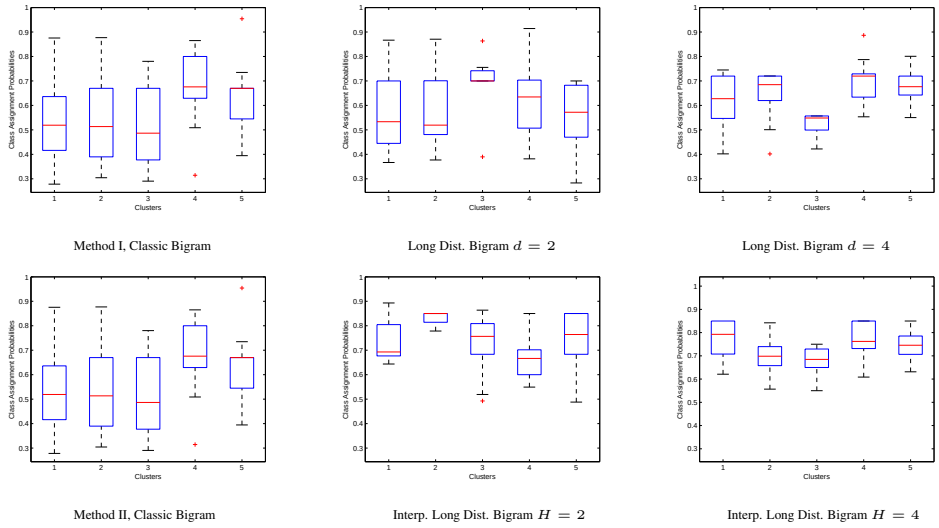

Figure 2. Intra-cluster statistics (dispersion, outliers) of the cluster assignment probabilities for sample clusters derived from the NPL corpus when Method I and Method II are applied to the baseline bigrams, the long distance bigrams at distance $d = 2$ or $d = 4$, and the interpolated long distance bigrams at distance $H = 2$ or $H = 4$.

$H = 2$	**debye**, **divide** (-ed, -er, -ers, -ing), **domain**, **equilibrium**, **filter** (-s, -ed, -ing), **fourier**, **geiger**, **harmonic** (-ically, -ics), **method** (-s), **maxima**, **medium**, **numerical** (-ically, -ous), **inverse** (-ely, -ion), **nonlinearly**, **nonuniform** (-ity), **subharmonic** (-s), **substrate**, **unitary**, **unsymmetric** (-ical, -ically)
$H = 3$	**asymmetrical** (-ically), **divide** (-ed, -er, -ers, -ing), **domain**, **equilibrium**, **filter** (-s, -ed, -ing), **fourier**, **harmonic** (-ically, -ics), **image**, **imaginary**, **inverse** (-ely, -ion), **linearly**, **logarithm** (-ic, -s), **lorenz**, **maxima**, **medium**, **method**, **minkowski**, **nonequilibrium**, **nonlinearly**, **nonuniform** (-ity), **numerical** (-ically, -ous), paid, **subharmonic** (-s), **substrate**, **unitary**, **unsymmetric** (-ical, -ically)
$H = 4$	**asymmetrical** (-ically), **divide** (-ed, -er, -ers, -ing), **domain**, **equilibrium**, **filter** (-s, -ed, -ing), **fourier**, **harmonic** (-ically, -ics), **image**, **imaginary**, **inverse** (-ely, -ion), **linearly**, **lorenz**, **maxima**, **medium**, **method**, **minkowski**, **nonequilibrium**, **numerical** (-ically, -ous), **nonlinearly**, **nonuniform** (-ity), **subharmonic** (-s), **substrate**, **unitary**, **unsymmetric** (-ical, -ically)

Table 1. Algebra-related Word clusters, that are produced by the PLSA Method II in the NPL corpus for $H = 2, 3, 4$. Boldface letters denote the word stems.

respectively. In addition, the convergence criterion for the EM algorithm of the PLSA word clustering requests the relative log-likelihood change between two successive EM-steps to be less than 10^{-4}, a condition that was satisfied after approximately 100 iterations. It is also worth mentioning that the PLSA-based algorithms have executed 10 times for each language model in order to guarantee that the results are not affected by the EM convergence to local extrema.

To select among the Q^2 possible long distance word-

pairs (trigger pairs) at the same distances as the ones employed in the long distance bigrams ($d = 2, 3, 4$ and $d = 6$), a probability threshold $p_0 = 3.5/Q$ was set. A trigger interaction between two words would thus be allowed only if the corresponding word pair probability in the bigram model were below p_0. The conditional probabilities of the extended model (i.e. bigram with trigger pairs) were then estimated by a back-off technique as described in [10].

