Graph Based Semi-Supervised Learning for Tamil POS Tagging

Mokanarangan Thayaparan, Surangika Ranathunga, Uthayasanker Thayasivam
Dept. of Computer Science and Engineering
University of Moratuwa, Katubedda 10400, Sri Lanka
{mokanarangan.11, surangika, rtuthaya }@cse.mrt.ac.lk

Abstract

Parts of Speech (POS) tagging is an important pre-requisite for various Natural Language Processing tasks. POS tagging is rather challenging for morphologically rich languages such as Tamil. Being low-resourced, Tamil does not have a large POS annotated corpus to build good quality POS taggers using supervised machine learning techniques. In order to gain the maximum out of the existing Tamil POS tagged corpora, we have developed a graph-based semi-supervised learning approach to classify unlabelled data by exploiting a small sized POS labelled data set. In this approach, both labelled and unlabelled data are converted to vectors using word embeddings and a weighted graph is constructed using Mahalanobis distance. Then semi-supervised learning (SSL) algorithms are used to classify the unlabelled data. We were able to gain an accuracy of 0.8743 over an accuracy of 0.7333 produced by a CRF tagger for the same limited size corpus.

Keywords: Semi-Supervised Learning, Low-resourced languages, Graph-based SSL, Word Embedding, POS tagging

1. Introduction

In the recent past, supervised learning methods have produced high accuracies for Parts-of-Speech (POS) tagging (Gimenez and Marquez, 2004). In particular, sequence models such as hidden Markov models (HMM) and conditional random fields (CRF) have given good results (Huang et al., 2015). However, these techniques rely on the availability of relatively large amounts of annotated data. Hence, building an accurate domain insensitive POS tagger is challenging for low resourced languages.

Tamil is one such low resourced language, which is widely used in South India and Sri Lanka. There have been several POS taggers developed for Tamil language using supervised learning techniques (Dhanalakshmi et al., 2009)(Pandian and Geetha, 2009). Since the annotated corpora used in this research have been of small size and from a single domain, these supervised techniques greatly suffer from accuracy and domain adaptability (Rani et al., 2016). For example, FIRE corpus (Forum for Information Retrieval Evaluation, 2014), a widely used freely available Tamil POS annotated corpus contains only 80k words. In contrast, the Wall Street corpus, which is an English POS-annotated corpus has a word count of 1,173K words (Gimenez and Marquez, 2004), meaning that the size of the FIRE corpus is approximately 15 times smaller than the Wall Street corpus. Thus, when using a small corpus such as FIRE, we cannot expect similar accuracy to that of English when supervised techniques are used. Moreover, these approaches depend on language dependent features such as morphological tags (Dhanalakshmi et al., 2009) thus limiting the scalability for adapting to other low resourced languages.

In contrast to supervised approaches, semi-supervised approaches such as graph based semi-supervised learning and manifold regularization (Niyogi, 2013) use both labeled and unlabelled data for their classification, and have proven to work with a small data sets (Zhu et al., 2003). Despite having smaller sized POS-tagged data for Tamil, there has been only two research leveraging the opportunity presented by semi-supervised learning. Ganesh et al. (2014) have used segmentation patterns to implement a bootstrapping approach for POS tagging. This approach relies on language dependent data such as suffix context patterns. Rani et al. (2016) use small annotated training data to build a classifier model using context-based association rule mining. This approach neither includes any language-specific linguistic information nor requires a large corpus. However, they collect all possible words occurring in the same context from the untagged data into a list called context-based list, thus limiting it from scaling to large monolingual corpus.

Graph based semi-supervised learning (SSL) has gained traction in Natural Language Processing tasks such as question answering (Celikyilmaz et al., 2009), structural tagging (Subramanya et al., 2010), and speech language recognition (Liu et al., 2016). Graph based SSL builds a meaningful graph using labelled and unlabelled instances. It then employs an SSL algorithm such as harmonic functions (Zhu et al., 2005) or label propagation (Zhu et al., 2003) to label the unlabelled instances. Graph based SSL is easily parallelizable and scalable to large data (Zhu et al., 2005).

