Granular Syntax Processing with Multi-task and Curriculum Learning

Xulang Zhang, Rui Mao, Erik Cambria*

School of Computer Science and Engineering, Nanyang Technological University, Singapore.

*Corresponding author(s). E-mail(s): cambria@ntu.edu.sg;
Contributing authors: xulang001@e.ntu.edu.sg; rui.mao@ntu.edu.sg;

Abstract

Syntactic processing techniques are the foundation of Natural Language Processing (NLP), supporting many downstream NLP tasks. In this paper, we conduct pair-wise Multi-Task Learning (MTL) on syntactic tasks with different granularity, namely Sentence Boundary Detection (SBD), text chunking, and Part-of-Speech(PoS) tagging, so as to investigate the extent to which they complement each other. We propose a novel soft parameter sharing mechanism to share local and global dependency information that is learned from both target tasks. We also propose a Curriculum Learning (CL) mechanism to improve MTL with non-parallel labeled data. Using non-parallel labeled data in MTL is a common practice, whereas it has not received enough attention before. For example, our employed PoS tagging data do not have text chunking labels. When learning PoS tagging and text chunking together, the proposed CL mechanism aims to select complementary samples from the two tasks to update the parameters of the MTL model in the same training batch. Such a method yields better performance and learning stability. We conclude that the fine-grained tasks can provide complementary features to coarse-grained ones, while the most coarse-grained task, SBD, provides useful information for the most fine-grained one, PoS tagging. Additionally, the text chunking task achieves state-of-the-art performance when joint learning with PoS tagging. Our analytical experiments also show the effectiveness of the proposed soft parameter sharing and CL mechanisms.

Keywords: Text chunking, Part-of-speech tagging, Sentence Boundary Detection, Multi-task learning, Granularity computing, Curriculum learning

PoS label:	PRP	VBD	PRP$	NNS	.
Chunk label:	S-NP	S-VP	B-NP	E-NP	O
Sentence Boundary label:	O	O	O	PERIOD	O
Example sentence:	I	received	her	flowers	.

Fig. 1 An example that showcases the PoS, chunk, sentence boundary labels of a given sentence. The PoS tagging labels employ the Penn Treebank [27] annotation schema. The chunk labels employ BIOES annotation schema [28], where "B" represents the beginning of a chunk that immediately follows another chunk; "I" represents the word is inside a chunk; "O" represents outside of any chunk; "E" represents the ending word of a chunk; "S" represents a chunk phrase that contains only a single token. The sentence boundary word is labeled as PERIOD, following the tagging schema of the IWSLT dataset [29]. Note that there are no punctuation marks in the IWSLT dataset.

1 Introduction

Syntactic processing is a generalization of natural language processing (NLP) subtasks that are concerned with the structure of phrases and sentences, as well as the relation of words to each other within the phrase or sentence [1]. There is a multitude in the granularity of syntactic processing. For instance, Sentence Boundary Detection (SBD), text chunking, and Part-of-Speech (PoS) tagging are all fundamental syntactic tasks, ranging from coarse-grained to fine-grained. The interplay between these tasks ensures a granular understanding of the syntactic and structural aspects of natural language, enabling more sophisticated language processing applications [2]. SBD aims to distinguish where sentences begin and end in raw texts. Downstream tasks such as machine translation [3, 4], information retrieval [5], and document summarization [6, 7] rely on predetermined sentence boundaries for good performance. In sentiment analysis, SBD can help identify negation scope to improve the performance [8]. Text chunking splits sentences into non-overlapping segments, such as Noun Phrase (NP) and Verb Phrase (VP). It helps to understand a sentence structure and the relation between words, e.g., recognizing names and syntactic components. It supports Natural Language Processing (NLP) tasks that require a general understanding of sentence components, such as text summarization [9] and sentiment analysis [10]. PoS tagging aims to label each word in a given text with its PoS tag, e.g., noun, verb, adjective, adverb, etc. It parses input text to assist downstream tasks, including syntactic tasks, e.g., text chunking [11, 12] and dependency parsing [13, 14], as well as high-level NLP tasks, e.g., information retrieval [15] and sentiment analysis [16, 17], and metaphor interpretation [18–20]. These three tasks are all commonly regarded as a sequence labeling problem. Figure 1 shows an example of the different task labels given an input sentence.

Although current works have achieved very high accuracy on these tasks [21–23], these fundamental tasks are still worth investigating for the improvement of downstream applications. However, there has been limited research on how syntactic tasks of different granularity affect each other. In traditional feature-engineering-based approaches, PoS tags are commonly used as input features for coarser-grained syntactic tasks including SBD and text chunking [24]. Modern neural-network-based techniques such as multi-task learning (MTL) [11, 25] and transfer learning [26] also show that PoS tagging is a complimentary task for text chunking, but the reverse is inconclusive.

In this work, we conduct pair-wise MTL on SBD, text chunking, and PoS tagging respectively to study the correlations between task granularity and the complementary effect to each other. The adoption MTL is motivated by the advantage of joint learning to mitigate the error propagation problem [30]. We propose an effective local and global dependency sharing (LGDS) mechanism. This is inspired by the finding in MTL that soft parameter sharing allows task-specific towers to absorb useful features that are learned from their neighbour towers [31, 32].

The employed PoS tagging dataset Wall Street Journal (WSJ) [27], text chunking dataset CoNLL 2000 (CoNLL00) [33], and SBD dataset IWSLT [29] have different labels. In the case of CoNLL00, chunking, PoS, and sentence boundary labels are present. Thus, an MTL model can be trained with parallel labeled data. That is, given an input with corresponding sets of ground-truth labels for the two tasks, the model can update the parameters of the task-specific towers simultaneously. On the other hand, the WSJ dataset lacks annotated chunk labels, and the IWSLT dataset is limited to sentence boundary labels. This poses a challenge for employing MTL. For instance, in MTL for chunking and PoS tagging, the model, when presented with a WSJ training sample, cannot simultaneously update both task towers. This scenario is termed training with non-parallel labeled data in our MTL paradigm. Addressing MTL with non-parallel labeled data is significant, as optimizing a neural network-based model on input instances with non-parallel labels may introduce bias toward a specific task, potentially causing instability in the training of the other task. For example, a WSJ input instance optimizes the parameters by its associated PoS labels. As such, the neural network tends to yield PoS-tagging-efficient parameters and lower the accuracy of text chunking. MTL with non-parallel labeled data is a common MTL paradigm, however, previous research did not pay enough attention on this [34, 35].

To address this challenge, we propose to incorporate parallel labeled data to balance the biased learning on non-parallel labeled data, assuming that a strategic combination of two instances can achieve effective learning for both tasks at each batch training step. To this end, we present a Curriculum Learning (CL) mechanism that selects complementary training instances from both datasets to be packed in the same training batch. The hypothesis (H1) is that a model can achieve more robust MTL with non-parallel labeled data, if the task-encoded input instances from two different datasets are in similar vector spaces. We use cross-entropy to measure the similarity between two vector spaces to select complementary samples, because cross-entropy is a classic and intuitive measure for quantifying the information difference between two probability distributions [36]. We select and train the pair of instances from two datasets in a same batch by the curriculum criterion of minimizing the cross-entropy of the hidden states from two task-specific towers.

We examine the pair-wise performance of SBD, chunking, and PoS tagging on our MTL model using three public datasets, and study how syntactic tasks of different granularity affect each other. The text chunking task obtains impressive performance when jointly learned with PoS tagging, achieving state-of-the-art performance (98.43% Micro-F1), outperforming the strongest published baseline by 1.13%. The PoS tagging task and SBD task both demonstrate performance gains when MTL with each other, compared to when MTL with chunking. It may be concluded that, pair-wise MTL

among syntactic tasks of different granularity can bring performance gain compared to single-task learning. Nevertheless, fine-grained task such as PoS tagging tend to be more helpful for coarse-grained tasks. Whereas coarse-grained task, i.e., SBD, provides useful structural information for PoS tagging. We also conduct an ablation study to demonstrate the effectiveness of our proposed LGDS and CL mechanisms.

