http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 12 (20): 5955-5961, 2013
ISSN 1812-5638 / DOI: 10.3923/itj.2013.5955.5961

Simple Semi-supervised Learning for Chinese Word Segmentation and Pos Tagging

Xinxin Li, Xuan Wang and Muhammad Waqas Anwar
Harbin Institute of Technology Shenzhen Graduate School, Shenzhen 518055, China

Abstract: Strategies of unlabeled data selection are important for semi-supervised learning of natural language processing tasks. To increase the accuracy and diversity of new labeled data, plenty of methods have been proposed, such as ensemble-based self-training, co-training and tri-training methods. In this paper, we propose a simple and effective semi-supervised algorithm for Chinese word segmentation and part-of-speech tagging problem which selects new labeled data agreed by two different approaches: character-based and word-based models. Theoretical and experimental analysis verifies that sentences with same annotation on both models are more accurate than those generated by single models and are suitable for semi-supervised learning as additional data. Experimental results on Chinese Treebank 5.0 demonstrate that our semi-supervised approach is comparable with the best reported semi-supervised approach which employs complex feature engineering.

Key words: Chinese word segmentation and POS tagging, semi-supervised learning, new agreed data, joint decoding

INTRODCUTION

Chinese Word Segmentation (CWS) and Part Of Speech (POS) tagging are prerequisite steps for deep understanding of Chinese, such as syntactic and semantic parsing. The two tasks can be processed in a pipeline, where the boundary of each word is determined first and then its POS tag is labeled. Using the cascaded approach, both tasks are solved separately as sequence labeling problems (Ng and Low, 2004). However, there are two drawbacks for this approach. The first one is error propagation along the pipeline. POS tags on mistakenly recognized words will never be correctly labeled. The second disadvantage is that useful information on both tasks can't be shared. Word segmentation cannot benefit from POS tagging which it does.

Current approaches treat CWS and POS tagging problem as a joint task. Ng and Low proposed a character-based model to label every character incrementally, where the label of each character is a combination of the boundary and POS tag of the word (Ng and Low, 2004). Zhang introduced a word-based model treating word as a basic unit and performed CWS and POS tagging simultaneously (Zhang and Clark, 2010). However, POS tagging for Chinese is still more inaccurate than English. In one hand, Chinese words are lack of rich morphological features. In the other hand, the basic unit of Chinese, word, has a higher out of vocabulary rate than English. With only raw text given, the accuracy of POS tagging after automatic segmentation is even lower (Zhang and Clark, 2008; Jiang *et al*., 2008; Kruengkrai *et al*., 2009).

Semi-supervised methods use unlabeled data to improve the accuracy of classifiers trained only on labeled data (Yarowsky, 1995). Self-training method incrementally adds new labeled data by the classifier trained on annotated data, which has been successfully applied on POS tagging and parsing (McClosky *et al*., 2008). However, the performance for self-training approach heavily depends on the performance of original classifier. Ensemble-based self-training method uses a cascaded classifier that combines multiple base classifiers or a reranking model to select the optimal one from multiple candidates as additional data (Spoustova *et al*., 2009). Co-training method requires two classifiers, each one with different view of the data (Blum and Mitchell, 1998). Abney improved co-training method by selecting each classifier maximizing the agreement with another classifier on unlabeled data (Abney, 2002).

In this study, we proposed a new semi-supervised algorithm for Chinese word segmentation and POS tagging problem by selecting new labeled sentences annotated same by character-based model and word-based model as additional training data. We name these additional data as new agreed data or new agreed sentences. It operates similar as self-training method, but use a different strategy of additional data selection. Both baseline character-based and word-based models are obtained on the training data and new agreed data. Then these two models are combined in a joint model using a linear function (Li *et al*., 2011). Experiments on Chinese Treebank 5.0 show that our approach is comparable to the best semi-supervised approach which extracts rich features from new annotated data to their model.