In this paper, we present a novel graph-based semi-supervised approach to produce an accurate POS tagger for Tamil using a limited size corpus. Our idea is inspired by Talukdar and Pereira (2010)'s case study on modified absorption, which is a label propagation algorithm. They have implemented a Named Entity recognizer by building a connected word graph. Similarity between words is measured using WordNet. Then they employ label propagation to assign labels to all the unlabelled nodes.

Since Tamil is a low resourced language with no proper WordNet, we built a connected word graph using word vectors and employed label propagation. Our method is based on the clustering hypothesis that relative distance of word vectors of similar categories is lower than those between different categories. We use neural word embedding (Word2Vec (Mikolov et al., 2013), FastText (Joulin et al., 2016)) to create word vectors. Mahalanobis distance is used for measuring the distance (metric learning) between these vectors in order to construct the graph. Mahalanobis distance generalizes the standard Euclidean distance, and has proven to be more effective (Davis et al., 2007). We empirically tested with four different metric learning algorithms (Information Theoretic Metric Learning (ITML) (Davis et al., 2007), Sparse Determinant Metric Learning (SDML) (Qi et al., 2009), Least Squares Metric Learning (LSML) (Liu et al., 2012), and Local Fisher Discriminant Analysis (LFDA) (Sugiyama, 2006)) to calculate Mahalanobis distance. Once the graph is constructed with labeled and unlabeled nodes, to assign labels to unlabeled nodes, we experimented with three different SSL algorithms (LP-ZGL) (Zhu et al., 2003), Absorption (Talukdar et al., 2008) and Modified Absorption (MAD) (Talukdar and Pereira, 2010)). Local Fisher Discriminant Analysis (LFDA) metric learning coupled with Label Propagation(LP-ZGL) yielded a maximum accuracy of 0.8743 for the FIRE corpus against a baseline accuracy of 0.7338 achieved by using a traditional CRF model. Unlike supervised learning approaches, our approach does not require a large high quality annotated data set, or language dependent features.

Thus the contributions of this paper are: (1) converting words to vectors using neural word embedding and building meaningful word graphs, (2) using Mahalanobis distance to measure relationships between word vectors, hence measuring the correlation between variables, and (3) using a language independent graph based semi-supervised approach for POS tagging in Tamil.

The rest of the paper is organized as follows. Section 2 discusses graph based semi supervised learning techniques and previous attempts on Tamil POS tagging. Section 3 details the data set used in our experiment. Section 4 discusses the methodology and how we implemented the system. Section 5 details the experiments carried out and the relevant results. Section 6 and Section 7 document the conclusion and future work, respectively.

2. Related Work

2.1. Graph based Semi-supervised Learning

Graph theory and Natural Language Processing are well studied disciplines, but are commonly perceived as distinct with different algorithms and with different applications. But recent research has shown that these disciplines are connected and graph-theoretical approaches can be employed to find efficient solutions for NLP problems. Entities are connected by a range of relations in many NLP problems and graph is a natural way to capture the relationship between the entities. Graph based approaches have been used in word sense disambiguation, entity disambiguation, thesaurus construction, textual entailment and semantic classification (Mihalcea and Radev, 2011).

Graph based semi-supervised learning builds graphs connecting labeled and unlabeled data points, and perform classification by propagating the labels. The graph is constructed to reflect our prior knowledge about the domain. The intuition is that similar data points have similar labels. We let the hidden/observed labels be random variables on the nodes of this graph. Labels are injected to unlabeled nodes from labeled nodes. Graphs provide a uniform representation for heterogeneous data and are easily parallelizable (Zhu et al., 2005).

