PAPER

Character-Level Dependency Model for Joint Word Segmentation, POS Tagging, and Dependency Parsing in Chinese

Zhen GUO[†], **Yujie ZHANG**[†a)], **Chen SU**[†], **Jinan XU**[†], ***Nonmembers***, ***and*** **Hitoshi ISAHARA**[††], ***Member***

SUMMARY Recent work on joint word segmentation, POS (Part Of Speech) tagging, and dependency parsing in Chinese has two key problems: the first is that word segmentation based on character and dependency parsing based on word were not combined well in the transition-based framework, and the second is that the joint model suffers from the insufficiency of annotated corpus. In order to resolve the first problem, we propose to transform the traditional word-based dependency tree into character-based dependency tree by using the internal structure of words and then propose a novel character-level joint model for the three tasks. In order to resolve the second problem, we propose a novel semi-supervised joint model for exploiting n-gram feature and dependency subtree feature from partially-annotated corpus. Experimental results on the Chinese Treebank show that our joint model achieved 98.31%, 94.84% and 81.71% for Chinese word segmentation, POS tagging, and dependency parsing, respectively. Our model outperforms the pipeline model of the three tasks by 0.92%, 1.77% and 3.95%, respectively. Particularly, the F1 value of word segmentation and POS tagging achieved the best result compared with those reported until now.

key words: *joint model, Chinese word segmentation and POS tagging, dependency parsing, word internal dependency structure, semi-supervised learning*

1. Introduction

Chinese word segmentation, POS tagging and dependency parsing are the three basic tasks of Chinese natural language processing. In almost all Natural Language Processing (NLP) applications (such as information retrieval, machine translation), Chinese sentences should be processed by the tasks at first. In traditional studies, Chinese word segmentation, POS tagging and dependency parsing are treated independently, as three single task models. In model training of POS tagging and dependency parsing, standard word segmentation data and standard POS tagged data are used, while in practical usage erroneous word segmentation and POS tagging output from the former task are fed into the later task. The single task models therefore have the following defects in practical application. 1) Error propagation between tasks: in practical application, for example, the output of Chinese word segmentation task is the input of the POS tagging task. Errors of Chinese word segmentation will seriously affect the accuracy of part-of-speech tagging. 2) Multi-level features unavailable: in Chinese, some ambiguous POS tags can be resolved by considering long-range syntactic information, while the necessary features are unavailable in traditional POS tagging model.

Manuscript received April 3, 2015.

Manuscript revised August 15, 2015.

Manuscript publicized October 6, 2015.

[†]The authors are with Beijing Jiaotong University, Haidian District, Beijing, 100044 China.

[††]The author is with Toyohashi University of Technology, Toyohashi-shi, 441–8580 Japan.

a) E-mail: yjzhang@bjtu.edu.cn

DOI: 10.1587/transinf.2015EDP7118

Joint model for incorporating multiple tasks is an effective solution to solve the above problems. Because the tasks have strong interactions, many studies have been devoted to the joint model: Chinese word segmentation and POS tagging joint model [1], [2], Chinese POS tagging and dependency joint model [3], [4], Joint model for Chinese word segmentation, POS tagging and word-based dependency parsing [5], Chinese word segmentation, POS tagging and phrase-structure syntactic analysis combination model [6]. A series of studies have shown that the joint model can improve the performance of each task.

The input of the Chinese word segmentation task is character sequence and the input of the POS tagging and syntactic structure analysis is word sequence. Solving the conflict in processing between character and word is the key point of applying joint approach to Chinese word segmentation, POS tagging and dependency parsing. Hatori (2012) [5] presume that in addition to the word-to-word dependency arcs, each word implicitly has inter-character arcs. However, Hatori (2012) [5] did not use the real word internal structures to improve the joint model. Zhang (2013) [6] believe that Chinese Characters play an important role in the Chinese language. Zhang (2013) [6] manually annotate the internal structures of words in Chinese TreeBank 5 (CTB5) and build a character-based phrase-structure analyzer.

The advantage of joint model is that one task can be promoted by other tasks during processing by exploring the available internal results from the other tasks. For building joint model, manually annotated dependency Treebank are required as training data. However, a large amount of dependency Treebank with high quality is not available at present. Since a large-scale raw corpus is easy to obtain and it contains rich knowledge, some researchers used raw corpus to improve the performance of models to be trained [7]–[11]. In the case of joint model, it becomes a new issue how to extract the valuable knowledge from massive raw corpora and then how to combine the knowledge into the complex model.

In solving above problems, this paper made the following contributions.

