Institute of Science and Technology
Tribhuvan University

Seminar Report

On

Structural SVM for Multi-label Text Classification

Submitted to

Central Department of Computer Science and Information Technology

Tribhuvan University

Submitted by

Joshana Shakya (15/075)

In partial fulfillment of the requirement for Master's Degree in Computer Science and Information Technology, 2nd Semester

February, 2021

Institute of Science and Technology
Tribhuvan University

Date: ....

Recommendation Letter of Supervisor

I hereby recommend that this seminar report is prepared under my supervision by **Joshana Shakya** entitled **”Structural SVM for Multi-label Text Classification”** be accepted as fulfillment in partial requirement for the degree of Master’s of Science in Computer Science and Information Technology. In my best knowledge, this is an original work in computer science.

....

Asst. Prof. Bal Krishna Subedi
Supervisor
CDCSIT, TU

CERTIFICATE OF APPROVAL

Date: . . . .

This is certify that the seminar report prepared by Joshana Shakya entitled ”Structural SVM for Multi-label Text Classification” in partial fulfillment of the requirements for the degree of M.Sc. in Computer Science and Information technology has been well studied. In our opinion, it is satisfactory in the scope and quality as a project for the required degree.

. . . .

Asst. Prof. Bal Krishna Subedi
Supervisor
CDCSIT
TU

. . . .

Internal Examiner

. . . .

Asst. Prof. Nawaraj Paudel
Head of Department CDCSIT
TU

ACKNOWLEDGEMENT

It gives us immense pleasure to express my deepest sense of gratitude and sincere thanks to our highly respected and esteemed guide Asst. Prof. Bal Krishna Subedi for his valuable guidance, encouragement and help for completing this work. His useful suggestions for this whole work and co-operative behavior are sincerely acknowledged.

I would like to express my sincere thanks to our Department Head Asst. Prof. Nawaraj Paudel, Central Department of Computer Science and Information Technology, for whole hearted support.

I am also grateful to our teachers for their constant support and guidance. At the end I would like to express our sincere thanks to all my friends and others who helped me directly or indirectly during this seminar work.

Joshana Shakya (15/075)

ABSTRACT

Machine learning methods usually solve classification and regression problems with single output. To solve problems that include complex output spaces, structured output prediction methods such as structural SVM are used. Multi-label classification problem involves predicting zero or more mutually non-exclusive class labels. Therefore, this problem has complex output space. The pystruct implementation of 1-slack structural SVM is used to perform multi-label classification of text documents. The performance evaluation shows small loss even for small vocabulary size.

Keywords: *Hamming Loss, Multi-label Classification, Structural Support Vector Machine*

TABLE OF CONTENTS

LIST OF TABLES

ABBREVIATIONS

CC	Correlation Coefficient
IG	Information Gain
MlKNN	Multi-label K-Nearest Neighbor
QPBO	Quadratic Pseudo-Boolean Optimization
OP	Optimization Problem
QP	Quadratic Programming
SSVM	Structural Support Vector Machine
SVM	Support Vector Machine
SVMHMM	Structural SVM Hidden Markov Models
SU	Symmetrical Uncertainty
TF	Term Frequency
TF-IDF	Term Frequency Inverse Document Frequency

CHAPTER 1 INTRODUCTION

1.1 Background

We usually use machine learning methods for classification and regression problems with single output. But what if we need to predict multiple, correlated outputs simultaneously? Such problems include diverse set of tasks, for example, natural language parsing, protein structure prediction, or image segmentation. Structured output prediction is the branch of machine learning that solves these problems. It is the problem of teaching machines to predict complex structured objects and is very different from traditional machine learning problems, where the output set contains only simple categorical labels or real values.

Given a set of training examples $S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\} \in (X \times Y)^n$, the structured output prediction learns a hypothesis $h : X \to Y$ in the following form [8]:

$$h(x) = \text{argmax}_{y \in Y} f(x, y) \tag{1.1}$$

where the function f evaluates how well a particular output structure $y \in Y$ matches an input $x \in X$, and it is parameterized by the vector w, and the argmax over all possible $y \in Y$ extracts the highest scoring output as the prediction for x and is usually computed using the combinatorial optimization algorithms.

1.2 Structural Support Vector Machine

Structural Support Vector Machine (SSVM) is a discriminative structured output prediction algorithm which allows flexible construction of features to predict complex objects. It inherits the properties of regular SVM such as flexibility in the choice of loss function and the opportunity to learn non-linear rules via kernels.