One of the challenges of graph based approach is building the graph that reflects the relationship between entities. Depending on the task, the nodes and edges may represent a variety of language related units and links. Different NLP tasks have approached this challenge in different ways. For the task of opinion summarization, Zhu et al. (2013) constructed a graph of sentences linked by edges whose weight combines the term similarity and objective orientation similarity. And to perform discourse analysis in chat, Elsner and Charniak (2010) predicted the probabilities for pair of utterance as belonging the same conversation thread or not based on lexical, timing and discourse-based features. Then constructed a graph with each nodes representing the utterances and the edges representing the probability score between the nodes. Although these approaches are evidences for the versatility of graph based approaches, these cannot be adopted to a word level problem like sequential tagging. Using graph methods for sequential tagging relies on the belief that similar words will have the same tag. Unlike the aforementioned approaches, here the nodes in these graph represents words or phrases and the the edges will indicate the similarity between nodes. Talukdar and Pereira (2010) tag words with NER information through a label propagation algorithm on a word similarity graph built using WordNet information. Words are represented are the graph vertices and the edge denotes the WordNet relationship. This approach cannot be adopted for a low resource language which doesn't have a proper WordNet.Subramanya et al. (2010) POS tags on a similarity graph where local sequence contexts (n-grams) are vertices. The similarity function between graphs is the cosine distance between the point-wise mutual information vectors (PMI) representing each node. The point-wise mutual information is calculated between n-gram and set of context features. These context features includes suffixes, left word and right word contexts. The challenge of this approach is the scalability for a morphologically complex language like Tamil.

2.2. Tamil POS tagging

Tamil is a low resourced, morphologically rich language with many inflections and a complex grammatical structure. Thus, automatic POS tagging for Tamil is a challenging task. Supervised learning approaches have been heavily undertaken in Tamil for POS tagging. These include CRF models using morphological information (Pandian and Geetha, 2009) and Support Vector Machines (SVM) using semantic information (Dhanalakshmi et al., 2009). These models had been trained using different corpora containing approximately 200k annotated words. These annotated corpora or taggers are not publicly available.

There have been very few attempts in using semi-supervised approaches for Tamil language to develop POS taggers. Ganesh et al. (2014) have used language features with a bootstrapping approach to obtain a precision of 86.74%. They have presented a pattern based bootstrapping approach using only a small set of POS labelled suffix context patterns. The patterns consist of a stem and a sequence of suffixes, obtained by segmentation using a manually created suffix list. This bootstrapping technique generates new patterns by iteratively masking suffixes with low probability of occurrences in the suffix context, and replacing them with other co-occurring suffixes. This approach relies on language specific information.

Rani et al. (2016) have employed a semi-supervised rule mining approach using morphological features for Hindi, Tamil, and Telugu languages. They have used a combination of a small annotated and untagged training data to build a classifier model using a concept of context-based association rule mining. These association rules work as context-based tagging rules.

3. Data set

For our experiment, we used the FIRE Tamil Corpus. The FIRE Tamil corpus contains 80k POS tagged words with 21 different tags as shown in Table 1.

NN	Noun
NNC	Compound Noun
RB	Adverb
VM	Verb Main
SYM	Symbol
PRP	Personal Pronoun
JJ	Adjective
NNP	Pronoun
PSP	Prepositions
QC	Quantity Count
VAUX	Verb Auxiliary
DEM	Determiners
QF	Quantifiers
NEG	Negatives
QO	Quantity Order
WQ	Word Question
INTF	Intensifier
NNPC	Compound Pro Noun
CC	Coordinating Conjunction
RBP	Adverb Phrase

Table 1: POS tagsets for FIRE Tamil Corpus

4. Methodology

Our work is inspired by Talukdar and Pereira (2010)'s case study on the performance of different algorithms for classification in graphs. In this work, words are represented as nodes and the similarity between nodes are measured using WordNet distance. Since Tamil is a low resourced language, this approach was not viable for us. Another approach was to represent words by converting them to vectors and computing the similarity. Subramanya et al. (2010) had employed a point wise mutual information (PMI) based approach to convert the word to vectors and compute the similarity by measuring the cosine distance. His approach used hand-crafted features that will not work with same efficiency across different languages.

Hence, an efficient way of representing a word in the vector space has to be determined. In addition, it is required to identify mechanisms for (1) constructing a meaningful graph based on the word vector, and (2) classifying unlabelled words based on the constructed graph by measuring the similarity.