Our research scope does not target to achieve state-of-the-art performance for all the involved tasks. Instead, we aim to propose a MTL framework where the the complementary effects between syntactic tasks of different granularity can be reliably evaluated using their respective benchmark datasets. Thus, the contribution of this work can be summarized as: (1) We propose an MTL framework with a novel soft parameter sharing mechanism that shares local and global dependency information for sequence-labeling-based syntactic processing tasks; (2) We propose a CL mechanism to improve the stability of the loss convergence for MTL with non-parallel-labeled data; (3) We study how syntactic tasks with different granularity complement each other through pair-wise MTL.

2 Related Work

2.1 Sentence Boundary Detection

SBD is an important yet overlooked pre-processing task. It is seemingly easy through identifying punctuation marks. However, the presence of period may cause notable ambiguities, e.g., abbreviations and decimal points. Early methods focus on the disambiguation of period usage in text. Recent task definition, motivated by automatic speech recognition, becomes more challenging by aiming to classify whether a word is followed by a sentence boundary punctuation mark in unsegmented speech transcripts [29]. Rule-based approach [37–39], despite the difficulties of constructing a comprehensive enough rule set, is still employed in recent years and achieves competitive performance. Whereas deep learning approaches [40–42] is the most widely used for the SBD task nowadays. Notably, early feature engineering methods often incorporate PoS tags as a useful feature for SBD [43–45], which suggests that PoS tagging might be a complementary joint learning task for SBD.

2.2 Text Chunking

Text chunking is normally formulated as a sequence labeling task since the work of Ramshaw and Marcus [46]. Early feature engineering methods utilized graphical models, e.g., Conditional Random Fields (CRF) [47–50]. In the era of deep learning, recent works utilized neural networks to automatically capture relevant features. The most widely-used architecture is the combination of CRF and Recurrent Neural Network (RNN) variants such as Bidirectional Long Short-Term Memory (BiLSTM) [21, 51–53] and Gated Recurrent Units (GRU) [54]. However, there is no study on the effectiveness of learning both tasks with dependency information sharing. We believe that PoS dependency features are useful for chunking.

2.3 Part-of-Speech Tagging

PoS tagging is a well-studied problem. Similar to text chunking, most feature engineering methods utilized graphical models [55–57], whereas Convolutional Neural Network (CNN) [58–60] and the abovementioned RNN variants are used to learn features nowadays. The accuracy of PoS tagging has been pushed to its near limit. Hence, improvement of this task should prioritize its effectiveness on aiding downstream tasks.

2.4 Multi-task Learning

MTL takes the advantage of sharing parameters and learns features from two or more tasks. It helps a machine improve performance and mitigate overfitting [61]. There has been studies where MTL is applied to multiple syntactic and semantic tasks [11, 25], e.g., PoS tagging, text chunking, name entity recognition, and semantic role labeling. However, the motivation behind task selection stems from similar task formulation (sequence labeling), instead of linguistic granularity. As such, the tasks involved are jointly learned together indiscriminately, and their complementarity to each other is not explored.

Furthermore, there are only a few studies that effectively address the non-parallel labeled data learning issue in MTL [35]. Chen et al. [25], Liu et al. [62] proposed LSTM-based architectures that can handle non-parallel labeled data by using a shared LSTM layer between two task towers. Their limitation is that such approach is more suitable for highly similar tasks, e.g., sentiment classification tasks in different domains or annotations. Similarly, Zhao et al. [63] employed shared stacked Bi-LSTM-CNNs with inter-task feedback strategy to adopt hierarchical tasks for parallel multi-task learning. However, such architecture runs the risk of biasing towards one task when the volume of task datasets are imbalanced.

2.5 Curriculum Learning

CL aims to automatically select the most suitable samples for each training step [64]. The curriculum is a sequence of training criteria that rely solely on the data, the model, and the task objective. CL is widely used to select training samples from easy to difficult for efficient learning [65–67]. However, to the best of our knowledge, CL has not been used to address the issue of MTL with non-parallel labeled data.

3 Methodology

We conduct pair-wise MTL on three syntactic processing tasks, namely, SBD, text chunking, and PoS tagging. Our hypothesis is that syntactic tasks of varying granularity tend to display different level of compatibility with one other. We also hypothesize that by controlling the combination of input data from different sources without parallel-annotated labels, an MTL model can achieve higher overall accuracy and smooth learning loss convergence. This is because the feature-alignment of multiple task inputs can help the neural network learn features from similar spaces. In contrast, features from very different spaces may lead to unstable learning. This is particularly

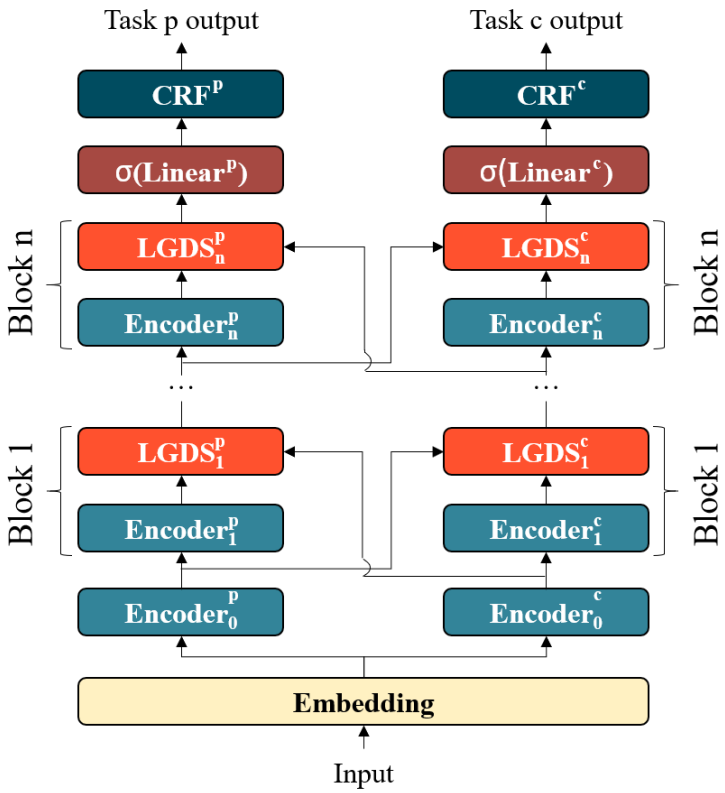

Fig. 2 Architecture of the multitask learning framework. σ denotes a SoftMax function.

important for MTL in the context where the sub-tasks do not have parallel labels for the same input sentence.

In light of this, we propose an MTL framework (the first subsection) with a novel soft parameter sharing mechanism to pass linguistic features learned from one task to the other, so as to investigate the pair-wise complementarity of the involved syntactic tasks. The soft parameter sharing mechanism (LGDS in the second subsection) means to share local and global dependency information between two tasks. Meanwhile, given the fact that some task datasets do not have ground-truth labels for the other target task, e.g., the SBD dataset not having PoS labels nor chunk labels, we introduce a CL mechanism (the third subsection) to improve the accuracy and learning stabilization by controlling the combination of input data from both tasks in the same batch.

3.1 Multi-task Learning

We denote the two tasks involved in our MTL framework as p and c, respectively. Then, given an input sentence $w = (w_1, w_2, \ldots, w_l)$, the goal of the MTL model is to predict its task p labels $p = (p_1, p_2, \ldots, p_l)$ and its task c labels $c = (c_1, c_2, \ldots, c_l)$, where l is the sequence length.