Structural SVM tries to learn the scoring function f in Equation 1.1 such that good output structures score higher than bad output structures for a fixed input x. It uses disciminant function f as:

$$f(x, y) = w^T \cdot \psi(x, y) \tag{1.2}$$

where the combined feature vector function ψ relates inputs and outputs, and w are model parameters. The structural SVM formulation presented in Section 3.3.1 is used to learn model parameters w.

1.3 Multi-label Classification

Multi-label classification is a variant of the classification problem which involves predicting zero or more mutually non-exclusive class labels. In multi-label classification, the training set is composed of instances each associated with a set of labels, and the task is to predict the label sets of unseen instances through analyzing training samples with known label sets. There are two main strategies for multi-label classification:

i. Problem transformation method

ii. Algorithm adaptation method

The problem transformation methods transform multi-label problem into single-label problem whereas the algorithm adaptation method adapts the algorithm to directly perform multi-label classification.

1.4 Problem Statement

Multi-label classification problem can be treated as n classes independent binary classification problems. But taking class labels as atomic entities without analyzing the correlation between input and output, there is no learning, only memorization of class labels. It is necessary to enable learning across input and output space. Structural SVM extracts features from inputs and outputs, learns from the combined features to solve multi-label classification problem.

1.5 Objective

The main objective of this study is to use structural SVM to solve multi-label classification problem on text data.

CHAPTER 2 LITERATURE REVIEW

2.1 Background Study

In 2001, Conditional random field was introduced by Lafferty et al. for segmenting and sequencing labeled data. Later, the large margin methods for structured and interdependent output variables [7] were presented. This developed the concept of Structural SVM for label sequence learning, sequence alignment, multi-label classification, etc. Structural SVM for these different problems were implemented by Thorsten Joachims in C, and it was later implemented in python by Andreas Christian Muller.

2.2 Literature Review

A comparison between structural SVM and conventional SVM has been done for reference parsing in [9]. The authors compared structural SVM and conventional SVM in token-level and chunk-level reference parsing of medical journal articles using two types of contextual features: observation features and label features. They extracted the observation features from the neighboring tokens and treated the labels assigned to those tokens as the label features. Because structural SVM for reference parsing is implemented as a sequence learning algorithm, and the joint feature presentation function includes two kinds of features: state transition features and observation features extracted from individual tokens within a sequence, the authors used structural SVM designed specially for sequence labeling called SVMHMM. The authors concluded that although SVM performance improved when the second order contextual observation features were used, structural SVM achieved higher overall token-level and chunk-level accuracies than the SVM method. Both methods achieved above 98% token classification accuracy and an overall chunk-level accuracy of over 95%.

The authors in [1] has presented the construction of a classification model in multi-label scenario for the classification of product review documents. Their work dealt with the text classification problem using an approach of multi-label classification using structural support vector machine. They performed the experiment on a collection of product reviews of various electronic gadgets including mobile phones, tablets, laptops, pen-drives, etc. The process was carried out using python implementation of Struct SVM. With 85.4% accuracy, the authors concluded that the system that they built was an optimized method in the case of a multi-label text classification scenario. In their experiment, they observed that the training time for multi-label classification is considerably high for large datasets and hence they are extending their work with core vector machines considering the scalability aspects of the algorithm.

In [6], the authors explored the two main approaches in machine learning to solve multi-label classification problem: problem transformation and algorithm adaptation. The writers focused on the four different factors, i.e., feature weighting method, feature selection method, multi-label classification approach, and single-label classification algorithm to obtain the best combination. They performed the experiment using Binary, TF, and TF-IDF for feature selection, Information Gain(IG), Symmetrical Uncertainty(SU), and Correlation Coefficient(CC) for feature selection, problem transformation and algorithm adaptation for multi-label classification, and Nave Bayes, J48, SVM, Adaboost.MH, and MlkNN for single-label algorithm. The result showed that the problem transformation approach, generally, gave better result than algorithm adaptation. Problem transformation approaches that ignore label correlation obtain better performances than methods that consider label correlation because of the low correlation of data used in the experiment. They built a model for automatic multi-label classification of Indonesian news articles using the best combination obtained from the experiments. The combination consists of TF-IDF feature weighting method, Symmetrical Uncertainty feature selection method, Calibrated Label Ranking, which belongs to problem transformation approach, and SVM algorithm. This combination gave F-measure of 85.13% in 10-fold cross-validation. The F-measure decreased to 76.73% in testing mode.