4.1. Representing a word in the vector space

We adopted the Word2Vec model proposed by Mikolov et al. (2013) and convert the word into the vector space to construct the graph. To the best of our knowledge, Word2Vec has never been used to construct weighted word graphs to be used in SSL. Similarly we also experimented with Fast Text skipgram (Bojanowski et al., 2016) and bag of words models (Joulin et al., 2016). The key difference between Word2Vec and FastText is that Word2Vec treats each word in corpus as an atomic entity and generates a vector for each word. In contrast, FastText treats each word as composed of ngrams and the vector word is made of the sum of these vectors.

4.2. Constructing a meaningful graph based on the word vector

Each word is converted to a d dimensional vector space. Out of the n words in the list, n_l are labelled($n >>> n_l$). We employ 32 different tags to denote each POS entity (Dhanalakshmi et al., 2009). $G = (V, E, W)$ is the graph we are interested in constructing; where V is the set of vertices with $|V| = n$, E is the set of edges. W is the symmetric $n \times n$ matrix of edge weights we want to learn. Usually we could choose a standard distance metric (Euclidean, City-Block, Cosine, etc.). Instead, Mahalanobis distance has proven to be effective with clustering problems over the standard metrics (De Maesschalck et al., 2000).

We use a supervised method for learning the Mahalanobis distance. For this purpose, we need to calculate the positive definite matrix A of size $d \times n$ that parametrizes the Mahalanobis distance, $d_A(x_i, x_j)$ (Dhillon et al., 2010; Davis et al., 2007; Sugiyama, 2006) between words x_i and x_j as shown in Equation (1).

$$d_A(x_i, x_j) = (x_i - xj)^T A(x_i - x_j) \qquad (1)$$

Since A is positive definite, it can be decomposed into $P^T P$, where P is another matrix of size $d \times d$

$$\begin{aligned} d_A(x_i, x_j) &= (x_i - x_j)^T P^T P(x_i - x_j) \\ &= (Px_i - Px_j)^T (Px_i - Px_j) \qquad (2) \\ &= d_I(Px_i, Px_j) \end{aligned}$$

There are many proposed methods for calculating the transformation matrix P. We empirically experimented with different metric learning algorithms, including Information Theoretic Metric Learning (ITML) (Davis et al., 2007), Sparse Determinant Metric Learning (SDML) (Qi et al., 2009), Least Squares Metric Learning (LSML) (Liu et al., 2012), and Local Fisher Discriminant Analysis (LFDA) (Sugiyama, 2006).Researches in link prediction in networks (Shaw et al., 2011), music recommendation (McFee et al., 2011) and bio metrics verification (Ben et al., 2012) has shown that metric learning plays a vital role increasing accuracy of the system.

ITML minimizes the differential entropy between multivariate Gaussian under constraints on the distance function. Davis et al. (2007) have expressed the problem as that of minimizing the LogDet divergence subject to linear constraints. SDML uses l_1-penalized log-determinant regularization to calculate the metric. This algorithm exploits the sparsity nature underlying the intrinsic high dimensional feature space. LSML uses an algorithm that minimizes a convex objective function corresponding to the sum of squared residuals of constraints. Finally LFDA, is a linear supervised dimensionality reduction method which is particularly useful when dealing with cases where one or more core classes consist of separate clusters in input space.

We calculate P using each of these metric learning algorithms and project the words into a new space to calculate Px_i. Based on Equation 2, we compute the Euclidean distance in the linearly transformed matrix. Gaussian kernel $[2, 16]$ was used to compute the similarity between words as shown in Equation 3 (Dhillon et al., 2010). We then sparsify the graph by selecting k neighbors for each node and set the weights to zero for all others (Zhu et al., 2003).

$$W_{ij} = exp(\frac{-d_A(x_i, x_j)}{2\sigma^2}) \qquad (3)$$

The culmination of all these steps results in a meaningful graph where relative distances of word vectors of similar categories will be lower than those between different categories.