Treating PLSA-based clustering as a "defuzzification" of a fuzzy clustering result, the partition entropy coefficient (PE) is estimated as a figure of merit for the clustering assessment [1]. The partition entropy coefficient admits values in $[0, \log_2 R]$. A small value of the partition entropy coefficient indicates the existence of a clustering structure in the data as well as the ability of the clustering technique to create efficiently hard clusters. Figure 1 plots the partition entropy coefficient for the PLSA-based clusters derived when the long distance bigrams, the interpolated long distance bigrams, and the bigrams with trigger-pairs are employed in each dataset. The partition entropy coefficient values for all datasets are small, quite closer to the lower bound than to the upper bound that is depicted in the y-axis of the plot. As it can also be seen, the partition entropy coefficient values are significantly smaller for the interpolated long distance bigrams rather than the long distance bigrams. Moreover, the PE coefficient values when trigger-pair bigrams are used are smaller than those when the long distance bigrams are used, but greater than the PE values measured when the interpolated long distance bigrams are employed.

A comparison of the clustering methods in terms of their intra-cluster dispersion is illustrated using box plots in Figure 2. Due to space limitations, results are shown for histories $d = 1, 2$ and $d = 4$ and the NPL corpus only. More precisely, Figure 2 depicts the cluster assignment probability of each word for five sample clusters derived by the clustering methods under study, when the baseline bigram model, the long distance bigrams at distance $d = 2$ or $d = 4$, and the interpolated long distance bigrams at distance $H = 2$ or $H = 4$ are used. By comparing the plots in the first row of Figure 2, it can be seen that the intra-cluster dispersion, which demonstrates the cluster compactness, is smaller when the long distance bigrams in Method I are used instead of the baseline bigrams. From the plots of the second row in Figure 2, it can be verified that Method II, which employs the interpolated long distance bigrams, reduces the outliers observed when long distance bigrams are used.

In Table 1, sample clusters from the NPL corpus produced by the clustering PLSA Method II under study are demonstrated. It has been found that although the PLSA method when applied to a bigram model extended with trigger-pairs selected from histories $d = 2, 3, 4$ and d=6 yields meaningful clusters as Method II does,it frequently splits a compact cluster related to a topic into more than one subclusters.

4 Conclusions

A technique that resorts to probabilistic latent semantic analysis is developed for word clustering. It employs long distance bigram models with and without interpolation in order to capture the long-term word dependencies with a few parameters. The validity assessment of the clustering results has demonstrated the superiority of the interpolated long distance bigrams against the long distance bigrams in producing more compact word clusters with less outliers. These clusters have also been proven to have less outliers than those produced when trigger pairs from various histories are employed in conjunction with the baseline bigram.

References

[1] J. C. Bezdek. Mathematical models for systematics and taxonomy. In G. Estabrook, editor, *Proc. 8th Int. Conf. Numerical Taxonomy*, pages 143–166, Freeman, San Francisco, CA, 1975.

[2] J. T. Goodman. A bit of progress in language modeling. *Computer Speech and Language*, 15(4):403–434, October 2001.

[3] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1-2):177–196, 2001.

[4] X. Huang, F. Alleva, H. Hon, M. Hwang, K. Lee, and R. Rosenfeld. The SPHINX-II speech recognition system: An overview. *Computer Speech and Language*, 2:137–148, 1993.

[5] A. K. Jain and R. C. Dubes. *Algorithms for Clustering Data*. Prentice Hall, Englewood Cliffs, N.J., 1988.

[6] F. Jelinek. *Statistical Methods for Speech Recognition*. MIT Press, Cambridge, MA, 1998.

[7] N. Bassiou and C. Kotropoulos. Interpolated Distanced Bigram Language Models for Robust Word Clustering. In *Proc. Nonlinear Signal and Image Processing*, pages 12–15, Sapporo, Japan, May 2005.

[8] M. F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, July 1980.

[9] R. Rosenfeld. *Adaptive Statistical Language Modeling: A Maximum Entropy Approach*. Ph.D. Thesis, Carnegie Mellon University, School of Computer Science, Pittsburgh, PA, April 1994.

[10] C. Tillmann and H. Ney. Word trigger and the EM algorithm. In *Proc. Workshop Computational Natural Language Learning*, pages 117–124, 1997.