# GloDETC : Global Document Embeddings for Multi-label Text Classification

*A Thesis Submitted*
*in Partial Fulfilment of the Requirements*
*for the Degree of*

**B.Tech. - M.Tech. (Dual Degree)**

*by*
**Nitish Gupta**
**Roll No. : 10327461**

*under the guidance of*
**Prof. Harish Karnick**
**Prof. Rajesh M. Hegde**


Department of Electrical Engineering
Indian Institute of Technology Kanpur
April, 2015

# CERTIFICATE

It is certified that the work contained in this thesis entitled "*Merging Word Senses*", by *Sumit Bhagwani(Roll No. Y8127515)*, has been carried out under my supervision and that this work has not been submitted elsewhere for a degree.

(Prof. Harish Karnick)
Department of Computer Science and Engineering,
Indian Institute of Technology Kanpur
Kanpur-208016

June, 2013

# Abstract

Currently used general purpose dictionaries are often too fine-grained, with narrow sense divisions that are not relevant for many Natural Language applications. WordNet, which is a widely used sense inventory for Word Sense Disambiguation, has the same problem. With different applications requiring different levels of sense granularities, producing sense clustered inventories of arbitrary sense granularity has evolved as a crucial task.

We try to exploit the resources available like human-labelled sense clusterings and semi-automatically generated domain labels of synsets, to estimate the similarity between synsets. Using supervision, we learn a model which predicts the probability of any two senses of a word to be merged. To learn a more generic model, we propose a graph based approach, which allows us to use the information learnt from supervision as well. Using this complete similarity measure, we propose a simple method for clustering synsets. We show that the coarse-grained sense inventory obtained significantly boosts the disambiguation of nouns on standard test sets.

*Dedicated to*
*My Family*

# Acknowledgement

I wouldn't like to express my sincere gratitude towards my thesis supervisor, Prof. Harish Karnick, for his constant support and encouragement. I am grateful for his patient guidance and advice in giving a proper direction to my efforts. I am also grateful to Prof. Rajesh M. Hegde for giving me the freedom to work on a topic of my liking.

I am fortunate for the bevy of my friends I am a part of - I would like to thank them immensely, especially Aditya, Abhinav, Shubhendu, Parul, Pranay, Arzoo, Saurabh, Prajyoti, Asheesh and Anjana for their pleasant company and their ability to infuse humor into any situation. I would particularly like to thank Shrutiranjan Satapathy for his constant support, invaluable discussions and needless coffee breaks.

Last, but not the least, I would like to thank my parents and brother for their love and encouragement. Without their support and patience this work would not have been possible.

***Nitish Gupta***

# Contents

# List of Tables

# List of Figures

# List of Algorithms

# Chapter 1

# Related Work

The task of text classification, i.e. classification of documents into a fixed number of predefined categories has been long studied in-depth for many years now. This multi-class classification problem has further evolved into a multi-label text classification task where each document can belong to multiple, exactly one or no category at all.

Supervised machine learning techniques that learn classifiers to perform this category assignment task can be broken down into two main components, namely, text representation and learning algorithm. Text representation involves converting the documents, that are usually strings of characters, into numerical vectors that are suitable inputs to the learning algorithm while the learning algorithm uses pairs of labeled input text representations and the categories it is belongs in, to learn a model so as to classify new documents into categories.

## 1.1 Text Representation

Any text-based classification system requires the documents to be represented in an appropriate manner dictacted by the task being performed [Lewis, 1992]. Moreover, [Quinlan, 1983] showed that the accuracy of the classification task depends as much on the document representation as on the learning algorithm being employed. Different from the data mining task, which deals with structureed documents, text classification deals with unstructured documents that need to be appropriately trans-

formed into numerical vectors, i.e. the need for text representation. In this section we introduce the most effective and widely-used techniques to represent documents for text classification.

### 1.1.1 Bag of Words

It is found in information retrieval research that word stems work well as representations units for documents and that their ordering in a document is of minor importance for many tasks. This is attributed by the fact that the most widely-used used model to represent documents for the classification task is the *Vector Space Model (VSM)* [Salton and Yang, 1973].

In the Vector Space Model, a document $d$ is represented as a vector in the term/word space, $d = (w_1, w_2, \ldots, w_{|V|})$ where $|V|$ is the size of the vocabulary. Each of the $w_i \in [0, 1]$, represents the weightage of the term $i$ in the document $d$. This is called the *bag-of-words* model as it ignores word ordering and each document is reduced to a bag of words that it contains or not.