4.3. Classifying Unlabelled Nodes based on the Constructed Graph

Once the graph is constructed, unlabelled words in the graph should be classified. For this, we experimented with Label Propagation(LP-ZGL), and Absorption and Modified Absorption (MAD) techniques. LP-ZGL (Zhu et al., 2003) was one of the first graph based SSL methods. LP-ZGL propagates the labels over the graph by penalizing any label assignment where two nodes connected by a highly weighted edge are assigned different labels. LP-ZGL prefers smooth labeling over the graph. This property is also shared by the other two algorithms. Absorption (Talukdar et al., 2008) has been used for open domain class-instance acquisition. Absorption is an iterative algorithm where label estimates depend on the previous iteration. Modified Absorption (MAD) (Talukdar and Pereira, 2010) shares the same properties of the Absorption algorithm but can be expressed as an unconstrained optimization problem. We experimented with all these algorithms to estimate the labels of the untagged words.

5. Experiments and Results

5.1. Experiments

We split the data into 60k words for training and 20k words for testing. To the best of our knowledge, there has been only Named Entity Recognition research (Abinaya et al., 2014) done in Tamil using FIRE corpus and no POS tagging research done.

We trained both Word2Vec and FastText models with a word window of three (the commonly used window size) using the Tamil Wikipedia corpus (Wikipedia, 2016) (about 1M words) after removing only the punctuation marks. We used these models to convert word to vector form. Each vector is of 300 dimensions. For graph construction, a subset of 3000 sentences with approximately 50k unlabelled words from the Tamil Wikipedia corpus were added to the set. We constructed the word graphs using the aforementioned four metric learning approaches and employed three labeled propagation approaches to identify the best combination.

Since most of the successful approaches related to Tamil POS tagging have been carried out using Conditional Random Fields (CRF) (Pandian and Geetha, 2009), we used the same approach with word trigram feature as our baseline method. Here, trigrams were selected because for Word2Vec and FastText models also, a word window of three was used.

5.2. Results

The following Tables 2-5 document the results obtained for each graph construction algorithm in combination with the classification methods.

Word To Vector Algorithm	MAD	Abs	LP-ZGL
Word2Vec (SkipGram)	0.7534	0.7531	0.7201
Word2Vec (Bag of words)	0.6945	0.6967	0.6754
Fasttext (SkipGram)	0.8146	0.814	0.822
Fasttext (Bag of Words)	0.795	0.7952	0.801

Table 2: Accuracy of Information Theoretic Metric Learning

Word To Vector Algorithm	MAD	Abs	LP-ZGL
Word2Vec (SkipGram)	0.7012	0.701	0.721
Word2Vec (Bag of words)	0.6641	0.6542	0.665
Fasttext (SkipGram)	0.7886	0.7935	0.7988
Fasttext (Bag of Words)	0.7712	0.775	0.7767

Table 3: Accuracy of Sparse Determinant Metric Learning

Word To Vector Algorithm	MAD	Abs	LP-ZGL
Word2Vec (SkipGram)	0.734	0.733	0.732
Word2Vec (Bag of words)	0.701	0.71	0.711
Fasttext (SkipGram)	0.8547	0.861	0.8634
Fasttext (Bag of Words)	0.823	0.834	0.845

Table 4: Accuracy of Least Squares Metric Learning

Word To Vector Algorithm	MAD	Abs	LP-ZGL
Word2Vec (SkipGram)	0.7678	0.7775	0.7757
Word2Vec (Bag of words)	0.7664	0.7567	0.7456
Fasttext (SkipGram)	0.8673	0.8573	**0.8743**
Fasttext (Bag of Words)	0.85	0.853	0.86

Table 5: Accuracy of Local Fisher Discriminant Analysis

As illustrated above, Local Fisher Discriminant Analysis(LFDA) combined with Label propagation yields the best accuracy of 0.8743. LFDA is a linear supervised dimensionality reduction method. It proved effective in our case since each of our words had a size of 300 dimensions. FastText(skipgram) in combination with label propagation consistently performed better than other algorithms in all graph construction methodologies.
To test the robustness of the approach, we trained the best performing combination (LFDA and LP-ZGL) with 20k words and tested with 60k words. It yielded an accuracy of 0.753. Meanwhile, the baseline CRF model only gave an accuracy score of 0.633. This proves that our approach is more robust even when the labelled data set is comparatively small.