Corresponding Author: Xinxin Li, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen 518055, China

SEMI-SUPERVISED APPROACH

Experimental results reveal that co-training method is beneficial for POS tagging on small training data, but doesn't study well on large training data (Clark *et al*., 2003). Instead of previous self-training and co-training methods, we propose a semi-supervised approach augmenting the training data with more confident new labeled data, which are annotated same by two different classifiers. The difference between Abney's co-training algorithm and our approach is that his algorithm selects each classifier which has the most agreement with the other classifier on unlabeled data and our method selects new agreed data as additional data directly. Our approach can select different sizes of new agreed data with or without iteration. In this paper, we set it with no iteration. The approach is shown in 1.

Algorithm 1: Semi-supervised approach with new example selection strategy
Input: L is labeled data
U is unlabeled sentences
Train classifiers f_1, f_2 using L
Label U with classifier f_1
Label U with classifier f_2
Generate new agreed data U' annotated same by two classifiers f_1, f_2
Train new semi-supervised classifiers f_{s1}, f_{s2} using $L+U'$

Analysis on binary classification: Abney gives a theoretical justification for standard co-training algorithm and the independence relax one (Abney, 2002). In this section, we provide the analysis for our new example selection strategy. Suppose there are two weak correct classifiers S and T for a binary classification problem. The two classifiers suffice the independence assumption. We use the same figure as Abney's work to exhibit the distribution of two classifiers. In Fig. 1, Y represents the true distribution of training data. +, - denote the positive and negative classes, one of which each example belongs to. Area A and D in Fig. 1 represent the examples that are classified as positive by both classifiers, in which A is positive examples and D is actually negative examples.

The accuracy of classifier S is calculated as:

$$\begin{aligned} P(S) &= P(S=+,Y=+)+P(S=-,Y=-) \\ &= N(S=+,Y=+)/N_{all}+P(S=-,Y=-)/N_{all} \\ &= (A+B+G+H)/(A+B+C+D+E+F+G+H)>0.5 \end{aligned}$$

The accuracy of classifier T is:

$$\begin{aligned} P(T) &= P(T=+,Y=+)+P(T=-,Y=-) \\ &= N(T=+,Y=+)/N_{all}+N(T=-,Y=-)/N_{all} \\ &= (A+C+F+H)/(A+B+C+D+E+F+G+H)>0.5 \end{aligned}$$

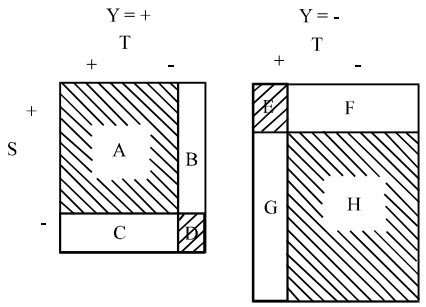

Fig. 1: Disagreement of two classifiers S and T

where, N_{all} represents the number of all data. Then the number of data agreed by classifiers S and T is N (S = T) = N (S = T = +)+N (S = T = -) = A+D+E+H, the shade area in Fig. 1, in which the number of correct part is N (S = T = Y) = A+H. Then the accuracy of agreed data is:

$$\begin{aligned} P(S=T=Y|S=T) &= N(S=T=Y)/N(S=T) \\ &= (A+H)/(A+D+E+H) \end{aligned}$$

where, suppose the unlabeled data has the same distribution as training data. We can prove the examples agreed by two classifiers follow Theorem 1:

$$(A+H)/(A+D+E+H) > \max(P(S),P(T))$$

where, since the two classifiers S, T are assumed independently, the prove is simple. The theorem shows that the data agreed by two classifiers is more accurate than both classifiers S and T.

Analysis on sequence labeling problem: The theorem can also be applied on sequence labeling problem. Assume two weak correct classifiers S and T for sequence labeling problem are independent and the annotations for different tokens in the same sentence are also independent. Suppose the accuracy of classifiers S and T are separately ps and pt (0.5<ps, pt = 1). Then the proportions of sentences with m correct annotations in all sentences with length n (0 = m = n) for two classifiers are separately:

$$P(N_S = m) = C_m^n * p_s^m * (1-p_s)^{(n-m)}$$

$$P(N_T = m) = C_m^n * p_t^m * (1-p_t)^{(n-m)}$$

where, each sentence has an accuracy m/n. The proportions of sentences that are correct on all tokens is $P(N_s = n) = p_s^n$ and $P(N_T = n) = p_t^n$.