- We transform the traditional word-based dependency tree into character-based dependency tree by using the internal structure of words and implement a novel character-level joint model for Chinese word segmentation, POS tagging and dependency parsing.

- We propose and implement a novel semi-supervised joint model for exploiting n-gram feature and dependency subtree feature from partially-annotated corpus.
- Experimental results on the Chinese Treebank show that our joint model achieved 98.31%, 94.84% and 81.71% for Chinese word segmentation, POS tagging and dependency parsing, respectively. Our model outperforms the pipeline model of the three tasks by 0.92%, 1.77% and 3.95%, respectively. Particularly, the F1 value of word segmentation and POS tagging achieved the best result compared with those reported until now.

The remainder of this paper is divided as follows: Section 2 describes the character-level dependency model for joint word segmentation, POS tagging and dependency parsing in Chinese. Section 3 describes a semi-supervised joint model by exploiting n-gram feature and dependency subtree feature from partially-annotated corpus. Section 4 presents our experimental results. Section 5 concludes with ideas for future research.

2. Character-Level Dependency Model for Joint Word Segmentation, POS Tagging and Dependency Parsing in Chinese

Beam-search and global models have been applied to transition-based NLP framework, leading to state-of-the-art accuracies that are comparable to other best strategies. We transform the traditional word-based dependency tree into character-based dependency tree using the internal structure of words annotated by Zhang (2013) [6] and then realize a novel character-level joint model for the three tasks. We use the online perceptron algorithm with early-update for global learning and beam search algorithm for decoding [13].

2.1 Character-Based Dependency Tree

Chinese characters are ideographic characters and play an important role in Chinese word. Chinese characters, consisting of one Chinese word, usually have syntactic structures. For example, the characters “理 (cut)” and “发 (hair)” in the Chinese word “理发 (cut hair)” form a verb-object, and modify the character “店 (shop)” to form the word “理发店 (The barber shop)”. The structure of the word between characters is similar to the structure of the sentence between words.

Zhang (2013) [6] manually annotate the structures of words that occur in the CTB5. An example is shown in Fig. 1. “l”, “r” and “c” are used to indicate the “left”, “right” and “coordination” head directions, respectively. We transform the structure of Fig. 1 (b) and Fig. 1 (d) into dependency structure in Fig. 1 (c) and Fig. 1 (e). For the structure of “c”, we select the right part as head node. We transform

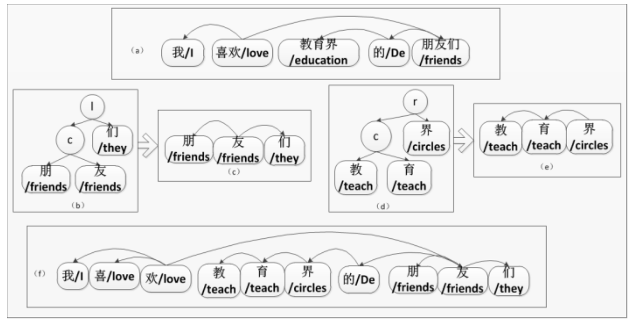


Fig. 1 Intra-word structure and character-based dependency tree

the traditional word-based dependency tree in Fig. 1 (a) into the character-based dependency tree in Fig. 1 (f) through this way.

Character-based dependency tree is more suitable for Transition-based joint model than word-based dependency tree owing to the following factors. First, we do not need a specially designed state alignment scheme for joint model. For a sentence with N character, the number of dependence arc is N-1 for character-based dependency tree, while the number of dependent arc is not fixed for word-based dependency tree. The joint model, therefore, can easily compare states with same number of characters and same number of arcs. Second, the internal structure of words is a useful source of information and some recent work has started to exploit this information. Li (2011) [15] studied the morphological structures of Chinese words, showing that 35% percent of the words in CTB5 can be treated as having morphemes and improve the performance of Chinese dependency parsing with such morphology. Zhang (2013) [6] demonstrate that both the annotated word structures and flat word structures, bring about improvement in Phrase structure syntactic analysis.

2.2 Transition Actions

In a transition-based system, an input sentence is processed in a linear left-to-right pass, and the output is constructed by a state-transition process. After an action is performed, the system transfers the current state T_i to the next one T_{i+1}. A state T consists of a stack S and a queue Q, where $S = \{\ldots S1, S0\}$ contains partially constructed parse trees, and $Q = \{Q0, Q1 \ldots\}$ is the sequence of input characters that have not been processed.

This paper redesign the following transition actions for combining the Chinese word segmentation, POS tagging and dependency parsing into a joint system, through which the features used in previous research can also be integrated into the model conveniently. The candidate transition actions at each step are defined as follows.