The architecture of the model is shown in Figure 2. The input sentence w is first embedded with pre-trained embeddings, then fed into the two respective towers for task p and task c. In each tower, the input is passed through an encoder ($Encoder_0$), whose output is denoted as H_0^p in the task p tower and H_0^c in the task c tower. In this work, we adopt Transformer [68] as the encoder, because it has been widely applied in diverse NLP tasks, presenting strong performance [69, 70].

Next, n blocks of encoders and soft parameter sharing mechanisms ($LGDS$) are employed. Here, we denote the output of block i in the task p tower as H_i^p, and the one in the task c tower as H_i^c. Then, for the task p tower, H_i^p is given by

$$T_i^p = Encoder_i^p(H_{i-1}^p), \tag{1}$$
$$H_i^p = LGDS_i^p(T_i^p, H_{i-1}^c). \tag{2}$$

For the task c tower, H_i^c is computed similarly to Equation 2 by incorporating H_{i-1}^p through LGDS.

Next, the final hidden states H_n^p and H_n^c are each fed into a linear layer ($L_{n+1}(\cdot)$) with SoftMax (σ):

$$E^p = \sigma(L_{n+1}^p(H_n^p)), \tag{3}$$
$$E^c = \sigma(L_{n+1}^c(H_n^c)). \tag{4}$$

Finally, during training, we use E^p, E^c, CRF and its loss function [57] to obtain losses for task p ($\mathcal{L}^p$) and task c ($\mathcal{L}^c$), respectively. The overall loss ($\mathcal{L}$) is given by

$$\mathcal{L} = \alpha\mathcal{L}^p + (1-\alpha)\mathcal{L}^c, \tag{5}$$

where α is a hyper-parameter. During inference, Viterbi decoding algorithm [71] is employed in CRF to predict the label sequences of task p and task c.

3.2 Local and Global Dependency Sharing

We propose a soft parameter sharing mechanism named LGDS to incorporate local and global dependencies that are learned from both tasks. As shown in Figure 3, LGDS combines CNN and Biaffine attention [13]. CNN, constricted by window size, is used to extract relevant information from the neighbour tower within the local context of the focal token. The output of the CNN in block i of the task c tower K_i^c can be computed as

$$\begin{aligned} K_i^{c'} = & ReLU(\text{Conv1D}(H_{i-1}^p, f=1) \\ & \oplus \text{Conv1D}(H_{i-1}^p, f=3) \\ & \oplus \text{Conv1D}(H_{i-1}^p, f=5)), \end{aligned} \tag{6}$$

$$K_i^c = tanh(W_k K_i^{c'} + b_k), \tag{7}$$

where $\oplus$ denotes concatenation. f denotes for filter width. W_k and b_k are learnable parameters.

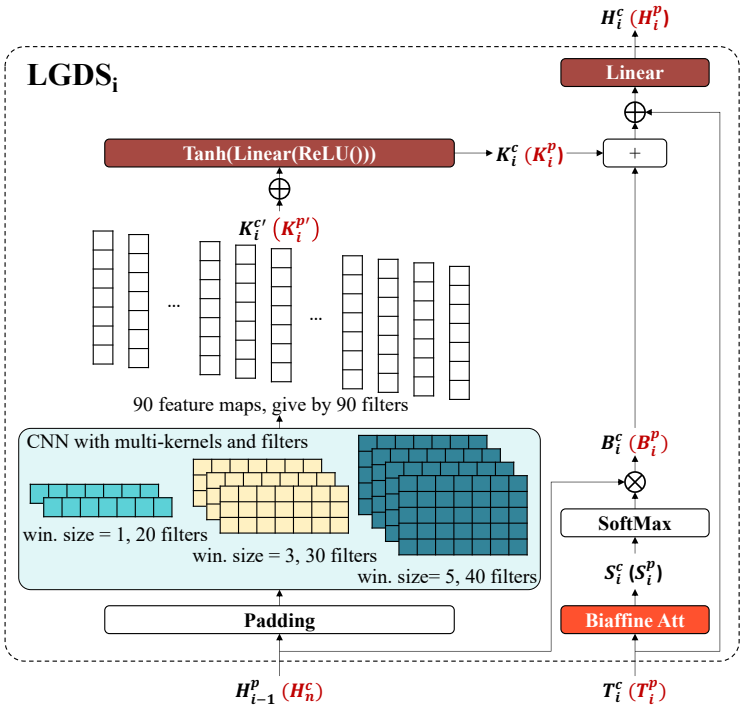

Fig. 3 Local and global dependency sharing (LGDS) mechanism. Bold italics denotes input and output variables, where black denotes the variables for learning task c, red with parentheses denotes variable for learning task p. Colored boxes denote layers with learnable parameters. + denotes a plus; $\oplus$ denotes concatenation; $\otimes$ denotes matrix multiplication.

Biaffine attention was used in dependency parsing to capture the dependencies between each word of a sentence [13]. Thus, we use Biaffine attention to capture the long-range dependencies. A Biaffine attention matrix ($S_i^c \in \mathbb{R}^{l \times l}$) in block i of task c is computed by

$$S_i^c = T_i^c U_i^c H_{i-1}^{p\top} + H_{i-1}^p e_i^c, \tag{8}$$

where U_i^c and e_i^c are learnable parameters. The Biaffine attention output (B_i^c) is the task p information (H_{i-1}^p) enhanced by its global task-c-dependent information (S_i^c)

$$B_i^c = tahn(softmax(S_i^c)H_{i-1}^p). \tag{9}$$

Finally, the LGDS output of the current block i from the task c tower (H_i^c) is computed as

$$H_i^c = L_i^c(T_i^c \oplus (K_i^c + B_i^c)). \tag{10}$$

The output for the LGDS in the task p tower (H_i^p) can be derived from similar procedures, while the input from the private tower is T_i^p instead of T_i^c; the input from the neighbour tower is H_{i-1}^c.

3.3 Curriculum Learning

The benchmark dataset for one task might not contain ground-truth labels for the other task, e.g., the PoS tagging dataset (WSJ) not being annotated with chunk labels, causing a non-parallel labeled data MTL issue. To alleviate this, we propose a CL mechanism to select complementary samples from the other task's dataset to optimize the training on the non-parallel-labeled task's data. Using the MTL of chunking and PoS tagging as an example, for an input WSJ instance ($w^{p,m}$), we randomly select J instances from CoNLL00 to feed into the model together for forward propagation. The length of $w^{p,m}$ is l. The CoNLL00 instances are padded or pruned to achieve the same length (l) in vector space. We denote the output of the first chunking encoder ($Encoder_0^c$) resulting from the j-th CoNLL00 instance as $T_0^{c,j}, (j \in \{1, \ldots, J\})$. Subsequently, we select the CoNLL00 instance whose $T_0^{c,j}$ is the most similar to $T_0^{p,m}$ (given by $w^{p,m}$ and $Encoder_0^p$) in vector space to balance the learning of $w^{p,m}$, according to cross-entropy

$$w^{c*} = \arg\min_j (-\sum_{k=1}^{l} T_{0,k}^{p,m} \log(T_{0,k}^{c,j})). \tag{11}$$

w^{c*} and $w^{p,m}$ are learned with both forward and backward propagation in the same batch to achieve stable CL for text chunking. We use hidden states from $Encoder_0^c$ and $Encoder_0^p$ rather than task losses as signals, because minimizing losses does not allow the model to learn useful information from the current input.

Similarly, we apply this CL procedure on the SBD dataset (IWSLT) when training with PoS tagging or text chunking, as it does not contain ground-truth labels for either task. We do not apply the CL mechanism when training on parallel annotated data, e.g., CoNLL00 with both chunking, PoS and SBD labels, and WSJ with both PoS and SBD labels.

4 Experiment

4.1 Baselines

To put the performance of our MTL framework into perspective, we include the following single task and multi-task baselines.

