Enhancing Opinion Role Labeling with Semantic-Aware Word Representations from Semantic Role Labeling

Meishan Zhang[1] and **Peili Liang**[2] and **Guohong Fu**[3*]
1. School of New Media and Communication, Tianjin University, China
2. School of Computer Science and Technology, Heilongjiang University, China
3. Institute of Artificial Intelligence, Soochow University, China
mason.zms@gmail.com,
imlpliang@gmail.com,
ghfu@hotmail.com

Abstract

Opinion role labeling (ORL) is an important task for fine-grained opinion mining, which identifies important opinion arguments such as holder and target for a given opinion trigger. The task is highly correlative with semantic role labeling (SRL), which identifies important semantic arguments such as agent and patient for a given predicate. As predicate agents and patients usually correspond to opinion holders and targets respectively, SRL could be valuable for ORL. In this work, we propose a simple and novel method to enhance ORL by utilizing SRL, presenting semantic-aware word representations which are learned from SRL. The representations are then fed into a baseline neural ORL model as basic inputs. We verify the proposed method on a benchmark MPQA corpus. Experimental results show that the proposed method is highly effective. In addition, we compare the method with two representative methods of SRL integration as well, finding that our method can outperform the two methods significantly, achieving 1.47% higher F-scores than the better one.

1 Introduction

Fine-grained opinion mining aims to detect structured user opinions in text, which has drawn much attention in the natural language processing (NLP) community (Kim and Hovy, 2006; Breck et al., 2007; Ruppenhofer et al., 2008; Wilson et al., 2009; Qiu et al., 2011; Irsoy and Cardie, 2013, 2014; Liu et al., 2015; Wiegand et al., 2016). A structured opinion includes the key arguments of one opinion, such as expressions, holders and targets (Breck et al., 2007; Yang and Cardie, 2012, 2013; Katiyar and Cardie, 2016). Here we focus on opinion role labeling (ORL) (Marasović and Frank, 2018), which identifies opinion holders and targets assuming that the opinion expressions are given. Figure 1 shows an example of the task.

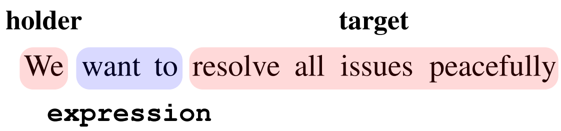

Figure 1: Examples of fine-grained opinion mining.

The focused task behaves very similar with semantic role labeling (SRL) which identifies the core semantic roles for given predicates. Earlier work attempts to exploit a well-trained SRL model to recognize possible semantic roles for a given opinion expression, and then map the semantic roles into opinion roles (Kim and Hovy, 2006; Ruppenhofer et al., 2008). The heuristic approach is unable to obtain high performance for ORL because there are large mismatches between SRL and ORL. For example, opinion expressions are different from verb/noun predicates in SRL, and meanwhile, opinion holders and targets may not always correspond to semantic agents (ARG0) and patients (ARG1), respectively.

We can exploit machine learning based method to solve the mismatching problem between ORL and SRL. With a small number of annotated ORL corpus, we can feed the SRL outputs as inputs to build a statistical model for ORL. By this way, the model can learn the consistencies and inconsistencies between SRL and ORL, arriving at a full exploration of SRL. The method is essentially a feature-based method, treating SRL outputs as a source of features for ORL. The main drawback of the method is that direct exploration of SRL outputs may lead to the error propagation problem. SRL errors can be further propagated into ORL outputs, resulting in degraded ORL performance.

In this work, we propose a simple and novel

*Corresponding author.

Proceedings of NAACL-HLT 2019, pages 641–646

method by using implicit semantic-aware word representations from SRL to enhance ORL. The method is referred to as **SRL-SAWR** for brief. Thanks to the recent advances of encoder-decoder neural SRL models (Zhou and Xu, 2015; He et al., 2017), we can extract implicit vectorized features from the intermediate encoder module instead, avoiding the direct exploration of the final one-best SRL outputs. The vectorized features from the encoder part are implicit semantic-aware representations for input sentences. By taking the semantic-aware representations from SRL as ORL inputs, we are able to make use of SRL information and meanwhile alleviate the error propagation problem.

Here we exploit a neural conditional random field (CRF) model with deep bi-directional long short-term memory networks (Bi-LSTMs) as a baseline, most of which is borrowed from Katiyar and Cardie (2016) and Marasović and Frank (2018). Our preliminary experiments show that the model is able to achieve state-of-the-art performances for both ORL and SRL. Based on this model, we study the proposed implicit semantic-aware word representations for ORL. In addition, we compare this method with two other representative methods of SRL integration as well: one uses discrete SRL outputs as features directly for ORL and the other one exploits a multi-task-learning (MTL) framework to benefit ORL by SRL information.