CHAPTER 3 METHODOLOGY

3.1 Pre-processing methods

The datasets are processed through the pre-processing task of text documents such as multi-label binarization, tokenization, stop word removal, lowercase conversion, and stemming using Porter stemmer. Multi-label binarization encodes multiple labels (0 or 1) per instance. Tokenization is the process of splitting a text into words, phrases, or other meaningful elements called tokens. Stop words usually refer to the most common words in a language without dependency to a particular topic. In lowercase conversion, all uppercase characters are converted to their lowercase forms. Stemming refers to the process of reducing inflected words to their word stem, base, or root form.

3.2 Vectorization

The term-weighting scheme is used to vectorized the text in the document. The term-weighting schemes applied in the experiment is unigram Term Frequency Inverse Document Frequency (TF-IDF). The TF-IDF measure is given by:

$$v_{ij} = \frac{f_{ij}}{fd_j} \times \log\left(\frac{|D|}{ft_i}\right)$$

where $|D|$ is the total number of documents, f_{ij} is the number of occurrences of term i in document j, fd_j is the total numbers of terms occurring in document j, and ft_i the total number of documents in which term i appears at least once.

3.3 Text Classification

In the experiment, 1-slack Structural SVM was used for the classification.

3.3.1 Structural SVM Optimization Problem

Given a discriminant function of the form $f(x, y) = \langle w, \psi(x, y)\rangle$, with hypotheses of the form $h(x) = \text{argmax}_{y \in Y} f(x, y)$, where $h(x)$ is a hypothesis parameterized by w with training pairs in the form $S = \{(x_i, y_i) \in X \times Y : i = 1, 2, \ldots, n\}$, the hypothesis $h(x)$ may be learned with the following quadratic program [3] :

$$\min_{w,\,\xi}\ \frac{1}{2}||w||^2 + \frac{C}{n}\sum_{i=1}^{n}\xi_i \tag{3.1}$$

$$s.t.\ \forall i : \xi_i \geq 0, \tag{3.2}$$

$$\forall i, \forall y \in Y : \langle w, \psi(x_i, y_i)\rangle \geq \langle w, \psi(x_i, y)\rangle + \Delta(y_i, y) - \xi_i \tag{3.3}$$

The loss $\Delta(y_i, y)$ of the constraints associated output y is incorporated as the margin between the discriminant function for the correct output $\langle w, \psi(x_i, y_i)\rangle$ and the incorrect output y's discriminant function $\langle w, \psi(x_i, y)\rangle$.

3.3.2 Cutting Plane Algorithm

The complication with SSVM optimization problem is that there are as many constraints as there are possible labels. For example, in multi-label classification, for a sample size m where there are l possible labels for each sample, there would be l number of constraints. Despite an immense number of constraints, structural SVM employs an Algorithm 1 to solve the optimization problem efficiently.

Algorithm 1 Structural SVM Cutting Plane Algorithm

Input: $(x_1, y_1), \ldots, (x_n, y_n), C, \epsilon$
$S_i \leftarrow \phi$ for all $i = 1, \ldots, n$
repeat
    for i=1, . . . , n **do**
        $H(y) \equiv \Delta(y_i, y) + \langle w, \psi(x_i, y)\rangle - \langle w, \psi(x_i, y_i)\rangle$
        compute $\hat{y} = \text{argmax}_{y \in Y} H(y)$
        compute $\xi_i = \max\{0, \max_{y \in S_i} H(y)\}$
        if $H(\hat{y}) > \xi_i + \epsilon$ **then**
            $S_i \leftarrow S_i \cup \{\hat{y}\}$
            $w \leftarrow$ optimize primal over $\cup_i S_i$
        end if
    end for
until no S_i has changed during iteration

The algorithm starts with an empty set of constraints, iteratively finds the most violated constraint, add this constraint into a working set, and reoptimizes the quadratic program with this additioinal constraint. The most violated constraint for a training example (x_i, y_i) is the constraint in the full QP that requires the highest slack ξ_i and is determined by solving

$$\hat{y} = \text{argmax}_{y \in Y} H(y) \tag{3.4}$$

where the cost function H is:

$$H(y) \equiv \langle w, \psi(x_i, y) \rangle - \langle w, \psi(x_i, y_i) \rangle + \Delta(y_i, y) \tag{3.5}$$

If the constraint is violated by more than a predefined tolerance ϵ, then the constraint is added, and otherwise it is ignored. The algorithm terminates when no valid constraint is added [3].