SBD:

- **T-BRNN** [72]: A bidirectional GRU model with attention mechanism, using GloVe embeddings.
- **BERT** [41]: A BERT-large model with Bi-LSTM-CRF stacked on top.
- **Roberta** [42]: A Roberta-large model with Bi-LSTM-CRF stacked on top.

Text chunking:

- **GRU-CRF** [54]: A deep hierarchical GRU model that encodes both character level and word level information, using fine-tuned SENNA embedding [11].
- **Star-C** [73]: A Star-Transformer-CRF model using GloVe embedding [74].

Table 1 Details of the WSJ, CoNLL 2000, and IWSLT datasets.

Dataset	Task	Data	# seq.	# token
IWSLT	SBD	Train	-	2,102,417
		Dev.	-	295,800
		Test	-	12,626
CoNLL 2022	Chunking	Train	8,937	211,727
		Test	2,013	47,377
WSJ	PoS tagging	Train	38,219	912,344
		Dev.	5,527	131,768
		Test	5,462	129,654

- **MVCRF** [12]: A multi-view CRF that extracts features from the word view as well as POS view, using SENNA embedding. **Flair-C** [21]: A BiLSTM-CRF model that utilizes Flair embedding, GloVe embedding and task-trained character embedding.
- **ACE** [22]: A BiLSTM-CRF model that automatically concatenates suitable embeddings including GloVe, Flair, BERT[75], etc.

POS tagging:

- **LSTM-CNN-CRF** [59]: A BiLSTM-CRF model that uses CNN to extract character level features, using GloVe embedding and task-trained character embedding.
- **GatedDualCNN** [60]: A deep CNN architecture that employs a dual path to alleviate the vanishing gradient problem, using GloVe embedding and task-trained character embedding.
- **Star-P** [73]: Same as Star-C.
- **Flair-P** [21]: Same as Flair-C.

MTL:

- **Meta-LSTM** [25]: A MTL framework with a task-shared Meta-LSTM layer for non-parallel labeled data, using GloVe embedding.
- **Gated** [76]: A MTL model with gated network as sharing mechanism using BERT embedding.
- **AUX** [77]: A BiLSTM-based model that concatenates the output of the auxiliary tower with input representation to feed into the primary tower.

We also conduct experiments by using BiLSTM instead of Transformer in our framework (**Ours-LSTM**) to achieve a fair comparison with other BiLSTM-CRF-based baselines.

4.2 Datasets

Our multitask learning framework is trained and evaluated with the WSJ dataset [78], the CoNLL00 dataset [33], and the IWSLT dataset [29]. The details of the used datasets can be found in Table 1. For the SDB task, we use the PERIOD class tagging schema provided by the IWSLT dataset [29], and use the Ref testing set for evaluation. For the chunking task, we use the BIOES tagging schema [28]. For the POS tagging

Table 2 SBD results when learning with PoS tagging (SBD w/ PoS) and with chunking (SBD w/ Chunking). The F1 scores of the single-task learning baselines as shown under the SBD w/ PoS column for readability. The bold and underlines denote the best and second best results. The numbers on the subscripts are standard deviations. * indicates the models re-implemented by us. w/CL means our CL mechanism is applied.

Model	SBD w/ PoS		SBD w/ Chunking	
	F1	Acc	F1	Acc
T-BRNN	72.9	-	-	-
BERT	84.1	-	-	-
Roberta	**88.6**	-	-	-
Gated*	81.31	81.42	79.22	79.53
Gated w/CL*	81.58	81.64	79.69	79.88
AUX*	79.99	80.28	78.93	79.41
AUX w/CL*	80.46	80.61	79.18	79.52
Ours-LSTM	86.74	87.01	84.43	84.86
Ours-Transformer	<u>87.25</u>	87.48	85.82	85.95

task, we use the Penn Treebank annotation schema [27]. The standard evaluation metrics are accuracy for POS tagging, F1 measure for text chunking, and PERIOD class F1 measure [29] for SBD, which are in line with our baselines.

4.3 Setups

For the hyper-parameters, we randomly select $J = 4$ chunking instances for CL. We adopt $\alpha = 0.6$. We set the initial learning rate to 0.0001, and adopt a learning schedule with the step size of 20 and decay factor $\gamma = 0.5$. We also adapt an early stop strategy, where the model stops training if the overall accuracy of the two tasks is not improved in five epochs. We use Adam [79] to optimize the model. We run 30 epochs with the batch size of 20 on NVIDIA Tesla P100-PCI-E. We use Flair embedding, GloVe embedding, and task-trained character embedding as embeddings, aligning with one of the strongest baselines for two of the target tasks [21]. There are 2 blocks (n=2 in Figure 2) of encoder with LGDS in each task specific tower. The Transformer-based encoders have 4 heads, 128 dimension hidden states. Additionally, we examine BiLSTM-based encoders with 200 dimension hidden states. We report micro-F1 and accuracy for both text chunking and PoS tagging tasks, based on the averaged results of 5 runs.

5 Results

From Table 2, we can see that jointly learning SBD with PoS tagging outperforms that with text chunking by 2.02% F1 score. Such a performance advantage can be observed in all the MTL methods, indicating that PoS tagging is a more complementary task for SBD than chunking. Additionally, Ours-Transformer outperforms BERT in both task pair settings, and only falls behind Roberta by 1.35% F1 score when paring with PoS tagging, despite using fewer Transformer layers than both baselines. It shows that our MTL framework can pass useful features from one task tower to the other and glean

Table 3 Text chunking results when MTL with PoS tagging (Chunking w/ PoS) and with SBD (Chunking w/ SBD). The F1 scores of the single-task learning baselines as shown under the Chunking w/ PoS column for readability. The bold and underlines denote the best and second best results. The numbers on the subscripts are standard deviations. * indicates the models re-implemented by us. w/CL means our CL mechanism is applied.

Model	Chunking w/ PoS		Chunking w/ SBD	
	F1	Acc	F1	Acc
GRU-CRF	95.41	-	-	-
MVCRF	95.44	-	-	-
Star-C	95.93	-	-	-
Flair-C	96.72	-	-	-
ACE	97.3	-	-	-
Meta	95.11	-	-	-
Gated*	97.50	97.63	97.11	97.36
Gated w/CL*	<u>97.82</u>	97.94	97.23	97.51
AUX*	97.18	97.51	96.69	96.83
AUX w/CL*	97.52	97.86	96.80	96.98
Ours-LSTM	97.96	98.04	97.40	97.73
Ours-Transformer	**98.13**	98.45	97.48	97.98

the benefits of joint learning. Applying our proposed CL mechanism also consistently improve the MTL baselines Gated and AUX in both task pair settings, proving its effectiveness in bringing performance gain.

Results shown in Table 3 indicate that text chunking achieves the best performance when jointly learned with PoS tagging. Although outperforming the best single-task baseline (ACE) when learned with SBD, the extend of improvement, 0.18%, is much less significant comparing to when learned with PoS tagging, which stands at 0.83%. This contrast of complementarity can also be observed among all the experimented MTL methods. We can also see that when jointly learned with PoS tagging, Ours-LSTM outperforms the LSTM-based ACE by 0.66% in F1 scores. It shows the effectiveness of our proposed MTL task pair, soft parameter sharing (LGDS) and CL mechanisms. Using Transformer can further improve the model, reaching 98.13% F1, achieving the best performance. It significantly outperforms Meta, showing that our approach for non-parallel labeled data in MTL is superior to existing works.