Experiments are conducted on the MPQA 2.0 dataset, which is a standard benchmark for opinion mining. Results show that SRL is highly effective for ORL, which is consistent with previous findings (Kim and Hovy, 2006; Ruppenhofer et al., 2008; Marasović and Frank, 2018). Meanwhile, our implicit SRL-SAWR method can achieve the best ORL performance, 2.23% higher F-scores than the second best method. All the codes and datasets are released publicly available for research purpose under Apache Licence 2.0 at https://github.com/zhangmeishan/SRL4ORL.

2 Method

2.1 Baseline

ORL aims to identify important opinion arguments for a given opinion expression. The task can be modeled as a sequence labeling problem, similar to SRL (Zhou and Xu, 2015; He et al., 2017). We adopt the {BMESO} schema to convert opinion arguments into a sequence of boundary tags for each word, where B, M and E denote the beginning, middle and ending words of an argument, S denotes a single-word argument, and O denotes the other words. Formally, given a sentence $w_1 \cdots w_n$ and a span of opinion expression $w_s \cdots w_e (1 \leq s \leq e \leq n)$, we aim to assign each word in the sentence by a tag, outputting $t_1 \cdots t_n$.

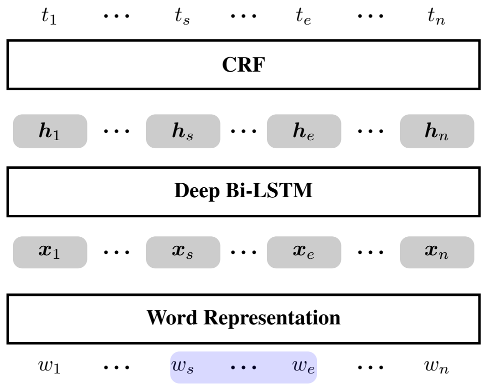

Figure 2: The overall architecture of the baseline.

Inspired by Katiyar and Cardie (2016) and Marasović and Frank (2018), we exploit a deep Bi-LSTM CRF model as the baseline. Figure 2 shows the overall architecture of the baseline model. This model can achieve state-of-the-art performances for both ORL and SRL, which facilitates our study. The key components of the baseline model include three parts: word representation, the deep Bi-LSTM encoder and the CRF decoder. The word representation takes sequential words and opinion expressions as input, mapping them into dense-valued feature vectors $\boldsymbol{x}_1 \cdots \boldsymbol{x}_n$. Following we extract high-level neural features based on the vectors by deep Bi-LSTM, arriving at $\boldsymbol{h}_1 \cdots \boldsymbol{h}_n$. And finally a CRF decoder is applied to output the ORL results $t_1 \cdots t_n$.

2.2 SRL Integration

SRL aims to find the core semantic arguments for a given predicate, which is highly correlative with the ORL task. The semantic roles agent (ARG0) and patient (ARG1) are often corresponding to the opinion holder and target, respectively. Several works even directly transfer semantic roles into opinion roles for ORL (Kim and Hovy, 2006; Ruppenhofer et al., 2008), treating opinion expressions as the major predicates. These systems can achieve good performances, indicating that SRL information can be greatly useful for ORL.

Here we propose a novel method to encode

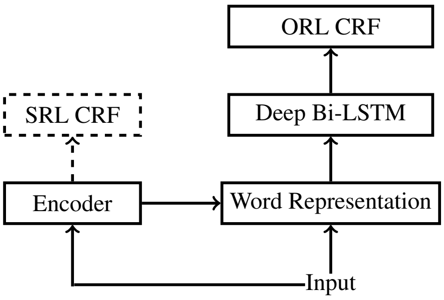

Figure 3: SRL integration methods for ORL.

the SRL information implicitly, enhancing ORL model with semantic-aware word representations from a neural SRL model (SRL-SAWR). Figure 3 shows the overall architectures of our SRL integration method. Instead of using the discrete outputs from the SRL model, the SRL-SAWR method exploits the intermediate encoder outputs as inputs for ORL, which can alleviate the problems in the above two methods. On the one hand, we do not rely on the discrete outputs of a well-trained SRL, reducing the error prorogation problem. And on the other hand, we handle ORL and SRL separately, avoiding the model structure dependencies between the two tasks.