Then the proportion of sentences agreed by two classifiers is:

$$P(N_S = N_T) = \sum_{m=0}^{n} C_m^n * p_s^m * (1-p_s)^{(n-m)} * p_t^m * (1-p_t)^{(n-m)}$$
$$= (p_s p_t + (1-p_s)(1-p_t))^n$$

where, the formula shows that the number of new agreed sentences will become larger with the increase of the accuracy of both classifiers S and T.

Theorem 2. The accuracy of these new agreed sentences is:

$$\sum_{m=0}^{n} (m/n * C_m^n * p_s^m * (1-p_s)^{(n-m)} * p_t^m * (1-p_t)^{(n-m)}) / P(N_S = N_T)$$
$$= p_s p_t * (p_s p_t + (1-p_s)(1-p_t))^{N-1} / P(N_S = N_T)$$
$$= p_s p_t / (p_s p_t + (1-p_s)(1-p_t))$$
$$\geq \max(p_t, p_t) \quad \Leftarrow 0.5 \leq p_s, p_t \leq 1$$

where, the accuracy in theorem 2 is a constant determined only by the accuracy of both classifiers S and T, not by the sentence length n. The theorem shows that the accuracy of new agreed sentences is better than both classifiers. More precise both classifiers are, more accurate the new agreed sentences will be. In section 4, we will perform experiments on CWS and POS tagging problem to validate the theorem.

MODELS FOR CHINESE WORD SEGMENTATION AND POS TAGGING

The two baseline classifiers for CWS and POS tagging in our semi-supervised approach are character-based and word-based models. Averaged perceptron algorithm is used to train both baseline classifiers (Collins, 2002). It can perform simple, efficient training and achieve the state-of-art performance for sequence labeling problems (Spoustova *et al*., 2009). For an input sentence x, the correct output is selected as:

$$F(x) = \arg\max_{y \in GEN(x)} \Phi(x, y) \times \bar{\alpha}$$

where GEN (x) is the set of all candidate tag sequences for x and formula represents the inner product of the features and their corresponding weights. The feature weight vector is updated followed by:

$$\bar{\alpha} = \bar{\alpha} + \Phi(x, y) - \Phi(x, F(x))$$

For CWS and POS tagging problem, the features, including words, characters and POS tags, can be incorporated into the model. It iteratively updates all features of a whole sentence at once instead of only features of a character or a word in the sentence. In the training phase, we employ a lazy update optimization which doesn't update the averaged weight of each feature in every iteration, only update when its weight changes. Results reveal that it significantly accelerates the training phase.

Table 1: Feature template for character-based model

1	c_n, c_0c_n (n = -2..2)
2	c_nc_{n+1},$c_0c_nc_{n+1}$ (n = -2..1)
3	$c_{-1}c_1$,$c_0c_{-1}c_1$

Table 2: Feature template for word-based model

Segmentation	POS tagging
w_0	w_0t_0
$w_{-1}w_0$	$t_{-1}t_0$
w_0, when len (w_0) = 0	$t_{-2}t_{-1}t_0$
start (w_0) len (w_0)	$t_{-1}w_0$
end (w_0) len (w_0)	$w_{-1}t_0$
end (w_{-1}) start (w_0)	w_0t_0 end (w_{-1})
end (w_0) start (w_1)	w_0t_0 start (w_1)
c_nc_{n+1} (n = 0, len (w_0)-2)	end (w_{-1}) $w_0$start (w_1)t_0, when len(w_0) = 1
start (w_0) end (w_0)	start (w_0)t_0
w_0 start (w_1)	end (w_0)t_0
end (w_{-1}) w_0	c_nt_0 (n = 1,len (w_0)-2)
start (w_0) start (w_1)	start (w_0)c_nt_0 (n = 1, len (w_0)-2)
end (w_{-1}) end (w_0)	end (w_0)c_nt_0 (n = 1, len (w_0)-2)
w_{-1} len (w_0)	$c_nc_{n+1}t_0$ ($c_n = c_n+1$)
len (w_{-1}) w_0	class (start (w_0))t_0
	class (end (w_0))t_0

Character-based model: For CWS and POS tagging problem, the character-based model converts the structure of two layers into one layer, where a tag is combined by a word boundary tag and a POS tag. The word boundary tag denotes the relative position of a character in a word. As described in Ng and Low's study, we use s indicating a single character word and b, m, e indicating the beginning, middle and end character of a word respectively (Ng and Low, 2004). For example, a three-character word with POS tag NN is represented as b_NN m_NN e_NN.