1) Transition action for Chinese word segmentation and POS tagging:

① SHIFT-B(t): Push the head character of Q onto S as first character of a word and assign POS tag t.

② SHIFT-M: Push the head character of Q onto S as middle character of a word.

③ SHIFT-E: Push the head character of Q onto S as tail character of a word.

④ SHIFT-S(t): Push the head character of Q onto S as single word and assign POS tag t.

Since there are 33 POS tags in CTB5, 33 possible actions are formed for SHIFT-B(t) and SHIFT-S(t) respectively.

Through the above design, we combine the transition-based and sequence labeling Chinese word segmentation frameworks and the features used in previous research can conveniently be integrated into the new joint model. The Semi-supervised joint model proposed in Sect. 3 makes full use of it.

2) Transition actions for intra-word dependencies

① REDUCE-SUBLEFT: Pop the top two nodes S1 and S0 off S and then push a subtree S1←S0. Note that S0 and S1 belong to the same word.

② REDUCE-SUBRIGHT: Pop the top two nodes S1 and S0 off S and then push a subtree S1→S0. Note that S0 and S1 belong to the same word.

The building-up of the internal dependency relations within a word is similar with those between words. The difference is the type of relative elements. The element of the former is a character but the one of the latter is a word.

3) Transition actions for inter-word dependencies

① REDUCE-LEFT: Pop the top two nodes S1 and S0 off S and then push a subtree S1←S0. Note that S0 and S1 belong to different words already formed, implying that their intra-word structure analysis have been finished.

② REDUCE-RIGHT: Pop the top two nodes S1 and S0 off S and then push a subtree S1→S0. Note that S0 and S1 belong to different words already formed, implying that their intra-word structure analysis have been finished.

Based on the above transfer strategy, there are 72 actions in total, and a sentence with N characters requires 2N-1 transitions from the initial state to the terminal state.

We illustrate the procedure of the joint model in Table 1, using the example in Fig. 1. As shown in Table 1, Transition 0 is initial state, S = Φ, Q = {我 (I), 喜 (love), 欢 (love), · · ·}. In Transition 1, Transition action is SHIFT-S(PN), pushing the head character "我 (I)" of Q onto S as single word and assigning POS tag PN, then S = {我 (I)/PN} and Q = {喜 (love), 欢 (love), · · ·}. In Transition 2, the head character "喜 (love)" of Q is pushed into S as first character of a word with POS tag VV. In Transition 3, the head character "欢 (love)" of Q is pushed into S as tail character of a word with POS tag VV. At this point, "喜 (love) 欢 (love)" is segmented as a word and tagged as VV. In Transition 4, Transition action is REDUCE-SUBLEFT, popping the top two nodes "喜 (love)" and "欢 (love)" off S and then pushing a subtree "喜 (love)"←"欢 (love)". At this point, a dependency arc between two characters, "喜 (love)" and "欢 (love)", is formed. In transition 5, Transition action is REDUCE-LEFT, pop the top two nodes "我 (I)" and "欢 (love) (喜欢 (love))" off S and then push a subtree "我 (I)"←"欢 (love) (喜欢 (love))". Note that S contains partially constructed dependency trees. At this point, a dependency arc between two words, "我 (I)" and "喜欢 (love)", is formed.

Table 1 Illustration of the procedure of the joint model for the sentence in Fig. 1

Transition number	Transition action	Stack	Queue	Dependency arc
0	-	Φ	我 喜...	
1	SHIFT-S(PN)	我/PN	喜 欢...	
2	SHIFT -B(VV)	我/PN 喜/VV	欢 教...	
3	SHIFT -E(VV)	...喜/VV 欢/VV	教 育...	
4	REDUCE-SUBLEFT	我/PN 欢/VV	教 育...	喜←欢
5	REDUCE-LEFT	欢/VV	教 育...	我←欢(喜欢)
6	SHIFT -B(NN)	欢/VV 教/NN	育 界...	
7	SHIFT -M(NN)	…教/NN 育/NN	界 的...	
8	REDUCE-SUBLEFT	欢/VV 育/NN	界 的...	教←育
9	SHIFT -E(NN)	…育/NN 界/NN	的 朋...	
10	REDUCE-SUBLEFT	欢/VV 界/NN	的 朋...	育(教育)←界
11	SHIFT-S(DEG)	…界/NN 的/DEG	朋 友...	
12	REDUCE-LEFT	欢/VV 的/DEG	朋 友...	界(教育界)←的
13	SHIFT -B(NN)	…的/DEG 朋/NN	友 们	
14	SHIFT -M(NN)	…朋/NN 友/NN	们	
15	REDUCE-SUBLEFT	…的/DEG 友/NN	们	朋←友
16	SHIFT -E(NN)	…友/NN 们/NN	Φ	
17	REDUCE-SUBRIGHT	…的/DEG 友/NN	Φ	友(朋友)→们
18	REDUCE-LEFT	欢/VV 友/NN	Φ	的←友(朋友们)
19	REDUCE-RIGHT	欢/VV	Φ	欢(喜欢)→友(朋友们)