3.3.3 1-Slack Structural SVM

1-slack structural SVM reformulates the structural SVM quadratic problem by combining examples in a training set $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ into a single training example. There is no slack vector $\xi = \xi_1, \ldots, \xi_n$ for every training example, but rather a single scalar slack variable ξ.

$$\min_{w,\xi} \frac{1}{2}||w||^2 + C\xi \tag{3.6}$$

$$s.t. \forall i : \xi_i \geq 0, \tag{3.7}$$

$$\forall (\bar{y_1}, \ldots, \bar{y_n}) \in Y^n : \left\langle w, \frac{1}{n} \sum_{i=1}^{n} \psi(x_i, y_i) \right\rangle \tag{3.8}$$

$$\geq \left\langle w, \frac{1}{n} \sum_{i=1}^{n} \psi(x_i, \bar{y_i}) \right\rangle + \frac{1}{n} \sum_{i=1}^{n} \Delta(y_i, \bar{y_i}) - \xi \tag{3.9}$$

The idea of this optimization problem (OP) is that there is a constraint for every single combination of outputs across all training examples. This is in contrast to OP previously defined, which has a family of constraints for each training example, with one constraint per example per output [3].

The cutting plane algorithm analogous to Algorithm 1 to solve the above OP is:

Algorithm 2 1-Slack Structural SVM Cutting Plane Algorithm

```
Input: (x_1, y_1), ..., (x_n, y_n), C, ϵ
S_i ← φ
for i=1, ..., n do
    {set up cost functions}
    H_i(y) ≡ Δ(y_i, y) + ⟨w, ψ(x_i, y)⟩ − ⟨w, ψ(x_i, y_i)⟩
end for
repeat
    for i=1, ..., n do
        compute ŷ_i = argmax_{y∈Y} H_i(y)
    end for
    compute ξ = (1/n) max_{(ȳ_1,...,ȳ_n)∈S} Σ_{i=1}^n max(0, H_i(ȳ_i))
    if (1/n) Σ_{i=1}^n H_i(ŷ_i) > ξ + ϵ then
        S ← S ∪ {(ŷ_1, ..., ŷ_n)}
        w ← optimize primal over S
    end if
until S has not changed during iteration
```

CHAPTER 4 EXPERIMENTATION

4.1 Dataset

The experiment is made on Reuters-21578 [4] dataset, in which five categories and 560 documents are used. The Reuters-21578 collection consists of 21578 documents and 135 categories. 392 documents are used as training samples and the remaining 168 documents are used as testing samples.

Table 4.1: Training set and Testing set

Category	corn	cotton	rice	soybean	wheat
Training set	149	43	43	77	209
Testing set	74	19	24	34	78

Table 4.2: Dataset and Categories

No. of categories	one	two	three	four	five
Training set	298	66	23	3	2
Testing set	132	18	13	3	2

Table 4.3: Sample Dataset

doc_id	text	corn	cotton	rice	soybean	wheat
1	USDA REPORTS 10.572 MLN ACRES IN CONSERVATION The U.S. Agriculture Department has accepted 10,572,402 more acres of highly erodable cropland into the Conservation Reserve Program, USDA announced. In the latest signup, farmers on 101,020 farms ...	1	0	0	0	0
13	TRADE SEES STEADY CORN/WHEAT EXPORT INSPECTIONS The USDA's weekly export inspection report is expected to show steady corn and wheat exports and lower soybean exports, according to CBT floor traders' forecasts. Traders ...	1	0	0	1	1

4.2 Implementation Environment

Anaconda distribution of Python 3.6 is used as programming language. The other libraries used are:

i. numpy: To manage list of data

ii. pandas: To manipulate data

iii. nltk: To preprocess data

iv. scikit-learn: To split dataset into training and testing set, to perform TF-IDF vectorization, and to compute hamming loss

v. pystruct: To perform multi-label classification using structured SVM

4.3 Parameters

i. The vocabulary size is 3827.

ii. The inference method method used is qpbo.

iii. The inference_cache is 50 as in pystruct documentation. If > 0 the most violating of the cached examples will be used to construct a global constraint.

iv. The cost of constraints violation, C is taken 0.1.

v. The convergence tolerance, tol is taken 0.01 as in pystruct documentation. If dual objective decreases less than tol, learning is stopped. The default corresponds to ignoring the behavior of the dual objective and stop only if no more constraints can be found.

vi. The maximum number of iterations defined in pystruct is 10000.