From results shown in Tables 2 and 4, we can conclude that our model achieves the best PoS tagging and SBD performance when they are jointly learned, whereas Table 3 and 4 indicate that chunking can benefit a lot from PoS tagging but not in reverse. It can be inferred that fine-grained syntactic tasks are complementary for MTL with the more coarse-grained ones, among which the most fine-grained task, namely PoS tagging, consistently contributes the most to the improvement of the other tasks. On the other hand, PoS tagging receives limited benefits from MTL with more coarse-grained tasks. Jointly learning PoS tagging with SBD achieves better performance than with chunking, and comparable performance with the strongest single-task baseline. This might be due to the fact that SBD provides global sequence structure information, but is more challenging to learn in the fine-coarse processing of PoS tagging.

Table 4 PoS tagging results when learning with SBD (PoS w/ SBD) and with chunking (PoS w/ chunk). The accuracy of the single-task learning baselines as shown under the PoS w/ SBD column for readability. The bold and underlines denote the best and second best results. The numbers on the subscripts are standard deviations. * indicates the models re-implemented by us. w/CL means our CL mechanism is applied.

Model	PoS w/ SBD		PoS w/ chunk	
	F1	Acc	F1	Acc
LSTM-CNN-CRF	-	97.55	-	-
GatedDualCNN	-	97.59	-	-
Star-P	-	97.68	-	-
Flair-P	-	**97.85**	-	-
Meta	-	-	-	97.45
Gated*	97.67	97.64	97.47	97.40
Gated w/CL*	97.74	97.70	97.60	97.52
AUX*	97.61	97.59	97.60	97.53
AUX w/CL*	97.71	97.66	97.62	97.56
Ours-LSTM	97.78	97.70	97.58	97.52
Ours-Transformer	97.86	<u>97.79</u>	97.64	97.59

Combining the results in Tables 2, 3, and 4, we can further draw the conclusions that, 1) when paired with the most complementary task, Our-Transformer significantly outperforms the strongest baseline in text chunking, and obtains comparable performance in SBD and PoS tagging; 2) applying our proposed CL mechanism consistently bring significant improvement to the MTL baselines in all experiment settings, indicating its robustness. Based on the former observation, Our-Transformer is our main model of investigation in the following experiments.

5.1 Ablation Study

We conduct an ablation study using the best performing MTL setups for each task, reported in Tables 2, 3, and 4, i.e., chunking paired with PoS tagging, and PoS tagging and SBD paired together. Specially, the following variants are studied:

- **w/o MTL** denotes that the two target tasks are trained on the base tower structure of Transformer encoders and CRF using single-task learning.
- **w/o LGDS** denotes a hard parameter sharing model without LGDS and CL, where the two tasks share the same encoder layers and keep individual output layers.
- **Finetune** denotes pre-training the task tower without parallel labels using its corresponding dataset first, then fine-tuning the two task towers with the parallel labeled dataset.
- **w/o CL** denotes that chunking and PoS tagging are trained on the proposed LGDS-based MTL architecture without CL.
- **w/o non-parallel** denotes a w/o CL model that is trained solely on parallel labeled dataset, but is evaluated on testing sets of both tasks.

As seen in Table 5, a hard parameter sharing MTL model without LGDS (w/o LGDS) yields higher performance than the single task learning model (w/o MTL)

Table 5 Ablation study results. Chunking w/ PoS denotes chunking results when paired with PoS tagging. PoS w/ SBD denotes PoS tagging results when paired with SBD. SBD w/ PoS denotes SBD results when paired with PoS tagging.

Model	Chunking w/ PoS		PoS w/ SBD		SBD w/ PoS	
	F1	Acc	F1	Acc	F1	Acc
w/o MTL	95.77	96.64	97.51	97.46	85.04	85.18
w/o LGDS	96.47	97.20	97.61	97.55	86.42	86.57
Finetune	97.21	97.38	97.10	96.84	81.97	82.22
w/o CL	97.71	98.24	97.77	97.68	87.16	87.32
w/o non-parallel	97.35	97.96	96.32	96.33	79.98	80.26
Ours-Transformer	98.13	98.45	97.86	97.79	87.25	87.48

on all three tasks. Further comparisons between the performance of w/o MTL and Ours-Transformer in Tables 3, 4, and 2 show that all pair-wise MTL combinations perform better than the single task counterparts. This shows that joint training between syntactic tasks can provide useful features for each other. Learning multiple tasks simultaneously can also help the model against overfitting [61], because the model needs to learn robust representations to achieve the training targets of both tasks. The improvements of w/o CL over w/o LGDS are consistent across the three tasks, showing the effectiveness of our proposed soft parameter sharing mechanism and layer connections. Comparing the performance of Finetune with w/o CL and Our-Transformer on the three tasks, we can conclude that fine-tuning cannot achieve stable learning for MTL. Next, there is a sharp drop on SBD performance by simply using the WSJ dataset for joint learning with PoS tagging (w/o non-parallel). As a result, the PoS tagging performance also decreases. The same reason can be inferred to be responsible for the performance degradation of Chunking w/ PoS when solely using the CoNLL00 training set for the MTL of text chunking and PoS tagging. This shows the significance of introducing data from both tasks to support the pair-wise MTL. Finally, using the proposed LGDS, training strategies, and CL mechanism can help the model achieve further improvements on the three tasks, which is evidenced by the consistent improvements of Ours-Transformer over w/o CL and w/o WSJ models across different tasks.

5.2 Curriculum Learning Analysis

Figure 4 shows the loss curves of SBD and PoS tagging in pair-wise MTL, given by CL-4 (Figure 4a) and CL-1 (Figure 4b), respectively. CL-1 denotes that we randomly sample the equal number of instances from both task datasets in a batch without using any sample selection criterion. CL-4 employs our recommended CL sample size and sample selection criterion. Similarly, Figure 5 shows the loss curves of PoS tagging and text chunking, and Figure 6 the curves of SBD and chunking in pair-wise MTL. It can be observed in Figure 4 that the fluctuation of the PoS tagging loss curve of CL-4 is less than those of CL-1 (the blue lines). The same can be seen in the SBD curves (the green lines), albeit to a smaller extent. It shows that our proposed curriculum criterion (Eq. 11) is effective in selecting complementary PoS tagging instances to optimize and stabilize the learning of both PoS tagging and SBD. Such smoothing effect can also be observed in Figures 5 and 6 for the chunking loss curves (the red lines). The difference

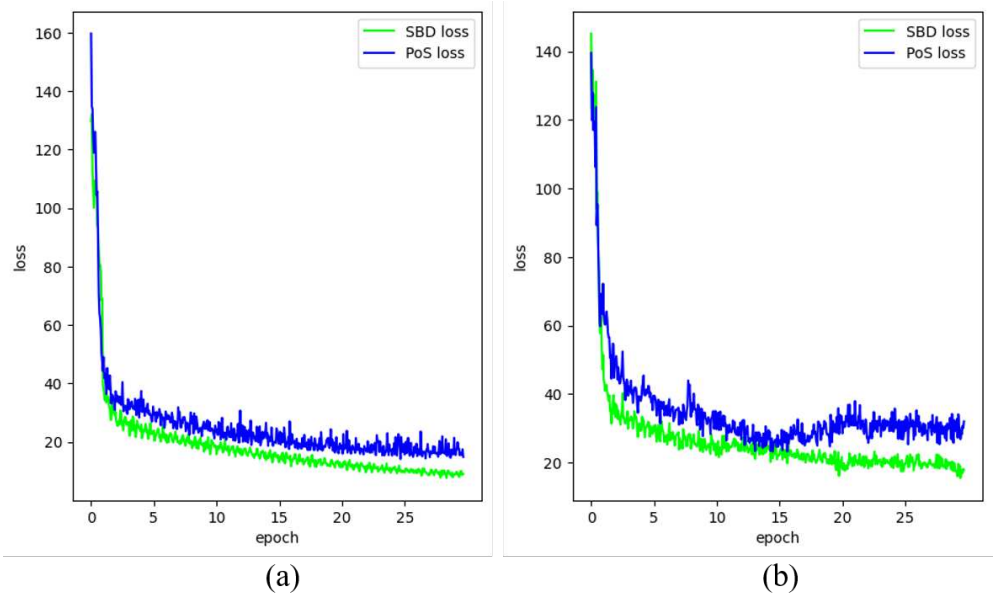

Fig. 4 Loss curves of SBD and PoS tagging, given by (a) our proposed CL mechanism with a sample size in 4; (b) the CL mechanism with a sample size in 1.