2.3 Feature Templates

Feature templates used in this paper are shown in Table 2. The feature set consists of two categories: (1) structure features, which encode the syntactic structure information; (2) string features, which encode the information of neighboring characters and words. Syntactic structure information includes word-based structure information and character-based structure information.

We use the features proposed by Hatori (2012) [5] and make adjustment on the features, including usage phase and usage way, in order to adapt to the proposed joint model. P01-P20 are syntactic structure features and "w" represents sub-word or word, according to the situation. W01–W20 are used to determine whether to segment at the current position or not. Because T01–T05 are used to determine the POS tag of the word being shifted, they are only applied for SHIFT-S(t) and SHIFT-B(t). C01-C08 are newly introduced word internal structure features in this paper.

Table 2 Feature template for joint model

Id	Feature
P01,P02,P03,P04	S0.w S0.t S1.w S1.t
P05,P06	S0.w-S0.t S1.w-S1.t
P07,P08	S0.w-S1.w S0.t-S1.t
P09	S0.w-S0.t-S1.t
P10	S0.t-S1.w-S1.t
P11	S0.w-S1.w-S1.t
P12	S0.w-S0.t-S1.w
P13	S0.w-S0.t-S1.w-S1.t
P14	S1.t- S1.rc.t-S0.t
P15	S1.t- S1.lc.t-S0.t
P16	S1.t- S1.rc.t-S0.w
P17	S1.t- S1.lc.t-S0.w
P18	S1.t-S0.t-S0.rc.t
P19	S1.t-S0.w-S0.lc.t
P20	S2.t- S1.t-S0.t
T01,T02	Q-1.t-T0 Q-1.w-T0
T03,T04	Q-2.t-Q-1.t-T0 C0-T0
T05	Q-1.t-Q-1.e-C0-T0
W01,W02	Q-1.w Q-2.w-Q-1.w
W03	Q-1.w(for single-char word)
W04	Q-1.b-len(Q-1.w)
W05	Q-1.e-len(Q-1.w)
W06,W07	Q-1.e-C0 Q-1.b-Q-1.e
W08,W09	Q-1.w-C0 Q-2.e-Q-1.w
W10,W11	Q-1.b-C0 Q-2.e-Q-1.e
W12	Q-2.w-len(Q-1.w)
W13	len(Q-2.w)-Q-1.w
W14,W15	Q-1.w-Q-1.t Q-2.t-Q-1.w
W16	Q-1.t-Q-1.w-Q-2.e
W17	Q-1.t-Q-1.w-C0
W18,W19	Q-1.t-Q-1.e Q-1.t-C0-C1
W20	Q-1.t-Q-1.-C(C∈Q-1.w|e)
C01,C02,C03,C04	S0.c S0.ct S1.c S1.ct
C05,C06	S0.lc.ct S0.rc.ct
C07,C08	S1.lc.ct S1.rc.ct

Notes: S0 and S1 are top two nodes on stack. Q-1 and Q-2 respectively denote the last-shifted word and the word shifted before Q-1. w denotes word form, t denotes POS tag and c denotes character, and Q-1.b/e denotes the beginning/ending characters of Q-1.w. lc and rc are the left most child and right most child. C0 and T0 denote the current character and POS tag. len() denotes the length of the word. P01-P13 are used in all actions. P14-P20 are used for all the actions except "Shfit_M" and "Shfit_E". W01-W20 are used in actions for Chinese word segmentation. T01-T05 are used in “SHIFT_B” and “SHIFT_S”. C01-C07 are used in actions for intra-word arc building.

3. Semi-Supervised Joint Model for Chinese Word Segmentation, POS Tagging and Dependency Parsing

Semi-supervised model is widely used in various natural language processing tasks and has obtained a significant effect, especially in the field where supervised learning is lack of manually annotated data. For Chinese word segmentation, POS tagging and dependency parsing, the training corpus is deeply manually annotated data. However, a large amount of dependency Treebank with high quality is not available at present. Meanwhile, raw data and word segmented data are easy to obtain. In joint model, one task can be promoted by other tasks during processing by exploring the available internal results from the other tasks. Thus, semi-supervised approach can achieve greater performance than single task model. In the case of joint model, it becomes a new issue how to extract the valuable knowledge from massive raw corpora and then how to combine the knowledge into the complex model. We implement a semi-supervised joint model by exploiting n-gram feature and dependency subtree feature from partially-annotated corpus.