An important requirement of such a representation is that, the terms that help in defining the semantic content of the document and play an important role in classification be given higher weightage than the others. Over the years, there has been much research in the information retrieval field on term weighting schemes. The most important term-weighting techniques are described below :

1. **One Hot Representation** : This is the most trivial representation, where each document is represented by a vector that is size of the vocabulary. Each element in the vector is either a 0 or a 1 to denote the absence or presence of a specific term in the document.

2. **Term Frequency (tf))** : The term frequency representation weighs the terms present in the document relative to their occurence frequency in the document. Hence a document $d$ is represented as, $d = (w_1, w_2, \ldots, w_{|V|})$, where, $w_k$ is the number of times the term $k$ appears in the document $d$.

3. **Inverse Document Frequency (idf)** : Though using *tf* as a term weighting scheme is a good starting point, it faces a challenge when high frequency terms are not concentrated in a few particular documents but are prevalent in the whole collection. Those terms then stop being characteristic of the semantic content of a few documents and need not be given high weightage. To overcome this problem, Salton and Buckley [1988] suggested a new term weighting called the inverse document frequency (idf). The *idf* weight of a term varies inversely with the number of documents $n$ it belongs to in a collection of total $N$ documents. A typical *idf* vector can be computed as

$$w_k = \log \frac{N}{n} \tag{1.1}$$

4. **Term Frequency Inverse Document Frequency (tf-idf)** : Given the above two term weighing schemes, it is clear that an important term in a document should have high *tf* but a low overall collection frequency (*idf*). This suggests that a reasonable measure for term importance may be then obtained by the *tf* and the *idf* ($tf \times idf$). As we will see in the results section, the *tf-idf* weighed bag-of-words document representation gives one of the best accuracies in the multi-label text classification task.

A common feature in the bag-of-words document representation is the *normalization factor*[Salton and Buckley, 1988] introduced to reduce the effect of varying document lengths and give equal weightage to documents of all lengths when learning the classifier for text categorization. **TODO:** Do we put how normalization is done? Another feature added to the bag-of-words representation is the removal of stop-words (short function words that do not add to the semantic content of the document) and words that occur infrequently to make the document vector more meaningful.

### 1.1.2 Dimensionality Reduction / Feature Selection

The bag-of-words representation scheme has several drawbacks but the most important drawback it suffers from is that document vectors are very sparse and high dimensional. Typical vocabulary sizes of a moderate-sized document collection ranges from tens to hundereds of thousands of terms which is prohibitively high for many learning algorithms. To overcome this issue of high-dimensional bag-of-words document representations, automatic feature selection is performed that removes uninformative terms according to corpus statistics and constructs new orthogonal features by combining several lower level features (terms/words). Several techniques used in practice are discussed below,

1. **Information Gain** : Information Gain is widely used as a term-goodness criterion in the field of machine learning, mainly in decision trees [Quinlan, 1986] and also in text classification [Lewis and Ringuette, 1994], [Moulinier et al., 1996]. It is a feature space pruning technique that measures the number of bits of information obtained(entropy) for category prediction by knowing the presence or absence of a term in a document. For terms where the information gain was below some predefined threshold are not considered in the document vector representation. The information gain of a term $t$ is defined as

$$G(t) = -\sum_{i=1}^{|C|} P(c_i) \log P(c_i) + P(t) \sum_{i=1}^{|C|} P(c_i|t) \log P(c_i|t) + P(\,t) \sum_{i=1}^{|C|} P(c_i|\,t) \log P(c_i|\,t) \tag{1.2}$$

2. **Mutual Information** : Similar to the Information Gain scheme, Mutual Information estimates the information shared between a term and a category and prunes terms that are below a specific threshold. The mutual information between a term $t$ and a category $c$ is estimated in the following fashion,

$$I(t, c) = \log \frac{P(t \wedge c)}{P(t) \times P(c)} \tag{1.3}$$

To measure the goodness of a term in global feature selection, the category specific scores of a term are combined using,

$$I_{avg}(t) = \sum_{i=1}^{|C|} P(c_i) I(t, c_i) \tag{1.4}$$

3. $\chi^2$ **Statistic** : The $\chi^2$ statistic measures the lack of independence a term $t$ and a category $c$ and can be compared to the $\chi^2$ distribution with one degree of freedom. The term-goodness factor is calculated for each term-category pair and is averaged as above. The major difference between Mutual Information and $\chi^2$ statistic is that the later is a normalized value and the goodness factors across terms are comparable for the same category.