The character-based model is trained using basic character features. In Table 1, c_{-n}, c_n represents the character n positions before or after current character.

Word-based model: The word-based model treats word as basic unit and performs word segmentation and POS tagging in a single perceptron algorithm. We re-implement Zhang's word-based model with same feature set, including word, character and POS tag features (Zhang and Clark, 2008). These features are listed in Table 2.

In Table 2, w, c, t denote the word, character and POS tag separately and start (w), end (w) represent the start

character and end character of current word. w-n, wn are defined similar as c-n, cn in character-based model. Features 15 and 16 in POS tagging are CTBMorph features.

Joint model: Since we have baseline character-based and word-based models, a joint decoding model can be used to improve the performance (Li *et al*., 2011). It employs a multi-beam search strategy, similar with word-based model. The difference is that when adding a character in the sequence to construct a new word, the joint model uses both character-based and word-based model to determine its score of current word, *nscore*.

For the word-based model, the *nscore* is the score of new word and its POS tag. For the character-based model, it is calculated as:

$$\text{nscore} = \sum_{c_i \in \text{word}} s_c(c_i, t_{c_i})$$

were, the nscore is the sum of all characters C_i in the word with its tags t_{ci}. To combine the strength of two baseline models, a log-linear interpolation combination algorithm is used. Adding a parameter α to weight the character-based model and word-based model, the score of the new added word is calculated as:

$$\text{nscore} = \alpha * \sum_{c_i \in \text{word}} s_c(c_i, t_{c_i}) + (1-\alpha) * s_w(w, t)$$

where, the parameter α $(0.0<\alpha<1.0)$ is adjusted according to the performance of the model on development set. The model equals the character-based model when $\alpha = 1.0$ and equals word-based model when $\alpha = 0.0$. The character-based model and word-based model can be trained separately, but used for joint decoding together. The decoding time of joint model is linear with baseline models.

EXPERIMENTS AND ANALYSIS

Previous studies on joint CWS and POS tagging have used Penn Chinese Treebank (CTB) in experiments. In this study, we also perform our experiments on CTB 5.0 (Xue *et al*., 2005), where the distribution of the training, development and test dataset are shown in Table 3. For semi-supervised learning, the unlabeled data is taken from People's Daily newswire. Experimental results will be evaluated on both word segmentation (Seg) and joint CWS and POS tagging (S and T).

Experiments on new agreed data: In section 2, we give the theoretical analysis on the accuracy of new labeled data

Table 3: Distribution of training, development and test dataset

Dataset	Large training	Small training	Devel	Test
Chapter IDs	1-270, 400-931, 1001-1151	1-270	301-325	271 - 300
#sentences	18089	3480	352	348
#words	493939	85105	6821	8008
#POS tags	35	32		

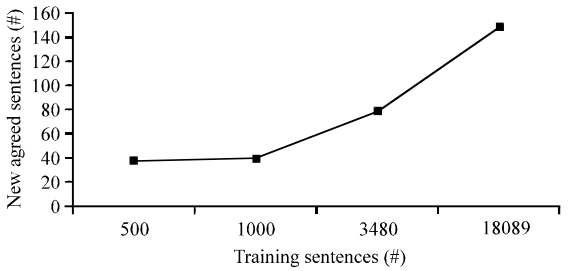

Fig. 2: Numbers of new agreed sentences on development dataset with different sizes of training data