6. Conclusion

Our research establishes the fact that graph based semi-supervised approaches are more robust than supervised classification algorithms for POS tagging when the data set is relatively small. Thus graph based semi supervised data can be employed in the early stages of creating POS tagged data sets. Human annotators can correct the automatically annotated corpus with less effort, and the corrected annotated data set can be used in an iterative manner to re-train the tagger. Thus, graph based semi-supervised approaches are particularly useful for POS tagging of low-resourced languages such as Tamil. We used neural word embedding to create a vector representation of words, and Mahanalobis distance to measure distance between word vectors in order to build the graph. This shows that word embedding provides an excellent alternative for WordNet in measuring similarity between words, especially for languages that do not have a WordNet. This is useful not only for graph building, but for any task that requires measuring the similarity of words.

7. Future work

Our language independent work has shown promise with low resources. We have only done the research for one language, and this research should be extended to other languages to verify the general applicability of the presented methodology. We hope to extend this idea for other low resourced sequential tagging problems such as Named Entity Recognition. This research can also be extended to improve and incorporate other word embedding techniques such as VarEmbed that uses morphological priors for probabilistic neural word embedding (Bhatia et al., 2016). We can also experiment with other graph construction algorithms such as b-matching (Jebara et al., 2009). The main limitation of this technique is the amount of time taken to build the graph. Thus we intend to look into different code optimization methods. While we have compared our approach with the pure CRF implementation, Lample et al. (2016) has shown that CRF in combination with LSTM can provide a higher accuracy for Named entity recognition but that approach has not been tried for POS tagging in morphologically complex languages such as Tamil. We are eager to see how our approach stacks up with them.

8. Acknowledgement

This research is partially funded by the National Languages Processing Centre of University of Moratuwa, Sri Lanka. The authors would also like to thank the LKDomain registry for partially funding this publication.

9. Bibliographical References

Abinaya, N., John, N., Ganesh, B. H., Kumar, A. M., and Soman, K. (2014). Amrita_cen@ fire-2014: Named entity recognition for indian languages using rich features. In *Proceedings of the Forum for Information Retrieval Evaluation*, pages 103–111. ACM.

Ben, X., Meng, W., Yan, R., and Wang, K. (2012). An improved biometrics technique based on metric learning approach. *Neurocomputing*, 97:44 – 51.

Bhatia, P., Guthrie, R., and Eisenstein, J. (2016). Morphological priors for probabilistic neural word embeddings. 3 August.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2016). Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606.*

Celikyilmaz, A., Thint, M., and Huang, Z. (2009). A graph-based semi-supervised learning for question-answering. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*, ACL '09, pages 719–727, Stroudsburg, PA, USA. Association for Computational Linguistics.

Davis, J. V., Kulis, B., Jain, P., Sra, S., and Dhillon, I. S. (2007). Information-theoretic metric learning. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, pages 209–216, New York, NY, USA. ACM.

De Maesschalck, R., Jouan-Rimbaud, D., and Massart, D. L. (2000). The mahalanobis distance. *Chemometrics and intelligent laboratory systems*, 50(1):1–18.

Dhanalakshmi, V., Rajendran, S., Soman, K. P., and Edu, K. (2009). POS tagger and chunker for tamil language.

Dhillon, P. S., Talukdar, P. P., and Crammer, K. (2010). Inference driven metric learning (idml) for graph construction.

Elsner, M. and Charniak, E. (2010). Disentangling chat. *Comput. Linguist.*, 36(3):389–409, September.

Ganesh, J., Parthasarathi, R., Geetha, T. V., and Balaji, J. (2014). Pattern based bootstrapping technique for tamil POS tagging. In *Mining Intelligence and Knowledge Exploration*, Lecture Notes in Computer Science, pages 256–267. Springer, Cham.