4. **Latent Semantic Indexing (LSI)** : LSI first introduced by Deerwester et al. [1990], is a popular linear algebraic dimensionality reduction technique that uses the term co-occurence statistics to capture the latent semantic structure of the documents and represent them using low-dimensional vectors. It is an efficient technique to deal with synonymy and polysemy. LSI aims to find the best subspace approximation to the original document bag-of-word vector space using Singular Value Decomposition. Given a term-document matrix $X = \left[x_1, x_2, \ldots, x_{|D|}\right] \in \mathbb{R}^{|V|}$, its k-rank approximation as found using SVD, can be expressed as,

$$X = TSD^T \tag{1.5}$$

where, $T \in \mathbb{R}^{|V| \times k}$ and $D \in \mathbb{R}^{|D| \times k}$ are orthonormal matrices called the left and right singular vectors respectively. The matrix $S \in \mathbb{R}^{k \times k}$ is a diagonal matrix of singular values arranged in descending order. The $k$-dimensional rows of the matrix $D$ contain the dimensionality reduced representations of the $|D|$ documents in the collection. The representations obtained using LSI alleviate the issue of data sparsity and high-dimensionality in bag-of-words representations and also helps unfold the latent semantic structure of the documents.

## 1.2 Learning Algorithms

Multi-label text classification has seen growing number of statistical learning methods being applied to it. Over the years, various larning algorithms like, Regression models ([Cooper et al., 1994], [Fuhr et al., 1991]), Conditional Random Field ([Ghamrawi and McCallum, 2005]), Nearest Neighbour techniques ([Yang, 1994], [Zhang and Zhou, 2005], [Zhang and Zhou, 2007]), Bayesian classifier and topic modelling ([Lewis and Ringuette, 1994], [McCallum, 1999], [Nigam et al., 2000], [Rubin et al., 2012], [Nigam et al., 1999], [Ueda and Saito, 2002]), SVM ([Joachims, 1998], [Elisseeff and Weston, 2001]), Neural Networks ([Wiener et al., 1995], [Ng et al., 1997]), Decision Trees ([Tong and Appelbaum, 1994]), Online learning algorithms ([Lewis et al., 1996], [Crammer and Singer, 2002]), Non-negative Matrix Factorization ([Liu et al., 2006]) etc. have been used or developed for Multi-label document categorization.

Earlier learning algorithms reduced the problem of multi-label classification into multiple binary classification problems and independently learned binary classifiers for each category. While these algorithms performed well, their drawback of considering correlation among categories led to the development of algorithms that learn a single classifier and jointly classify each document.

Multi-label classification problems can be also be classified into classification-based and ranking-based approaches, where the former assigns each test instance a $|L|$-sized label vector of ones and zeros indicating the presence and absence of labels. In the case of a ranking-based approach, the ranking system outputs the list of labels arranged in the increasing order of a ranking score which is then thresholded at an optimum and the top labels are considered appropriate label assigments for test instances.

Below we describe some of the famous learning algorithms for multi-label text classification,

### 1.2.1 With Multiple Binary Classifiers

The most common approach of multi-label text classification, treats each label independently and learns multiple binary classifiers, one for each category and then assigns to a test document all the categories for which the corresponding classfier says *'yes'*. In this section we describe some of the algorithms in the context of multi-label text classification.

1. **Logistic Regression (LR)** : Introduced by [Hosmer and Lemeshow, 1989], LR is a probabilistic binary classification regression model, that, for binary text classification learns a category weight vector and estimates the probability of a document belonging to the category using dot-product and the logistic link function. LR can be extended for multi-label document classification by learning multiple category vectors, specifically, one for each category. At test time, one would need to query all category vectors for each document to make the category assignments. In our work, we use logistic regression for multi-label text classification. The details for the model are given in **TODO:** Future link to description.

2. **Support Vector Machines (SVM)** : Support Vector Machines ([Cortes and Vapnik, 1995], [Vapnik, 2000]) based on the *Structural Risk Minimization* principle, are universal learners. In their basic form, SVMs learn linear threshold functions to find linear hyperplanes in the input data space to separate data of the two differnt classes. In the case, where data is not linearly separable, SVMs can be plugged-in with appropriate kernel functions to learn polyniomial classifiers, radial basic fucntions etc. For multi-label text classification, training data is treated separately for each category and maximum margin separating hyperplanes are found for each category independently [Joachims, 1998].

   Elisseeff and Weston [2001] study a ranking based variant of SVM, where the positive/negative distance from the separating hyperplane of a specific category is the score assigned to the particular instance for that category.

Their formulation then aims to maximize the margin between the score of a category that belongs to the document and a category that does not belong to do the document. This is also called the Rank-SVM.