Table 6 Curriculum learning sample size analysis by time costs and overall accuracy gains. Underlines denote baselines.

Setup	Time costs	Δ	Avg acc
w/o CL	0.86X	-0.03	97.70
CL-1	1.00X	-	97.73
CL-2	1.08X	+0.15	97.88
CL-4	1.31X	+0.24	97.97
CL-8	1.74X	+0.26	97.99

of PoS tagging loss curves (the blue lines) in the two figures is not conspicuous, as does the difference of SBD loss curves (the green lines). It might be that the learning of chunking does not cause significant biases for that of PoS tagging nor SBD. The comparatively stable loss curves in CL-4 (Figures 4a, 5a, and 6a) prove our hypothesis (H1 in introduction) that a model can achieve more robust MTL with non-parallel labeled data, if input instances from two different tasks are in similar vector spaces.

We further analyze the tradeoff between time costs and performance using text chunking and PoS tagging as a case study in Table 6. We use the averaged accuracy of chunking and PoS tagging as the overall accuracy measure, because our early stop point is determined by the condition that the highest overall accuracy (the sum of text chunking and PoS tagging accuracy) is not improved in 5 training epochs. We use the CL sample size of 1 (CL-1) as the time baseline, which means we randomly

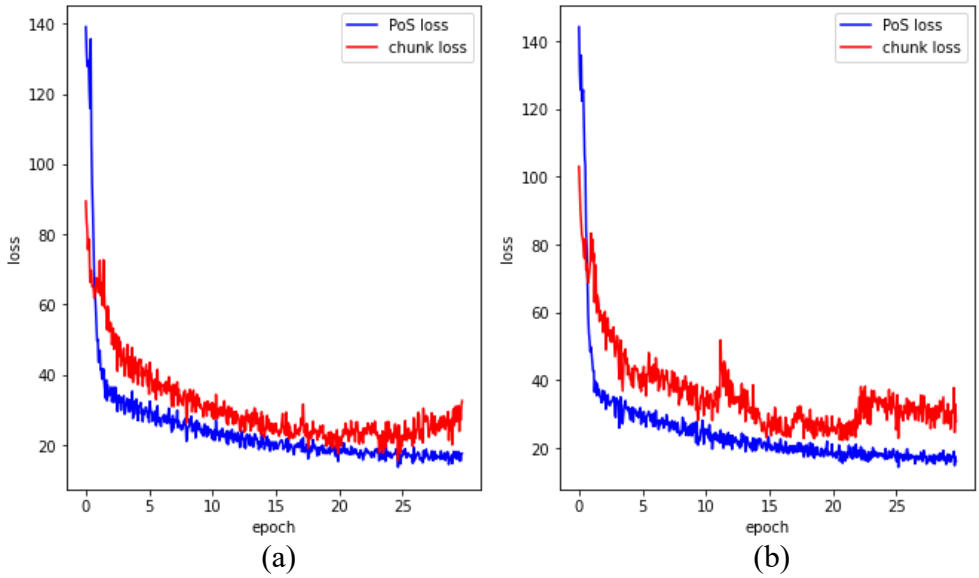

Fig. 5 Loss curves of PoS tagging and text chunking, given by (a) our proposed CL mechanism with a sample size in 4; (b) the CL mechanism with a sample size in 1.

sample a CoNLL00 instance to learn with a WSJ instance; CL-2 means the random sample size is 2, and so on. The accuracy shows an upper trend as the sample size grows. However, the time costs also increase, because the model needs to compute and compare more cross-entropy of hidden states when the sample size is larger. The CL improvements in different setups are marginal in our full model, because the model with LGDS and without CL (w/o CL) has achieved very high accuracy in both tasks. Improving model performance is very hard, given the w/o CL baseline has yielded an average accuracy by 97.70%. Compared with the improvement space (2.30%) to the ground-truth (100%), the gap (0.27%) between CL-4 (97.97%) and w/o CL (97.70%) and the gap (0.24%) between CL-4 (97.97%) and CL-1 (97.73%) are reasonable.

6 Conclusion

In this work, we propose a soft parameter sharing mechanism to share dependency information that is learned from the two involved tasks in pair-wise MTL. It consists of CNN and Biaffine attention to capture local and global dependency, respectively. Additionally, we propose a CL mechanism to achieve robust MTL with non-parallel labeled data. The addition of CL mitigates the learning bias given by the task with non-parallel data, so that the performance of both tasks may further improve. The employed curriculum criterion enables effective selection of complementary data, so that the learning loss of the tasks involved can converge more steadily.

Using the proposed MTL method, we conduct a study on how syntactic tasks of different granularity complement each other through pair-wise MTL. We conclude

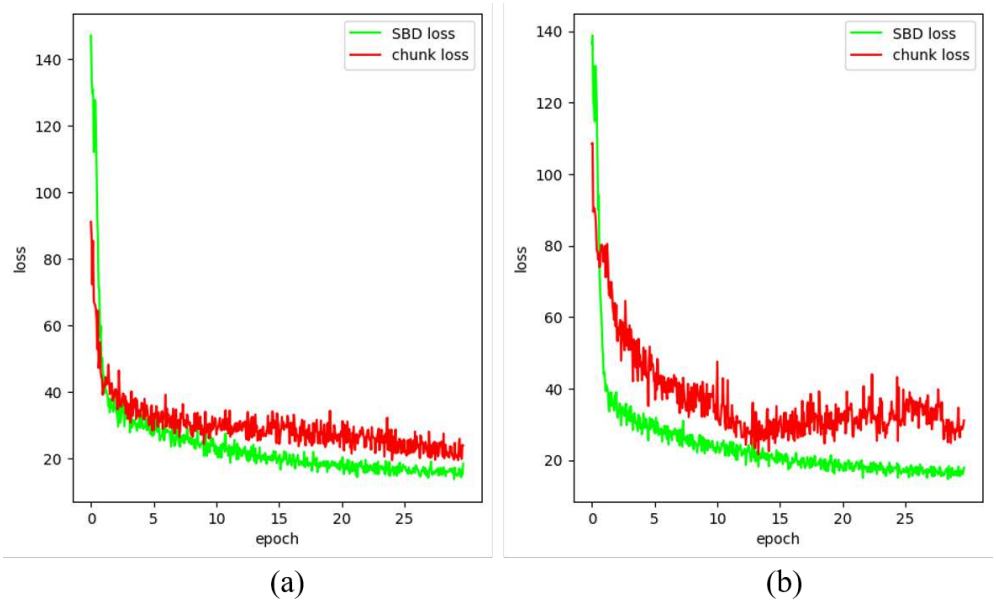

Fig. 6 Loss curves of SBD and text chunking, given by (a) our proposed CL mechanism with a sample size in 4; (b) the CL mechanism with a sample size in 1.

that fine-grained tasks can provide information that yield significant gains for coarse-grained tasks. On the other hand, the benefits that coarse-grained tasks bring to fine-grained tasks are limited, with the exception of SBD to PoS tagging, which is likely because the delineation of structure in an input sequence learned in the SBD task helps the PoS tagging tower focus on learning features specific to sentences, facilitating the learning of long-range label dependencies.

Additionally, our model achieves state-of-the-art performance on text chunking, and comparable performance on SBD and PoS tagging to the state-of-the-art baselines. We will test if our CL mechanism can relax task relevance requirement in MTL in future work.