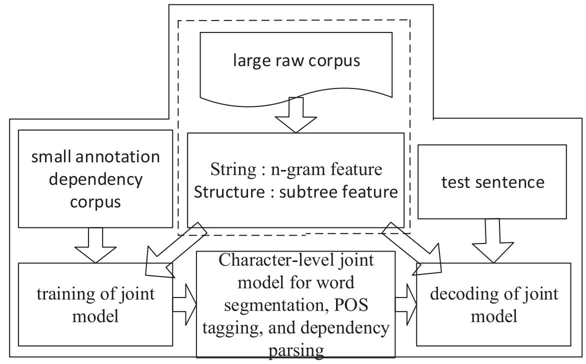


Fig. 2 Framework of semi-supervised joint model

The framework of semi-supervised joint model is shown in Fig. 2.

3.1 String Feature: N-Gram Feature

This section describes the extraction and use of n-gram for joint model. In this paper, we extract features from word segmentation annotated sentences following the strategy of Wang (2011) [7].

For a sentence given segmentation result, we can give each character C_i a tag T_i according to its position in the word it belongs to. In other words, a segmented sentence is a sequence $\{(C_0, T_0) \ldots, (C_n, T_n)\}$. Let g be a character n-gram (e.g., uni-gram C_i, bi-gram C_iC_{i+1}, trigram $C_{i-1}C_iC_{i+1}$ and so on), and seg be a segmentation profile for n-gram g observed at each position (e.g., T_i, T_iT_{i+1}, T_{i+1} and so on). Then, we can extract a list of (g, seg) from the segmented data. Next, we take statistics of the cases where n-gram g is segmented with the segmentation profile seg f(g, seg). In order to alleviate the sparseness of the data, we group all the (g, seg) into three sets: high-frequency (H), middle-frequency (M), and low-frequency (L). The grouping way are defined as follows: if the f(g, seg) is one of the top 10% of all the f(g, seg), the label of (g, seg) is represented as H; if it is between top 10% and 30%, it is represented as M, otherwise it is represented as L. In this way, n-gram {g, seg, and label} lists are produced. For the n-gram that does not exist in the list, we assign label none(N) to it.

When transition actions for Chinese word segmentation and POS tagging are being formed, i.e. when the head character of Q is being pushed onto S, we extract the n-gram string about the character and get label from the {g, seg, label} list described above. We combine the n-gram string and the label as the n-gram feature of the character. We only use the bi-gram string C_iC_{i+1} with seg = T_i for feature extraction in this paper.

3.2 Structure Feature: Dependency Subtree Feature

Dependency subtree features are generated from automatically parsed data. The most widely used subtree features are those subtrees containing two or three words. We only use

Table 3 Feature template added for semi-supervised joint model

Id	Feature
B01	f(C0-C1, seg)
S01	R-f(S0.w, S1.w, R)
S02	L-f(S0.w, S1.w, L)

Notes: C1 denotes the second character in the queue. C0, S0, S1 and w are same as Table2. f() denotes the label for frequency.

the subtrees containing two words for feature generation in this paper.

First, we use the open source graph-based dependency parser MSTParser† to parse large-scale raw data. In this paper, the raw data is partially annotated data, i.e. manually annotated with word segmentation and POS. For this reason, we call it as semi-supervised. Then we extract subtrees containing two words from the parsed dependency trees such as W1-W2-R/L. The order of W1 and W2 corresponds to the sequence of them in the original sentence, "R" and "L" indicating "right" and "left" head directions respectively. The frequency of W1-W2-R/L can be obtained from the automatically parsed data. We then group the subtrees into three sets corresponding to three levels of frequency: "high (H)", "middle (M)", and "low (L)". H, M, and L are used as set labels for the three sets. The following are the settings: if a subtree is one of the TOP-10% most frequent subtrees, it is in set H; else if a subtree is one of the TOP-30% subtrees, it is in set M; else it is in set L. In this way, subtree {W1-W2, R/L, label} lists are produced. For the subtree that does not appear in the parsed data, we assign label none(N) to it.

When we judge whether the top two nodes S1 and S0 on stack have dependency relationship, we get labels for all kinds of subtree between S1 and S0 as features using the subtree list obtained above.