We assume that the external SRL system is a neural-based encoder-decoder model. For fair comparisons with FS-MTL, here we use the same deep Bi-LSTM CRF model for SRL as well. Thus the encoder outputs are the hidden vectors from deep Bi-LSTMs. Assuming that the dumped hidden vector sequence from the SRL encoder is $\boldsymbol{h}_1^{\text{SRL}} \cdots \boldsymbol{h}_n^{\text{SRL}}$, we integrate it into the ORL model by the following equation:

$$\boldsymbol{x}_i^* = \boldsymbol{x}_i \oplus \boldsymbol{W}_{\text{SRL}} \boldsymbol{h}_i^{\text{SRL}}, \quad (1)$$

where $\boldsymbol{W}_{\text{SRL}}$ is a projection matrix which is a model parameter, $\boldsymbol{x}_i$ is the baseline word representation of word w_i, and $\boldsymbol{x}_i^*$ is the new word representation, which will be further fed into the deep Bi-LSTM layer of the ORL model. Noticeably, the model parameters of the SRL encoder are also fine tuned according to the ORL objective, as the preliminary results indicate that fine-tuning can bring better performance.

3 Experiments

3.1 ORL Data

We exploit the MPQA version 2.0 corpus (Wiebe et al., 2005; Wilson, 2008) to evaluate our models,[1] which has been widely adopted as a benchmark dataset for opinion mining (Yang and Cardie, 2013; Katiyar and Cardie, 2016; Marasović and Frank, 2018). There are 482 documents in the dataset. Following these work, we set aside 132 documents as the development set, and the remaining 350 documents are used as the test set in our experiments. We conduct experiments using five-fold cross-validation (CV) on the test set at the document level. Following Marasović and Frank (2018), we focus on opinion holders and targets only. The gold standard opinion expressions, holders and targets correspond to the direct subjective annotations, agent annotations and target annotations, respectively.

3.2 Evaluation Metrics

We use recall (R), precision (P) and their F1-measure value to measure our proposed models. The average values of the five-fold CV results are reported in this work. We exploit exact matching as the major metric. Following Marasović and Frank (2018), two kinds of soft evaluation methods are also adopted for evaluation, namely binary and proportional overlapping, Binary overlap treats an entity as correct if it contains an overlapped region with the gold-standard entity, and the proportional overlap assigns a partial score proportional to the ratio of the overlapped region.

3.3 Setting

There are several hyper-parameters to define our neural network structures. We simply set their values according to previous work (He et al., 2017; Marasović and Frank, 2018), without much tuning work. Concretely, we set the dimension size of all embeddings to 100, the output hidden size of LSTMs to 200 and the layer number of Bi-LSTM to 3. For external word embeddings, we use the pretrained 100-dimensional glove embeddings (Pennington et al., 2014).

We exploit online training to learn model parameters, and train on the entire training instances for 40 epochs, choosing the best-epoch model according to the performance on the development corpus. We use Adam (Kingma and Ba, 2014) with a learning rate 10^{-3} to update model parameters, and use gradient clipping by a max norm 1.0 and l_2-regularization by a parameter 10^{-8}. We apply dropout with a ratio of 0.2 over word represen-

[1]Available at http://www.cs.pitt.edu/mpqa.

tations, output layers of Bi-LSTMs to avoid overfitting (Srivastava et al., 2014).

3.4 SRL

For SRL, we use the large-scale dataset of CoNLL-2012 shared task, which is extracted from OntoNotes v5.0 corpus. The description and separation of train, development and test data set can be found in Pradhan et al. (2013). The training corpus contains over 250K predicates, which is much larger than the number of opinion expressions in the ORL training corpus (averaged 3.6K).

We exploit the same neural network model as the ORL for SRL, in order to make fair comparisons between our proposed model with FS-MTL. According to the preliminary experiments, the SRL model can reach an F-measure of 81.8%, which is comparable to the reported result (81.7%) in He et al. (2017).

3.5 Results

Table 1 shows the final results on the test dataset. We report the overall as well as the fine-grained performance in term of opinion arguments (i.e., holder and target). Compared with the baseline system, our final SRL-SAWR model can bring significantly better results ($p < 10^{-5}$ under pairwise t-test). For fine-grained evaluations, the final model outperforms the baseline model consistently on opinion holders and targets. The tendencies are similar by exploiting the binary and proportional matching methods. The results show that SRL information is very helpful for ORL, which is consistent with previous studies (Kim and Hovy, 2006; Ruppenhofer et al., 2008; Marasović and Frank, 2018). The implicit SRL-SAWR method is highly effective to integrate SRL information into the ORL model.