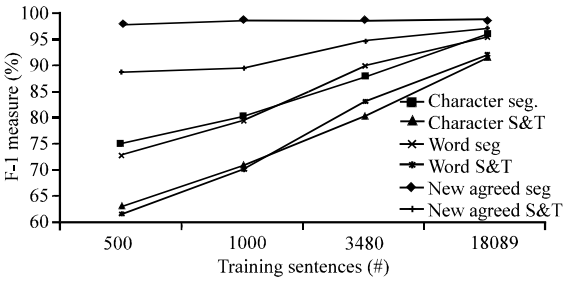

Fig. 3: Performance of baseline models and the accuracy of new agreed sentences on development dataset with different sizes of training data

agreed by two classifiers. In this section, we perform the experiments on CWS and POS tagging problem to validate the analysis. The first experiments are carried on with different sizes of training data. Besides large training dataset and small training dataset, we also trained our baseline models with 500, 1000 sentences. The new agreed sentences are selected by the character-based and word-based models trained with these initial sentences. The curve in Fig. 2 shows the numbers of new agreed sentences on development dataset. Along with more training data, the number of new agreed sentences becomes larger.

Figure 3 shows the performance of character-based and word-based models and the accuracy of new agreed sentences on development dataset generated by these models. In horizontal direction, with the addition of training data, F-1 measure of both models and new agreed sentences increase. In vertical direction, no matter how many sentences are used for training, the accuracy of new

Table 4: Statistics of new agreed sentences on development dataset

	#sentences	Seg F-1	S and T F-1
Character-based	352	96.00	91.60
Word-based	352	95.57	92.03
New agreed sentences			
#words = 0	148	98.63	97.12
#words = 5	120	98.56	96.99
#words = 10	78	98.48	96.85
#words = 15	41	98.45	96.81
#words = 20	27	98.36	96.73
#words = 30	8	99.13	97.97
#words = 40	6	99.45	97.97
#words = 50	2	100.00	97.06

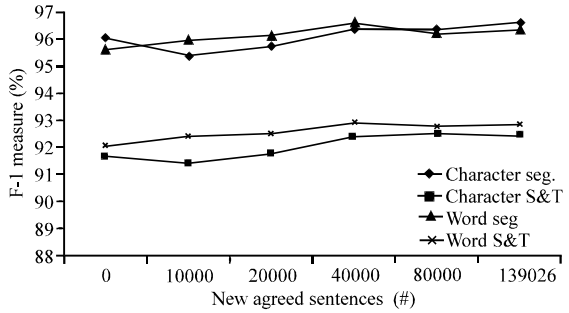

Fig. 4: Results for semi-supervised models on development data with different sizes of new agreed sentences

agreed sentences is always better than the performance of character-based and word-based models.

Table 4 contains the detailed results of models trained with large training data. For Chinese word segmentation, the F-1 measure of agreed new labeled sentences is 98.63 and 2.6% larger than character-based and word-based models, separately 96 and 95.57%. For joint segmentation and POS tagging, F-1 measure of all new labeled sentences is 97.12 and 5% larger than character-based and word-based models, separately 91.6 and 92.03%. The accuracy of new labeled data is also largely better than the accuracy of the best classifier in literature for Chinese word segmentation and POS tagging (Sun and Wan, 2012).

Experiments on development dataset: New agreed sentences are then selected from People's Daily newswire and added to the training dataset for baseline models. Figure 4 shows the results for our semi-supervised approach on large training dataset with different sizes of new agreed sentences. Both character-based and word-based models are evaluated on the development dataset. The number of new agreed data is 0 means the classifiers are trained using only original large training dataset without new agreed sentences.

From Fig. 4, we can see that semi-supervised algorithm improves both character-based and word-based models.

Table 5: Comparison with previous models on test dataset

Algorithm	Seg F1 (%)	S and T F1(%)
Our character semi	97.89	93.65
Our word semi	97.94	93.94
Our joint semi	98.23	94.16
Sun 12	-	94.68
Wang 11	98.11	94.18
Jiang 09	98.23	94.03
Sun 11	98.17	94.02
Zhang 08	97.78	93.67
K 09a	97.79	93.60
K 09b	97.87	93.67
Jiang 08a	97.85	93.41