The newly added feature templates are displayed in Table 3. We train joint model using feature templates both in Table 2 and in Table 3.

4. Experimental Results and Analyses

4.1 Data

We used Chinese Tree Bank (CTB5) as annotated corpus, and it was separated into several parts: Training data set (chapter: 1-270, 400-931 and 1001-1151), development data set (chapter 301-325) and test data set (chapter 271-300) [8]. As the names described, training data was used for training joint model, development data used for tuning parameters, and test data used for evaluation. We adopted Penn2Malt†† to transfer phrase structure tree to dependency tree. The word segmentation annotated People's Daily corpus (the first half of 1998 year) was regard as partially-annotated corpus†††, from which we extract N-gram feature and dependency subtree. We used a Conditional Random Field (CRF) –based POS tagger to conduct POS tagging on People's Daily corpus's and then parsed the data with graph-based dependency analyzer MSTParser (http://mstparser.sourceforge.net).

In this paper, we used F-score as the accuracy metric to measure the performance of Chinese word segmentation, POS tagging and dependency parsing. $F = 2PR/(P + R)$, where P is precision and R is recall. Note that a dependency relationship is correct only when the two related words are all recalled in word segmentation and the head direction is correct. Following conventions, the relationships containing any punctuation are ignored.

4.2 Comparison of Different Models

We implement a Character-level dependency model for joint word segmentation, POS tagging and dependency parsing and three semi-supervised joint model. The beam size is set as 64 in this paper. For comparison, we implement two pipeline systems as baseline

- SegTagDep: Character-level dependency model for joint word segmentation, POS tagging and dependency parsing
- SegTagDep+2-gram: Semi-supervised joint model with 2-gram feature described in Sect. 3.1
- SegTagDep+subtree: Semi-supervised joint model with dependency subtree feature described in Sect. 3.2
- SegTagDep+2-gram+subtree: Semi-supervised joint model with 2-gram feature and dependency subtree feature.
- CRF+MSTP: CRF-based Chinese word segmentation and POS tagging system, where the open source tool CRF++†††† is used and the feature templates are the same with Wang (2011) [8]. MSTP indicates open source dependency parser MSTParser, where the second order model is used for training and decoding.
- SegTag+MSTP: Joint model for Chinese word segmentation and POS tagging by Zhang (2010) [1] and Open source dependency parser MSTParser mentioned above.

4.3 Tables and Figures

The F1 scores of six systems' evaluation results are described in Table 4. The first line is the performance of our proposed joint model. According to Table 4, our joint model achieved better performance in all tasks than baselines. Compared with the pipeline system, CRF+MSTP, our joint model achieved higher F-score than it with 0.13%, 0.86% and 1.79% in word segmentation, POS tagging and dependency parsing respectively, demonstrating the effectiveness of the joint model. Compared with the partially joint model, SegTag+MSTP, our joint model achieved higher F-score than it with 0.68% and 1.66% in POS tagging and dependency parsing respectively. The improvement on POS tagging is due to the introduction of syntactic features since

†http://mstparser.sourceforge.net
††http://w3.msi.vxu.se/~nivre/research/Penn2Malt.html
†††http://www.icl.pku.edu.
††††http://crfpp.sourceforge.net/

Table 4 Comparison results of the different models on CTB5

Model	SEG	POS	DEP
SegTagDep	97.52	93.93	79.55
SegTagDep+2-gram	**98.38**	94.63	80.78
SegTagDep+subtree	97.74	94.25	80.40
SegTagDep+2-gram+subtree	98.31	**94.84**	**81.71**
CRF+MSTP	97.39	93.07	77.76
SegTag+MSTP	97.51	93.25	77.89
Hatori（2012）	98.26	94.64	--

Table 5 Comparison results of word segmentation and POS tagging

Model	SEG	POS
Kruengkrai09	97.87	93.67
Zhang10	97.78	93.67
Sun11	98.17	94.02
Wang11	98.11	94.18
Hatori12	98.26	94.64
Zhang13	97.84	94.80
Our	**98.31**	**94.84**

some ambiguities of POS tagging may be resolved by syntactic structure but hardly be resolved by character features. Meanwhile, the performance of dependency parsing also gets better as the F-score of POS tagging gets higher. However, as the high F-score of word segmentation, the performance of word segmentation in the proposed method is nearly same as baseline systems. The semi-supervised model with more untagged corpus will prove this suppose.