Gimenez, J. and Marquez, L. (2004). Svmtool: A general pos tagger generator based on support vector machines. In *In Proceedings of the 4th International Conference on Language Resources and Evaluation*.

Huang, Z., Xu, W., and Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. *CoRR*, abs/1508.01991.

Jebara, T., Wang, J., and Chang, S.-F. (2009). Graph construction and b-matching for semi-supervised learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 441–448, New York, NY, USA. ACM.

Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2016). Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. (2016). Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California, June. Association for Computational Linguistics.

Liu, E. Y., Guo, Z., Zhang, X., Jojic, V., and Wang, W. (2012). Metric learning from relative comparisons by minimizing squared residual. In *Proceedings of the 2012 IEEE 12th International Conference on Data Mining*, ICDM '12, pages 978–983, Washington, DC, USA. IEEE Computer Society.

Liu, Y., Kirchhoff, K., Liu, Y., and Kirchhoff, K. (2016). Graph-based semisupervised learning for acoustic modeling in automatic speech recognition. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 24(11):1946–1956, November.

McFee, B., Barrington, L., and Lanckriet, G. R. G. (2011). Learning content similarity for music recommendation. *CoRR*, abs/1105.2344.

Mihalcea, R. F. and Radev, D. R. (2011). *Graph-based Natural Language Processing and Information Retrieval*. Cambridge University Press, New York, NY, USA, 1st edition.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. 16 January.

Niyogi, P. (2013). Manifold regularization and semi-supervised learning: Some theoretical analyses. *Journal of Machine Learning Research*, 14:1229–1250.

Pandian, S. L. and Geetha, T. V. (2009). Crf models for tamil part of speech tagging and chunking. In *Proceedings of the 22Nd International Conference on Computer Processing of Oriental Languages. Language Technology for the Knowledge-based Economy*, ICCPOL '09, pages 11–22, Berlin, Heidelberg. Springer-Verlag.

Qi, G.-J., Tang, J., Zha, Z.-J., Chua, T.-S., and Zhang, H.-J. (2009). An efficient sparse metric learning in high-dimensional space via l1-penalized log-determinant regularization. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 841–848, New York, NY, USA. ACM.

Rani, P., Pudi, V., and Sharma, D. M. (2016). A semi-supervised associative classification method for POS tagging. *Int J Data Sci Anal*, 1(2):123–136, 1 July.

Shaw, B., Huang, B., and Jebara, T. (2011). Learning a distance metric from a network. In J. Shawe-Taylor, et al., editors, *Advances in Neural Information Processing Systems 24*, pages 1899–1907. Curran Associates, Inc.

Subramanya, A., Petrov, S., and Pereira, F. (2010). Efficient graph-based semi-supervised learning of structured tagging models. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 167–176, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sugiyama, M. (2006). Local fisher discriminant analysis for supervised dimensionality reduction. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 905–912, New York, NY, USA. ACM.

Talukdar, P. P. and Pereira, F. (2010). Experiments in graph-based semi-supervised learning methods for class-instance acquisition. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 1473–1481, Stroudsburg, PA, USA. Association for Computational Linguistics.

Talukdar, P. P., Reisinger, J., Paşca, M., Ravichandran, D., Bhagat, R., and Pereira, F. (2008). Weakly-supervised acquisition of labeled class instances using graph random walks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 582–590, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zhu, X., Ghahramani, Z., and Lafferty, J. (2003). Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ICML'03, pages 912–919. AAAI Press.

Zhu, X., Lafferty, J., and Rosenfeld, R. (2005). *Semi-supervised learning with graphs*. Ph.D. thesis, Carnegie Mellon University, language technologies institute, school of computer science.

Zhu, L., Gao, S., Pan, S. J., Li, H., Deng, D., and Shahabi, C. (2013). Graph-based informative-sentence selection for opinion summarization. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '13, pages 408–412, New York, NY, USA. ACM.

10. Language Resource References

Forum for Information Retrieval Evaluation. (2014). *FIRE Corpus*. Indian Institute of Science, Bangalore.

Wikipedia. (2016). *Tamil Wikipedia Corpus*. Wikipedia.