Declarations

- **Conflict of interest** The authors declare they have no conflict of interest.
- **Ethics approval** This paper does not contain any studies with human participants or animals performed by any of the authors.
- **Availability of data and materials** The datasets used in this study are publicly available.
- **Funding** No funding was obtained for this study.
- **Authors' contributions** Xulang Zhang: conceptualization, methodology, software, writing – original draft; Rui Mao: conceptualization, methodology, investigation,

writing – review & editing; Erik Cambria: resources, supervision, writing – review & editing.

References

[1] Woolf, B.P.: Chapter 5 - communication knowledge. In: Woolf, B.P. (ed.) Building Intelligent Interactive Tutors, pp. 136–182. Morgan Kaufmann, San Francisco (2009)

[2] Cambria, E., Mao, R., Chen, M., Wang, Z., Ho, S.-B.: Seven pillars for the future of AI. IEEE Intelligent Systems **38**(6) (2023)

[3] Matsoukas, S., Bulyko, I., Xiang, B., Nguyen, K., Schwartz, R., Makhoul, J.: Integrating speech recognition and machine translation. In: 2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07, vol. 4, p. 1281 (2007). IEEE

[4] Zhou, N., Wang, X., Aw, A.: Dynamic boundary detection for speech translation. In: 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pp. 651–656 (2017). IEEE

[5] Krallinger, M., Rabal, O., Lourenco, A., Oyarzabal, J., Valencia, A.: Information retrieval and text mining technologies for chemistry. Chemical reviews **117**(12), 7673–7761 (2017)

[6] Jing, H., Lopresti, D., Shih, C.: Summarization of noisy documents: a pilot study. In: Proceedings of the HLT-NAACL 03 Text Summarization Workshop, pp. 25–32 (2003)

[7] Boudin, F., Huet, S., Torres-Moreno, J.-M.: A graph-based approach to cross-language multi-document summarization. Polibits (43), 113–118 (2011)

[8] Councill, I., McDonald, R., Velikovich, L.: What's great and what's not: learning to classify the scope of negation for improved sentiment analysis. In: Proceedings of the Workshop on Negation and Speculation in Natural Language Processing, pp. 51–59 (2010)

[9] Gupta, H., Kottwani, A., Gogia, S., Chaudhari, S.: Text analysis and information retrieval of text data. In: 2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), pp. 788–792 (2016). IEEE

[10] Syed, A.Z., Aslam, M., Martinez-Enriquez, A.M.: Associating targets with sentiunits: a step forward in sentiment analysis of urdu text. Artificial intelligence review **41**(4), 535–561 (2014)

[11] Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.:

Natural language processing (almost) from scratch. Journal of machine learning research **12**(ARTICLE), 2493–2537 (2011)

[12] Sun, X., Sun, S., Yin, M., Yang, H.: Hybrid neural conditional random fields for multi-view sequence labeling. Knowledge-Based Systems **189**, 105151 (2020)

[13] Dozat, T., Manning, C.D.: Deep biaffine attention for neural dependency parsing. arXiv preprint arXiv:1611.01734 (2016)

[14] Zhou, H., Zhang, Y., Li, Z., Zhang, M.: Is POS Tagging Necessary or Even Helpful for Neural Dependency Parsing? (2020)

[15] Mahmood, A., Khan, H.U., Zahoor-ur-Rehman, Khan, W.: Query based information retrieval and knowledge extraction using hadith datasets. In: 2017 13th International Conference on Emerging Technologies (ICET), pp. 1–6 (2017). https://doi.org/10.1109/ICET.2017.8281714

[16] Asghar, M.Z., Khan, A., Ahmad, S., Kundi, F.M.: A review of feature extraction in sentiment analysis. Journal of Basic and Applied Scientific Research **4**(3), 181–186 (2014)

[17] Cambria, E., Liu, Q., Decherchi, S., Xing, F., Kwok, K.: SenticNet 7: A commonsense-based neurosymbolic AI framework for explainable sentiment analysis. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference, pp. 3829–3839 (2022)

[18] Mao, R., Lin, C., Guerin, F.: Word embedding and WordNet based metaphor identification and interpretation. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, vol. 1, pp. 1222–1231 (2018)

[19] Ge, M., Mao, R., Cambria, E.: Explainable metaphor identification inspired by conceptual metaphor theory. In: Proceedings of AAAI, pp. 10681–10689 (2022)

[20] Mao, R., Li, X., He, K., Ge, M., Cambria, E.: MetaPro Online: A computational metaphor processing online system. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), pp. 127–135. Association for Computational Linguistics, Toronto, Canada (2023). https://aclanthology.org/2023.acl-demo.12

[21] Akbik, A., Blythe, D., Vollgraf, R.: Contextual string embeddings for sequence labeling. In: Proceedings of the 27th International Conference on Computational Linguistics, pp. 1638–1649 (2018)

[22] Wang, X., Jiang, Y., Bach, N., Wang, T., Huang, Z., Huang, F., Tu, K.: Automated concatenation of embeddings for structured prediction. arXiv preprint arXiv:2010.05006 (2020)

[23] Wong, D.F., Chao, L.S., Zeng, X.: isentenizer-: Multilingual sentence boundary detection model. The Scientific World Journal **2014** (2014)

[24] Zhang, X., Mao, R., Cambria, E.: A survey on syntactic processing techniques. Artificial Intelligence Review **56**(6), 5645–5728 (2023)

[25] Chen, J., Qiu, X., Liu, P., Huang, X.: Meta multi-task learning for sequence modeling. In: McIlraith, S.A., Weinberger, K.Q. (eds.) Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th Innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018, pp. 5070–5077. AAAI Press, ??? (2018). https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17140

[26] Yang, Z., Salakhutdinov, R., Cohen, W.W.: Transfer learning for sequence tagging with hierarchical recurrent networks. In: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. OpenReview.net, ??? (2017). https://openreview.net/forum?id=ByxpMd9lx

[27] Marcus, M.P., Santorini, B., Marcinkiewicz, M.A.: Building a large annotated corpus of English: The Penn Treebank. Computational Linguistics **19**(2), 313–330 (1993)

[28] Ratinov, L., Roth, D.: Design challenges and misconceptions in named entity recognition. In: Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009), pp. 147–155 (2009)

[29] Che, X., Wang, C., Yang, H., Meinel, C.: Punctuation prediction for unsegmented transcript based on word vector. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pp. 654–658 (2016)

[30] Mao, R., Li, X.: Bridging towers of multi-task learning with a gating mechanism for aspect-based sentiment analysis and sequential metaphor identification. Proceedings of the AAAI Conference on Artificial Intelligence **35**, 13534–13542 (2021)

[31] Ruder, S.: An overview of multi-task learning in deep neural networks. CoRR **abs/1706.05098** (2017) 1706.05098

[32] Chen, S., Zhang, Y., Yang, Q.: Multi-task learning in natural language processing: An overview. arXiv preprint arXiv:2109.09138 (2021)

[33] Tjong Kim Sang, E.F., Buchholz, S.: Introduction to the conll-2000 shared task: Chunking. In: Fourth Conference on Computational Natural Language Learning and the Second Learning Language in Logic Workshop. ConLL '00, pp. 127–132.