3. :

# Bibliography


W Cooper, Aitao Chen, and F Gey. Full text retrieval based on probabilistic equations with coefficients fitted by logistic regression. *NIST SPECIAL PUBLICATION SP*, pages 57–57, 1994.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

Koby Crammer and Yoram Singer. A new family of online algorithms for category ranking. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 151–158. ACM, 2002.

Scott C. Deerwester, Susan T Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by latent semantic analysis. *JAsIs*, 41(6): 391–407, 1990.

André Elisseeff and Jason Weston. A kernel method for multi-labelled classification. In *Advances in neural information processing systems*, pages 681–687, 2001.

Norbert Fuhr, Stephan Hartmann, Gerhard Lustig, Michael Schwantner, Kostas Tzeras, and Gerhard Knorz. *AIR, X: a rule based multistage indexing system for large subject fields*. Citeseer, 1991.

Nadia Ghamrawi and Andrew McCallum. Collective multi-label classification. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 195–200. ACM, 2005.

David W Hosmer and Stanley Lemeshow. Applied logistic regression. 1989. *New York: Johns Wiley & Sons*, 1989.

Thorsten Joachims. *Text categorization with support vector machines: Learning with many relevant features*. Springer, 1998.

David D Lewis. Text representation for intelligent text retrieval: a classification-oriented view. *Text-based intelligent systems: current research and practice in information extraction and retrieval*, pages 179–197, 1992.

David D Lewis and Marc Ringuette. A comparison of two learning algorithms for text categorization. In *Third annual symposium on document analysis and information retrieval*, volume 33, pages 81–93, 1994.

David D Lewis, Robert E Schapire, James P Callan, and Ron Papka. Training algorithms for linear text classifiers. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 298–306. ACM, 1996.

Yi Liu, Rong Jin, and Liu Yang. Semi-supervised multi-label learning by constrained non-negative matrix factorization. In *Proceedings of the national conference on artificial intelligence*, volume 21, page 421. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006.

Andrew McCallum. Multi-label text classification with a mixture model trained by em. 1999.

Isabelle Moulinier, Gailius Raskinis, and J Ganascia. Text categorization: a symbolic approach. In *proceedings of the fifth annual symposium on document analysis and information retrieval*, pages 87–99, 1996.

Hwee Tou Ng, Wei Boon Goh, and Kok Leong Low. Feature selection, perceptron learning, and a usability case study for text categorization. In *ACM SIGIR Forum*, volume 31, pages 67–73. ACM, 1997.

Kamal Nigam, John Lafferty, and Andrew McCallum. Using maximum entropy for text classification. In *IJCAI-99 workshop on machine learning for information filtering*, volume 1, pages 61–67, 1999.

Kamal Nigam, Andrew Kachites McCallum, Sebastian Thrun, and Tom Mitchell. Text classification from labeled and unlabeled documents using em. *Machine learning*, 39(2-3):103–134, 2000.

J Ross Quinlan. Learning efficient classification procedures and their application to chess end games. In *Machine learning*, pages 463–482. Springer, 1983.

J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

Timothy N Rubin, America Chambers, Padhraic Smyth, and Mark Steyvers. Statistical topic models for multi-label document classification. *Machine Learning*, 88(1-2):157–208, 2012.

Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988.

Gerard Salton and Chung-Shu Yang. On the specification of term values in automatic indexing. *Journal of documentation*, 29(4):351–372, 1973.

Richard M Tong and Lee A Appelbaum. Machine learning for knowledge-based document routing (a report on the trec-2 experiment). *NIST SPECIAL PUBLICATION SP*, pages 253–253, 1994.

Naonori Ueda and Kazumi Saito. Parametric mixture models for multi-labeled text. In *Advances in neural information processing systems*, pages 721–728, 2002.

Vladimir Vapnik. *The nature of statistical learning theory*. Springer Science & Business Media, 2000.

Erik Wiener, Jan O Pedersen, Andreas S Weigend, et al. A neural network approach to topic spotting. In *Proceedings of SDAIR-95, 4th annual symposium on document analysis and information retrieval*, pages 317–332. Citeseer, 1995.

Yiming Yang. Expert network: Effective and efficient learning from human decisions in text categorization and retrieval. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 13–22. Springer-Verlag New York, Inc., 1994.

Min-Ling Zhang and Zhi-Hua Zhou. A k-nearest neighbor based algorithm for multi-label classification. In *Granular Computing, 2005 IEEE International Conference on*, volume 2, pages 718–721. IEEE, 2005.

Min-Ling Zhang and Zhi-Hua Zhou. Ml-knn: A lazy learning approach to multi-label learning. *Pattern recognition*, 40(7):2038–2048, 2007.