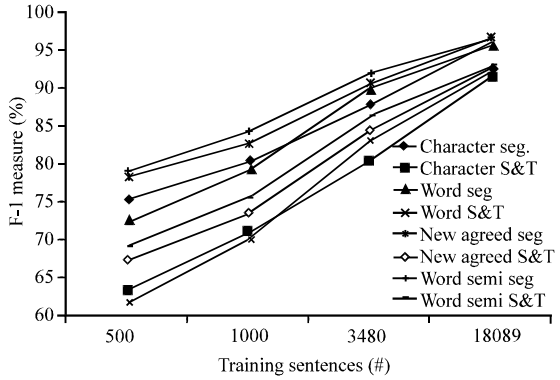

Fig. 5: Results for semi-supervised models on development data with different sizes of training data

For the character-based model, F-1 measure of word segmentation and joint CWS and POS tagging have a little decrease at beginning with only 10000 new agreed sentences, then increases with more sentences and Finally achieve the best with 80000 sentences. For the word-based model, the performance reaches the optimum with 40000 new agreed sentences and then decreases with more additional data. The detailed results in Table 5 show that the improvement of word-based model is larger than the character-based model, which increases 1 percent for word segmentation and 0.88% joint CWS and POS tagging.

The results of baseline and semi-supervised models with different sizes of initial training data on development dataset are shown in Fig. 5. The character-based and word-based semi-supervised models outperform both baseline models in four training data. The word-based model with 500 sentences improve the most, about 6% in word segmentation and 7% in joint CWS and POS tagging.

Then the character-based and word-based semi-supervised can be used in a joint decoding model described in section 3. The parameter á is adjusted to determine the weight of two models depending on their performance on development dataset. We compare our

Table 6: Comparison with previous models on test dataset

Algorithm	Seg F1 (%)	S and T F1(%)
Our character semi	97.89	93.65
Our word semi	97.94	93.94
Our joint semi	98.23	94.16
Sun 12	-	94.68
Wang 11	98.11	94.18
Jiang 09	98.23	94.03
Sun 11	98.17	94.02
Zhang 08	97.78	93.67
K 09a	97.79	93.60
K 09b	97.87	93.67
Jiang 08a	97.85	93.41

semi-supervised joint model with the supervised joint model and also models with self-training, co-training algorithm using both baseline classifiers. The unlabeled data for self-training and co-training methods are same as our semi-supervised model, from People's Daily newswire.

The experimental results of different models are compared. Using the semi-supervised approach, not only character-based and word-based models, but also the joint m odel improves their performance. Self-training and co-training methods also improve the performance, but still much lower than our semi-supervised joint method. Our model achieves 97.25% for word segmentation and 93.49% for joint CWS and POS tagging, outperforming Kruengkrai's word-character hybrid model and Wang's semi-supervised model (Kruengkrai *et al*., 2009; Wang *et al*., 2011).

Experiments on test dataset: We then make a comparison between our model with models on supervised, semi-supervised and domain adaptation approaches in literature. All results are evaluated on the test dataset. Table 6 shows that our semi-supervised character-based and word-based models are better than most state-of-art supervised models except the work of Sun (Sun, 2011) using a sub-word stacking method to combine word segmentation and POS tagging results. Wang introduced a semi-supervised approach exploiting rich statistical features derived from large new labeled data into their models (Wang *et al*., 2011). Jiang presented a strategy that uses a domain adaption method to improve the supervised model (Jiang *et al*., 2009). Our simple semi-supervised approach achieves comparable performance with these both models, lower than Sun's domain adaptation approaches because they employ an annotated corpus. (Sun and Wan, 2012).

CONCLUSION

In this study, we propose a semi-supervised approach for Chinese word segmentation and POS tagging. It selects new unlabeled sentences agreed by two classifiers: character-based and word-based models. The theoretical and experimental analysis shows new agreed labeled sentences have a better accuracy than the performance of both classifiers and are suitable for semi-supervised learning. A joint decoding model is used to combine character-based and word-based semi-supervised models. Experimental results show that our joint semi-supervised method achieves a comparable performance with current semi-supervised methods on large training dataset. In this paper, we studied the semi-supervised approach with no iteration. How the iteration affects the semi-supervised approach and improves the models with small size of training data can be further investigated.

REFERENCES

Abney, S., 2002. Bootstrapping. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA., USA., pp: 360-367.