4.4 Example

A news entitled China maintains strong demand for U.S. corn, soy: USDA is taken as sample document from Reuters website[2] to perform structured SVM multi-label classification. The sample is:
CHICAGO (Reuters) - Chinese demand for U.S. corn and soybeans remained robust in the latest week, U.S. Agriculture Department (USDA) data showed, and traders expect the recent surge of deals will cause the U.S. government to boost its export forecast for both commodities.

A USDA report released Friday showed that export sales of soybeans to China totaled 1.608 million tonnes in the week ended Sept. 3, the latest reporting period. Weekly corn export sales to China were 1.137 million tonnes. [EXP/SOY] [EXP/COR]

Separately, the government said private exporters reported the sale of 262,000 tonnes of soybeans to China, the sixth straight trading day that an announcement of a sale to the worlds top buyer of the oilseed has been announced.

China also has been ramping up its U.S. corn imports, as the country faces its first real corn shortfall of corn in years. A sharp price surge in corn - critical for Chinas mammoth hog, dairy and poultry sectors - is the latest in a series of ructions that include a devastating pig disease, pandemic-driven upsets for international suppliers and warnings of a growing food supply gap.

The trade is pricing in expectations that USDA will cut the size of this years corn and soybean crops while increasing demand estimates due to the recent Chinese buying spree, Arlan Suderman, chief commodities economist for broker StoneX, said in note to clients.

USDA will give an update on the export outlook in its monthly World Agricultural Supply and Demand Estimates report at 11 a.m. CDT (1600 GMT) on Friday.

China vowed to import $36.5 billion of U.S. agricultural goods annually as part of a Phase 1 trade deal signed in January. Chinese purchases through the first seven months of the year totaled just $8.559 billion, according to U.S. Census Bureau trade data.

The outputs of 3 steps in structured SVM multi-label classification are as follows:

Step 1: Preprocessing (tokenize, stopword removal, porter stemming)

chicago reuter chines demand us corn soybean remain robust latest week us agricultur depart usda data show trader expect recent surg deal caus us govern boost export forecast commoditiesa usda report releas friday show export sale soybean china total million tonn week end sept latest report period weekli corn export sale china million tonn expsoy expcorsepar govern said privat export report sale tonn soybean china sixth straight trade day announc sale world top buyer oilse announcedchina also ramp us corn import countri face first real corn shortfal corn year sharp price surg corn critic china mammoth hog dairi poultri sector latest seri ruction includ devast pig diseas pandemicdriven upset intern supplier warn grow food suppli gapth trade price expect usda cut size year corn soybean crop increas demand estim due recent chines buy spree arlan suderman chief commod economist broker stonex said note clientsusda give updat export outlook monthli world agricultur suppli demand estim report cdt gmt fridaychina vow import billion us agricultur good annual part phase trade deal sign januari chines purchas first seven month year total billion accord us censu bureau trade data

Step 2: Vectorization

Non-zero TF-IDF values of the words are:

0.0517, 0.0757, 0.0414, 0.0510, 0.0594, 0.1003, 0.1101, 0.0717, 0.0939, 0.0763, 0.0440, 0.0717, 0.0560, 0.1003, 0.0661, 0.1003, 0.2115, 0.2113, 0.0450, 0.2532, 0.0429, 0.0731, 0.0360, 0.0560, 0.0828, 0.1564, 0.0606, 0.1385, 0.1763, 0.0329, 0.0893, 0.0803, 0.0489, 0.0893, 0.0460, 0.0871, 0.0844, 0.1418, 0.0717, 0.1066, 0.0565, 0.0546, 0.0782, 0.0692, 0.0893, 0.0546, 0.0880, 0.0620, 0.1003, 0.0767, 0.0463, 0.0420, 0.0613, 0.0628, 0.1984, 0.2007, 0.0420, 0.0782, 0.0582, 0.0652, 0.0746, 0.0537, 0.0560, 0.1003, 0.1003, 0.0939, 0.0696, 0.0542, 0.0455, 0.0893, 0.1026, 0.0661, 0.0565, 0.1464, 0.0161, 0.0365, 0.1723, 0.0746, 0.0939, 0.0893, 0.0576, 0.0782, 0.0939, 0.1241, 0.0704, 0.0939, 0.0857, 0.1588, 0.1003, 0.0985, 0.0717, 0.1877, 0.0764, 0.0803, 0.0801, 0.1489, 0.0571, 0.1003, 0.1435, 0.1072, 0.0828, 0.0798, 0.0652, 0.0972, 0.0904