Association for Computational Linguistics, USA (2000). https://doi.org/10.3115/1117601.1117631 . https://doi.org/10.3115/1117601.1117631

[34] Le, D., Thai, M., Nguyen, T.: Multi-task learning for metaphor detection with graph convolutional neural networks and word sense disambiguation. In: Proceedings of the 34th AAAI Conference on Artificial Intelligence, pp. 8139–8146 (2020)

[35] Zhang, Z., Yu, W., Yu, M., Guo, Z., Jiang, M.: A survey of multi-task learning in natural language processing: Regarding task relatedness and training methods. arXiv preprint arXiv:2204.03508 (2022)

[36] Bhat, S., Debnath, A., Banerjee, S., Shrivastava, M.: Word embeddings as tuples of feature probabilities. In: Proceedings of the 5th Workshop on Representation Learning for NLP, pp. 24–33. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.repl4nlp-1.4 . https://aclanthology.org/2020.repl4nlp-1.4

[37] Grefenstette, G., Tapanainen, P.: What is a word, what is a sentence?: Problems of tokenisation. Report, Grenoble Laboratory (1994)

[38] Stamatatos, E., Fakotakis, N., Kokkinakis, G.: Automatic extraction of rules for sentence boundary disambiguation. In: Proceedings of the Workshop on Machine Learning in Human Language Technology, pp. 88–92 (1999). Citeseer

[39] Sadvilkar, N., Neumann, M.: PySBD: Pragmatic sentence boundary disambiguation. arXiv preprint arXiv:2010.09657 (2020)

[40] Knoll, B.C., Lindemann, E.A., Albert, A.L., Melton, G.B., Pakhomov, S.V.S.: Recurrent deep network models for clinical nlp tasks: Use case with sentence boundary disambiguation. Studies in health technology and informatics **264**(31437913), 198–202 (2019) https://doi.org/10.3233/SHTI190211

[41] Makhija, K., Ho, T.-N., Chng, E.-S.: Transfer learning for punctuation prediction. In: 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pp. 268–273 (2019). IEEE

[42] Alam, T., Khan, A., Alam, F.: Punctuation restoration using transformer models for high-and low-resource languages. In: Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020), pp. 132–142 (2020)

[43] Palmer, D.D., Hearst, M.A.: Adaptive multilingual sentence boundary disambiguation. Comput. Linguist. **23**(2), 241–267 (1997)

[44] Mikheev, A.: Tagging sentence boundaries. In: 1st Meeting of the North American Chapter of the Association for Computational Linguistics (2000)

[45] Agarwal, N., Ford, K.H., Shneider, M.: Sentence boundary detection using a maxEnt classifier. In: Proceedings of MISC, pp. 1–6 (2005)

[46] Ramshaw, L.A., Marcus, M.: Text chunking using transformation-based learning. In: Yarowsky, D., Church, K. (eds.) Third Workshop on Very Large Corpora (1995). https://aclanthology.org/W95-0107/

[47] Sutton, C., McCallum, A., Rohanimanesh, K.: Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data. Journal of Machine Learning Research **8**(3) (2007)

[48] Sun, X., Morency, L.-P., Okanohara, D., Tsuruoka, Y., Tsujii, J.: Modeling latent-dynamic in shallow parsing: A latent conditional model with improved inference. In: Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), pp. 841–848 (2008)

[49] Lin, J.C.-W., Shao, Y., Zhang, J., Yun, U.: Enhanced sequence labeling based on latent variable conditional random fields. Neurocomputing **403**, 431–440 (2020)

[50] Liu, Y., Li, G., Zhang, X.: Semi-Markov CRF model based on stacked neural Bi-LSTM for sequence labeling. In: 2020 IEEE 3rd International Conference of Safe Production and Informatization (IICSPI), pp. 19–23 (2020). https://doi.org/10.1109/IICSPI51290.2020.9332321

[51] Huang, Z., Xu, W., Yu, K.: Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv:1508.01991 (2015)

[52] Rei, M.: Semi-supervised multitask learning for sequence labeling. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 2121–2130. Association for Computational Linguistics, Vancouver, Canada (2017). https://doi.org/10.18653/v1/P17-1194 . https://aclanthology.org/P17-1194

[53] Zhai, F., Potdar, S., Xiang, B., Zhou, B.: Neural models for sequence chunking. arXiv preprint arXiv:1701.04027 (2017)

[54] Yang, Z., Salakhutdinov, R., Cohen, W.: Multi-task cross-lingual sequence tagging from scratch. arXiv preprint arXiv:1603.06270 (2016)

[55] Brants, T.: TnT-a statistical part-of-speech tagger. arXiv preprint cs/0003055 (2000)

[56] McCallum, A., Freitag, D., Pereira, F.C.: Maximum entropy markov models for information extraction and segmentation. In: Icml, vol. 17, pp. 591–598 (2000)

[57] Lafferty, J.D., McCallum, A., Pereira, F.C.N.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of

the Eighteenth International Conference on Machine Learning. ICML '01, pp. 282–289. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2001)

[58] Dos Santos, C., Zadrozny, B.: Learning character-level representations for part-of-speech tagging. In: International Conference on Machine Learning, pp. 1818–1826 (2014). PMLR

[59] Ma, X., Hovy, E.: End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. arXiv preprint arXiv:1603.01354 (2016)

[60] Zhao, L., Qiu, X., Zhang, Q., Huang, X.: Sequence labeling with deep gated dual path CNN. IEEE/ACM Transactions on Audio, Speech, and Language Processing **27**(12), 2326–2335 (2019)

[61] Ruder, S.: An overview of multi-task learning in deep neural networks. arXiv preprint arXiv:1706.05098 (2017)

[62] Liu, P., Qiu, X., Huang, X.: Recurrent neural network for text classification with multi-task learning. arXiv preprint arXiv:1605.05101 (2016)

[63] Zhao, S., Liu, T., Zhao, S., Wang, F.: A neural multi-task learning framework to jointly model medical named entity recognition and normalization. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 817–824 (2019)

[64] Soviany, P., Ionescu, R.T., Rota, P., Sebe, N.: Curriculum learning: A survey. International Journal of Computer Vision, 1–40 (2022)

[65] Ma, F., Meng, D., Xie, Q., Li, Z., Dong, X.: Self-paced co-training. In: International Conference on Machine Learning, pp. 2275–2284 (2017). PMLR

[66] Zhang, X., Kumar, G., Khayrallah, H., Murray, K., Gwinnup, J., Martindale, M.J., McNamee, P., Duh, K., Carpuat, M.: An empirical exploration of curriculum learning for neural machine translation. arXiv preprint arXiv:1811.00739 (2018)

[67] Wang, W., Caswell, I., Chelba, C.: Dynamically composing domain-data selection with clean-data selection by "co-curricular learning" for neural machine translation. arXiv preprint arXiv:1906.01130 (2019)

[68] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in Neural Information Processing Systems **30** (2017)

[69] Khan, S., Naseer, M., Hayat, M., Zamir, S.W., Khan, F.S., Shah, M.: Transformers in vision: A survey. ACM Computing Surveys (CSUR) (2021)

[70] Mao, R., Li, X., Ge, M., Cambria, E.: Metapro: A computational metaphor processing model for text pre-processing. Information Fusion **86-87**, 30–43 (2022)

https://doi.org/10.1016/j.inffus.2022.06.002

[71] Forney, G.D.: The viterbi algorithm. Proceedings of the IEEE **61**(3), 268–278 (1973)

[72] Tilk, O., Alumäe, T.: Bidirectional recurrent neural network with attention mechanism for punctuation restoration. In: Interspeech, vol. 3, p. 9 (2016)

[73] Guo, Q., Qiu, X., Liu, P., Shao, Y., Xue, X., Zhang, Z.: Star-transformer. arXiv preprint arXiv:1902.09113 (2019)

[74] Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014)

[75] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)

[76] Dankers, V., Rei, M., Lewis, M., Shutova, E.: Modelling the interplay of metaphor and emotion through multitask learning. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 2218–2229 (2019)

[77] Alqahtani, S., Mishra, A., Diab, M.: A multitask learning approach for diacritic restoration. arXiv preprint arXiv:2006.04016 (2020)

[78] Collins, M.: Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In: Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002), pp. 1–8. Association for Computational Linguistics, ??? (2002). https://doi.org/10.3115/1118693.1118694 . https://www.aclweb.org/anthology/W02-1001

[79] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)