Blum, A. and T. Mitchell, 1998. Combining labeled and unlabeled data with co-training. Proceedings of the 11th Annual Conference on Computational Learning Theory, July 24-26, 1998, Wisconsin, USA., pp: 92-100.

Clark, S., J.R. Curran and M. Osborne, 2003. Bootstrapping POS taggers using unlabelled data. Proceedings of the 7th Conference on Natural Language Learning at HLT-NAACL, May 31-June 1, 2003, Edmonton, Canada, pp: 49-55.

Collins, M., 2002. Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms. Proceedings of the Conference on Empirical Methods in Natural Language Processing, Volume 10, July 6-7, 2002, Philadelphia, PA., USA., pp: 1-8.

Jiang, W., L. Huang, Q. Liu and Y. Lu, 2008. A cascaded linear model for joint Chinese word segmentation and part-of-speech tagging. Proceedings of the ACL-08: HLT, June 2008, Columbus, Ohio, pp: 897-904.

Jiang, W., L. Huang and Q. Liu, 2009. Automatic adaptation of annotation standards: Chinese word segmentation and POS tagging: A case study. Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and 4th International Joint Conference on Natural Language Processing of the AFNLP, Volume 1, August 2-7, 2009, Suntec, Singapore, pp: 522-530.

Kruengkrai, C., K. Uchimoto, J. Kazamam, Y. Wang, K. Torisawa and H. Isahara, 2009. An error-driven word-character hybrid model for joint Chinese word segmentation and pos tagging. Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and 4th International Joint Conference on Natural Language Processing of the AFNLP, Volume 1, August 2-7, 2009, Suntec, Singapore, pp: 513-521.

Li, X., X. Wang and L. Yao, 2011. Joint decoding for Chinese word segmentation and POS tagging using character-based and word-based discriminative models. Proceedings of the International Conference on Asian Language Processing, November 15-17, 2011, Penang, Malaysia, pp: 11-14.

McClosky, D., E. Charniak and M. Johnson, 2008. When is self-training effective for parsing? Proceedings of the 22nd International Conference on Computational Linguistics, August 18-22, 2008, Manchester, UK., pp: 561-568.

Ng, H.T. and J.K. Low, 2004. Chinese part-of-speech tagging: One-at-a-time or all-at-once? word-based or character-based? Proceedings of the Conference on Empirical Methods in Natural Language Processing, July 27-31, 2004, Barcelona, Spain, pp: 277-284.

Spoustova, D., J. Hajic, J. Raab and M. Spousta, 2009. Semi-supervised training for the averaged perceptron POS tagger. Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, March 30-April 3, 2009, Athens, Greece, pp: 763-771.

Sun, W., 2011. A stacked sub-word model for joint Chinese word segmentation and part-of-speech tagging. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, June 19-24, 2011, Portland, Oregon, USA., pp: 1385-1394.

Sun, W. and X. Wan, 2012. Reducing approximation and estimation errors for Chinese lexical processing with heterogeneous annotations. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, July 8-14, 2012, Jeju Island, Korea, pp: 232-241.

Wang, Y., J. Kazama, Y. Tsuruoka, W. Chen, Y. Zhang and K. Torisawa, 2011. Improving Chinese word segmentation and pos tagging with semi-supervised methods using large auto-analyzed data. Proceedings of 5th International Joint Conference on Natural Language Processing, November 8-13, 2011, Chiang Mai, Thailand, pp: 309-317.

Xue, N., F. Xia, F. Chiou and M. Palmer, 2005. The Penn Chinese Treebank: Phrase structure annotation of a large corpus. Nat. Language Eng., 11: 207-238.

Yarowsky, D., 1995. Unsupervised word sense disambiguation rivaling supervised methods. Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics, June 26-30, 1995, Cambridge, MA., USA., pp: 189-196.

Zhang, Y. and S. Clark, 2008. Joint word segmentation and POS tagging using a single perceptron. Proceedings of the ACL-08: HLT, June 2008, Columbus, Ohio, pp: 888-896.

Zhang, Y. and S. Clark, 2010. A fast decoder for joint word segmentation and POS-tagging using a single discriminative model. Proceedings of the Conference on Empirical Methods in Natural Language Processing, October 9-11, 2010, Cambridge, MA., USA., pp: 843-852.