We now observe the effect of semi-supervised joint model. The results show that both n-gram feature and subtree feature contributed to the improvement in performance from the view of each metric, and that the combination of n-gram feature and subtree feature achieved further improvements on POS tagging and dependency parsing. The semi-supervised joint model, SegTagDep+2-gram+subree, achieved 98.31%, 94.84% and 81.71% for Chinese word segmentation, POS tagging and dependency parsing, respectively. Our model outperforms the pipeline system, CRF+MSTP on the three tasks by 0.92%, 1.77% and 3.95%, respectively. Particularly, the F1 value of word segmentation and POS tagging achieved the best result compared with those reported until now.

Statistical significances were tested by McNemar's Test ($p < 0.01$) for the above comparison results between our joint models and pipeline systems.

In Table 4, we also list the results from Hatori [5], where on the same test data set only results of word segmentation and POS tagging were reported.

Table 5 shows a comparison of the word segmentation and POS tagging accuracies with other state-of-the-art models. "Kruengkrai09" is a lattice-based model by Kruengkrai (2009) [2]. "Zhang10" is the incremental model by Zhang (2010) [1]. These two systems use no external resources other than the CTB corpora. "Sun11" is a CRF-based model by Sun (2011) [18] in which several models are combined and a idiom dictionary is used. "Wang11" is a semi-supervised model by Wang (2011) [8], where the Chinese Gigaword Corpus is used. "Hatori12" is a joint model by Hatori (2012) [5], where intra-word dependencies are not considered but a rich dictionary is used. "Zhang13" is a joint model for Chinese word segmentation, POS tagging and phrase-structure syntactic analyzer by Zhang (2013) [6]. Our model achieved the best accuracies compared with other models. The differences in accuracy between our model and "Zhang13", suggest that dependency parsing, in contrast to phrase structure parsing, tends to bring semantically related elements together and is better suited to lexicalized models.

We calculated the average processing time in sec. per sentence for our models. The evaluation result showed that 1.83, 1.89, 1.91 and 1.96 for SegTagDep, SegTagDep+2-gram, SegTagDep+subtree and SegTagDep+2-gram+subtree, respectively.

5. Conclusions

We transform the traditional word-based dependency tree into character-based dependency tree by using the internal structure of words and then implement a novel character-level joint model for Chinese word segmentation, POS tagging and dependency parsing. For Chinese word segmentation, we design 4 transition actions: Shfit_S, Shift_B, Shift_M and Shift_E, through which the features used in previous research, can be integrated into the model. We implement a novel semi-supervised joint model for exploiting n-gram feature and dependency subtree feature from partially-annotated corpus. Experimental results on the Chinese Treebank show that our joint model achieved 98.31%, 94.84% and 81.71% for Chinese word segmentation, POS tagging and dependency parsing, respectively. Our model outperforms the pipeline model of the three tasks by 0.92%, 1.77% and 3.95%, respectively. Particularly, the F1 value of word segmentation and POS tagging achieved the best result compared with those reported until now.

Acknowledgments

This work is sponsored by the International Science & Technology Cooperation Program of China under grant No. 2014DFA11350.

References

[1] Y. Zhang and S. Clark, "A fast decoder for joint word segmentation and POS-tagging using a single discriminative model," Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, pp.843–852, 2010.

[2] C. Kruengkrai, K. Uchimoto, J. Kazama, Y. Wang, K. Torisawa, and H. Isahara, "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 the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1, Association for Computational Linguistics, pp.513–521, 2009.

[3] J. Hatori, T. Matsuzaki, Y. Miyao, et al., "Incremental Joint POS Tagging and Dependency Parsing in Chinese," IJCNLP, pp.1216–1224, 2011.