Further, we compare the SRL-SAWR method with two other methods as well, namely SRL-TE and FS-MTL, respectively. The SRL-TE approach simply exploits the output SRL tags as inputs for ORL, embedding them as an additional source of word representations. The FS-MTL approach is exactly the proposed model by Marasović and Frank (2018). As shown in Table 1, all three methods can bring improved performance by integrating SRL, further demonstrating that SRL is indeed valuable for ORL. In addition, the SRL-SAWR method can achieve the best performance among the three methods, obtaining further significant improvements by at least $63.74 - 61.51 = 2.23$ points on overall F1-measure with exact matching ($p < 10^{-4}$). For fine-grained evaluations, the SRL-SAWR method can also give the best performance. The results demonstrate that SRL-SAWR is most effective to integrate the SRL information into a neural ORL model. The two methods, SRL-TE and FS-MTL, are comparable by evaluations based on the exact matching.

Model	Holder	Target	Overall
Exact F1			
Baseline	73.07	42.70	58.30
SRL-SAWR	**76.95**	**50.50**	**63.74**
SRL-TE	75.89	46.27	61.46
FS-MTL	75.58	46.40	61.51
Binary F1			
Baseline	81.57	68.34	75.15
SRL-SAWR	**84.91**	**73.29**	**79.10**
SRL-TE	83.47	68.79	76.33
FS-MTL	83.80	72.06	77.87
Proportional F1			
Baseline	79.35	61.22	70.55
SRL-SAWR	**82.82**	**67.31**	**75.08**
SRL-TE	81.56	64.74	72.40
FS-MTL	81.67	65.18	73.61

Table 1: Final results on the test dataset.

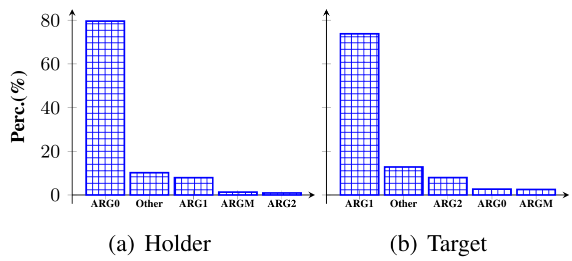

Figure 4: Percentages with respect to semantic roles.

3.6 Analysis

In this section, we conduct several experimental analysis on the test dataset to deeply understand the effectiveness of SRL information.

First, we examine the relationship between SRL and ORL. SRL identifies the semantic arguments for predicates, and ORL recognizes the opinion arguments for opinion expressions. Intuitively, in most cases, the opinion holders are corresponding to semantic agents/ARG0 of opinion triggers/expressions, and similarly, the opinion targets are usually corresponding to patients/ARG1. Figure 4 shows the percentages of opinion hold-

Model	Holder	Target	Overall
Baseline	**73.07**	**42.70**	**58.30**
SRL Mapping	68.56	25.33	46.29

Table 2: The performance of the SRL mapping method.

Model	Holder	Target	Overall
Consistent arguments			
Baseline	87.63	61.67	80.87
SRL-SAWR	**88.72**	**67.58**	**82.87**
SRL Mapping	82.57	40.36	63.77
SRL-TE	88.61	63.94	81.88
FS-MTL	88.16	66.80	82.28
Inconsistent arguments			
Baseline	42.90	36.28	38.37
SRL-SAWR	**49.09**	**44.24**	**44.65**
SRL Mapping	0.00	0.00	0.00
SRL-TE	42.30	39.23	40.14
FS-MTL	43.47	39.04	40.25

Table 3: Comparisons in terms of the consistent/inconsistent arguments between SRL and ORL.

ers/targets being corresponding to semantic roles, which are calculated according to the word-level mapping over the 1-best SRL outputs and the gold-standard ORL tags. We list only the five semantic roles with highest mapping percentages. As shown, the results are consistent with our intuition. Thus SRL and ORL are highly correlative. Considering the much larger scale of annotated SRL corpora, SRL can benefit ORL potentially.

According to the above findings, we design a simple system by mapping SRL outputs into ORL directly (Kim and Hovy, 2006; Ruppenhofer et al., 2008). We simply convert the semantic role ARG0 into holder, and ARG1 into target. Table 2 shows the performance. The results of the baseline system are shown for comparison. We can see that the simple mapping method is also one feasible alternative as a whole.