Step 3: Classification

The classification output of the sample document is ’corn’.

CHAPTER 5 RESULT AND ANALYSIS

5.1 Performance Evaluation

5.1.1 Hamming Loss

Hamming loss computes the average Hamming distance between two sets of samples. If $\hat{y}_i$ is the predicted value for the $i-th$ label of a given sample, y_i is the corresponding true value, and n is the number of labels, then the Hamming loss L between two samples is defined as [5]:

$$L_{Hamming}(y, \hat{y}) = \frac{1}{n_{labels}} \sum_{i=0}^{n_{labels}-1} 1(\hat{y}_i \neq y_i)$$

where 1(x) is the indicator function.

5.2 Result and Analysis

Table 5.1 shows the Hamming loss of the training and testing datasets obtained when using structured svm.

Table 5.1: Hamming loss of training set and testing set

Vocabulary size	Training Loss	Test Loss
3827	0.067347	0.134524

The results show that there is approximately 6% loss in training data and 13% loss in testing data.

CHAPTER 6 CONCLUSION

6.1 Conclusion

In this study, structural SVM was used for multi-label classification of Reuters-21578 dataset, where only five classes: corn, cotton, rice, soybean, and wheat were chosen. 70% of the datasets were used as training and the remaining 30% were used as test data. The performance of the classification was evaluated using Hamming Loss.

From the observations, it can be concluded that structural SVM has loss percentage approximately 6% in training and approximately 13% in testing data, even when number of vocabulary size is small.

REFERENCES

[1] B.Chrystal, J. and Joseph, S., "Multi-Label Classification of Product Reviews Using Structured SVM," *International Journal of Artificial Intelligence & Applications*, vol. 6, no. 3, pp. 61–68, 2015. doi: 10.5121/ijaia.2015.6306.

[2] "China maintains strong demand for U.S. corn, soy: USDA — Reuters," 2020. [Online]. Available: https://www.reuters.com/article/us-usa-trade-china-soybeans-idUSKBN26224K [Accessed: 02/03/2021].

[3] Finley, T. W., "Supervised Clustering With Structural Svms," *Development*, no. January, 2009.

[4] Lewis, D. D., "UCI Machine Learning Repository: Reuters-21578 Text Categorization Collection Data Set," 1987. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Reuters-21578+Text+Categorization+Collection [Accessed: 02/03/2021].

[5] "Metrics and scoring: Quantifying the quality of predictions," 2007. [Online]. Available: https://scikit-learn.org/stable/modules/model_evaluation.html#hamming-loss.

[6] Rahmawati, D. and Khodra, M. L., "Automatic multilabel classification for Indonesian news articles," *ICAICTA 2015 - 2015 International Conference on Advanced Informatics: Concepts, Theory and Applications*, pp. 1–6, 2015. doi: 10.1109/ICAICTA.2015.7335382.

[7] Tsochantaridis, I., Joachims, T., Hofmann, T., and Altun, Y., "Large margin methods for structured and interdependent output variables," *Journal of Machine Learning Research*, vol. 6, pp. 1453–1484, 2005, ISSN: 15337928.

[8] Yu, C., "Improved Learning of Structural Support Vector Machines: Training With Latent Variables and Nonlinear Kernels," no. January, pp. 1–189, 2011. [Online]. Available: http://webdocs.cs.ualberta.ca/$\sim$chunnam/papers/thesis_cnyu.pdf.

[9] Zhang, X., Zou, J., Le, D. X., and Thoma, G. R., "A structural SVM approach for reference parsing," *Proceedings - 9th International Conference on Machine Learning and Applications, ICMLA 2010*, pp. 479–484, 2010. doi: 10.1109/ICMLA.2010.77.