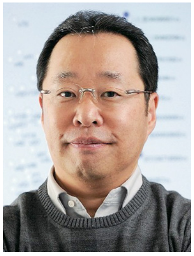
[4] Z. Li, M. Zhang, W. Che, et al., "Joint models for Chinese POS tagging and dependency parsing," Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, pp.1180–1191, 2011.
[5] J. Hatori, T. Matsuzaki, Y. Miyao, et al., "Incremental joint approach to word segmentation, pos tagging, and dependency parsing in chinese," Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1, Association for Computational Linguistics, pp.1045–1053, 2012.
[6] M. Zhang, Y. Zhang, W. Che, et al., Chinese parsing exploiting characters, 51st Annual Meeting of the Association for Computational Linguistics, 2013.
[7] Z. Guo, Y. Zhang, C. Su, and J. Xu, "Exploration of N-gram Features for the Domain Adaptation of Chinese Word Segmentation," Natural Language Processing and Chinese Computing, Springer Berlin Heidelberg, vol.333, pp.121–131, 2012.
[8] Y. Wang, Jun'ichi Y.T. Kazama, Y. Tsuruoka, et al., "Improving Chinese Word Segmentation and POS Tagging with Semi-supervised Methods Using Large Auto-Analyzed Data," IJCNLP, pp.309–317, 2011.
[9] T. Koo, X. Carreras, and M. Collins, "Simple semi-supervised dependency parsing," 46th Annual Meeting of the Association for Computational Linguistics, pp.595–603, 2008.
[10] W. Chen, J. Kazama, K. Uchimoto, and K. Torisawa, "Improving dependency parsing with subtrees from auto-parsed data," Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2, Association for Computational Linguistics, pp.570–579, 2009.
[11] W. Chen, J. Kazama, and K. Torisawa, "Bitext dependency parsing with bilingual subtree constraints," Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, pp.21–29, 2010.
[12] Y. Zhang and J. Nivre, "Analyzing the Effect of Global Learning and Beam-Search on Transition-Based Dependency Parsing," COLING (Posters), pp.1391–1400, 2012.
[13] M. Collins and B. Roark, "Incremental parsing with the perceptron algorithm," Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, p.111, 2004.
[14] M. Zhu, Y. Zhang, W. Chen, et al., Fast and Accurate Shift-Reduce Constituent Parsing.
[15] Z. Li and G. Zhou, "Unified dependency parsing of Chinese morphological and syntactic structures," Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Association for Computational Linguistics, pp.1445–1454, 2012.
[16] H. Zhao, C.N. Huang, M. Li, et al., "Effective tag set selection in Chinese word segmentation via conditional random field modeling, Proceedings of PACLIC, 20, pp.87–94, 2006.
[17] R. McDonald, K. Crammer, and F. Pereira, "Online large-margin training of dependency parsers," Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, pp.91–98, 2005.
[18] W. Sun, "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-Volume 1, Association for Computational Linguistics, pp.1385–1394, 2011.
[19] M. Zhang, Y. Zhang, W. Che, and T. Liu, "Character-Level Chinese Dependency Parsing," Association for Computational Linguistics, pp.1326–1336, 2014.

Zhen Guo is a MS student in School of Computer information and technology at Beijing Jiaotong University, China. His research focuses on Chinese dependency parsing and natural language processing.

Yujie Zhang is Professor of Computer and Information Technology at Beijing Jiaotong University. She received the B.Eng. degree from Beijing Jiaotong University, the M.Eng. degree from the Institute of Computing Technology, Chinese Academy of Science, and the Ph.D. degree from the University of Electro-Communications. With main research interests in natural language processing, machine translation, big text data processing, she has published more than 30 papers including over 10 papers in international journals such as a variety of IEEE and Natural Language Engineering. Prof. ZHANG received the first prize of National Science and Technology Progress Award. She served on the editorial board of numerous journals and several conferences, such as Journal of Chinese Information Processing, the IEICE Transactions, MT Summit 2011, Coling 2014.

Chen Su is a MS student in School of Computer information and technology at Beijing Jiaotong University, China. His research focuses on machine translation and natural language processing.

Jinan Xu is currently an Associate Professor in School of Computer Science and information technology,Beijing Jiaotong University, Beijing, China. He received the B.Eng. degree from Beijing Jiaotong University in 1992, the MS degree and Ph.D. degree in computer information from Hokkaido University, Sapporo, Japan, in 2003 and 2006, respectively. His research focuses on natural language processing, machine translation,information retrieve, text mining, and machine learning. He is a member of CCF, CIPSC, ACL and the ACM.



Hitoshi Isahara is Professor of Toyohashi University of Technology. He received the B.E., M.E., and Ph.D. degrees in electrical engineering from Kyoto University, Kyoto, Japan, in 1978, 1980, and 1995, respectively. His research interests include natural language processing, machine translation and lexical semantics. He was working at the Electrotechnical Laboratory of Japanese Ministry of International Trade and Industry from 1980 to 1995, and was working at the Communications Research Laboratory of Japanese Ministry of Posts and Telecommunications (currently, National Institute of Information and Communications Technology) from 1995 to 2009. He is a Professor and Director of the Information and Media Center at Toyohashi University of Technology from January 2010. He was the President of the International Association for Machine Translation (IAMT) from 2009 to 2011, and the President of the Asia-Pacific Association for Machine Translation (AAMT) from 2006 to 2012. He is a board member of Non-profit organization Gengo-Shigen-Kyokai (GSK) (literally "Language Resources Association") in Japan. He is a member of the Japanese national committee of ISO/TC37 (Terminology and Language Resource Management and its Application) and chair its SC3.