Further, we compare the SRL utilization capabilities of our proposed method and the other SRL-enhanced ORL systems, including the above SRL Mapping method. We categorize the opinion arguments by whether they can be directly mapped from the SRL outputs. The opinion arguments which can be directly mapped from SRL, referred to as consistent arguments, should be more easily identified by SRL enhanced models than the remaining inconsistent arguments. Table 3 shows the comparison results. We can see that all SRL-

gold	**holder**			**target**
	The white house	is said to be **embarrassed** by		the report
SRL	**ARG1**			**ARG0**
SRL-TE	**target**			**holder**
FS-MTL	**target**			**holder**
SRL-SAWR	**holder**			**target**

Figure 5: One example for case study.

enhanced supervised models can achieve better performances for consistent arguments. For the inconsistent arguments, the tendency is similar, except the holder performance of SRL-TE. In addition, our method can gain much larger improvements, which indicates that our method can better handle the inconsistencies between SRL and ORL.

Finally, we show one case study to illustrate the advantage of our SRL-SAWR method. Figure 5 shows one example. As shown, the SRL argument ARG0, which is more probably mapped onto holder, is annotated by target in the example. The SRL argument ARG1 is labeled as opinion holder, which is also one inconsistent case. Compared with SRL-TE and FS-MTL, our model can better handle these inconsistent cases. The observation further confirms our results in Table 3.

4 Conclusion

We proposed a simple and novel method (SRL-SAWR) to enhance ORL with SRL information by exploiting implicit semantic-aware word representations from SRL. The main idea is to export intermediate SRL encoder outputs as inputs to better word representations of an ORL model. This method does not impose any extra requirement for ORL, and meanwhile avoids the error proropagation problem from discrete SRL outputs. We conducted experiments to verify our method on a benchmark MPQA dataset. The results showed that our method can exploit SRL information effectively. We compared the proposed method with SRL-TE and FS-MTL, which are two representative approaches to enhance ORL by SRL. The results demonstrated our method can bring the best performance among the three approaches.

Acknowledgments

We thank all reviewers for their valuable comments. This work is supported by National Natural Science Foundation of China (NSFC) grants U1836222, 61602160 and 61672211.

References

Eric Breck, Yejin Choi, and Claire Cardie. 2007. Identifying expressions of opinion in context. In *IJCAI*, volume 7, pages 2683–2688.

Luheng He, Kenton Lee, Mike Lewis, and Luke Zettlemoyer. 2017. Deep semantic role labeling: What works and whats next. In *Proceedings of ACL)*, pages 473–483.

Ozan Irsoy and Claire Cardie. 2013. Bidirectional recursive neural networks for token-level labeling with structure. *arXiv preprint arXiv:1312.0493*.

Ozan Irsoy and Claire Cardie. 2014. Opinion mining with deep recurrent neural networks. In *EMNLP*, pages 720–728.

Arzoo Katiyar and Claire Cardie. 2016. Investigating lstms for joint extraction of opinion entities and relations. In *ACL*, pages 919–929.

Soo-Min Kim and Eduard Hovy. 2006. Extracting opinions, opinion holders, and topics expressed in online news media text. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, pages 1–8.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Pengfei Liu, Shafiq Joty, and Helen Meng. 2015. Fine-grained opinion mining with recurrent neural networks and word embeddings. In *EMNLP*, pages 1433–1443.

Ana Marasović and Anette Frank. 2018. Srl4orl: Improving opinion role labeling using multi-task learning with semantic role labeling. In *Proceedings of NAACL*, pages 583–594.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 EMNLP*, pages 1532–1543.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards robust linguistic analysis using ontonotes. In *Proceedings of the CoNLL*, pages 143–152.

Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. 2011. Opinion word expansion and target extraction through double propagation. *Computational Linguistics*, 37(1):9–27.

Josef Ruppenhofer, Swapna Somasundaran, and Janyce Wiebe. 2008. Finding the sources and targets of subjective expressions. In *LREC*.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *LRE*, 39(2-3):165–210.

Michael Wiegand, Christine Bocionek, and Josef Ruppenhofer. 2016. Opinion holder and target extraction on opinion compounds – a linguistic approach. In *Proceedings of NAACL*, pages 800–810.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2009. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational linguistics*, 35(3):399–433.

Theresa Ann Wilson. 2008. *Fine-grained subjectivity and sentiment analysis: recognizing the intensity, polarity, and attitudes of private states*.

Bishan Yang and Claire Cardie. 2012. Extracting opinion expressions with semi-markov conditional random fields. In *EMNLP-CoNLL*, pages 1335–1345. Association for Computational Linguistics.

Bishan Yang and Claire Cardie. 2013. Joint inference for fine-grained opinion extraction. In *ACL (1)*, pages 1640–1649.

Jie Zhou and Wei Xu. 2015. End-to-end learning of semantic role labeling using recurrent neural networks. In *Proceedings of ACL*, pages 1127–1137.