# SRL4ORL: Improving Opinion Role Labeling Using Multi-Task Learning With Semantic Role Labeling

**Ana Marasović** and **Anette Frank**
Research Training Group AIPHES
Department of Computational Linguistics
Heidelberg University
`{marasovic,frank}@cl.uni-heidelberg.de`

## Abstract

For over a decade, machine learning has been used to extract *opinion-holder-target structures* from text to answer the question *Who expressed what kind of sentiment towards what?*. Recent neural approaches do not outperform the state-of-the-art feature-based models for Opinion Role Labeling (ORL). We suspect this is due to the scarcity of labeled training data and address this issue using different multi-task learning (MTL) techniques with a related task which has substantially more data, i.e. Semantic Role Labeling (SRL). We show that two MTL models improve significantly over the single-task model for labeling of both holders and targets, on the development and the test sets. We found that the vanilla MTL model which makes predictions using only shared ORL and SRL features, performs the best. With deeper analysis we determine what works and what might be done to make further improvements for ORL.

## 1 Introduction

Fine-Grained Opinion Analysis (FGOA) aims to: (i) detect opinion expressions (O) that convey attitudes such as sentiments, agreements, beliefs or intentions (like *feared* in example (1)), (ii) measure their intensity (e.g. strong), (iii) identify their holders (H), i.e. entities that express an attitude (e.g. *it*), (iv) identify their targets (T), i.e. entities or propositions at which the attitude is directed (e.g. *violence*) and (v) classify their target-dependent attitude (e.g. negative sentiment)[1].

(1) Australia said [it]$_H$ [**feared**]$_{O_{neg}}$ [violence]$_T$ if voters thought the election had been stolen.

As the commonly accepted benchmark corpus MPQA (Wiebe et al., 2005) uses span-based annotations to represent *opinion entities* (opinions,

---

[1]Examples are drawn from MPQA (Wiebe et al., 2005).

holders and targets), the task is usually approached with sequence labeling techniques and the BIO encoding scheme (Choi et al., 2006; Yang and Cardie, 2013; Katiyar and Cardie, 2016). Initially pipeline models were proposed which first predict opinion expressions and then, given an opinion, label its *opinion roles*, i.e. holders and targets (Kim and Hovy, 2006; Johansson and Moschitti, 2013). Pipeline models have been substituted with so-called joint models that simultaneously identify all opinion entities, and predict which opinion role is related to which opinion (Choi et al., 2006; Yang and Cardie, 2013; Katiyar and Cardie, 2016). Recently an LSTM-based joint model was proposed (Katiyar and Cardie, 2016) that unlike the prior work (Choi et al., 2006; Yang and Cardie, 2013) does not depend on external resources (such as syntactic parsers or named entity recognizers). The neural variant does not outperform the feature-based CRF model (Yang and Cardie, 2013) in Opinion Role Labeling (ORL).

Both the neural and the CRF joint models achieve about 55% F1 score for predicting which targets relate to which opinions in MPQA. Thus, these models are not yet ready to answer the question this line of research is usually motivated with: *Who expressed what kind of sentiment towards what?*. Our goal is to investigate the limitations of neural models in solving different subtasks of FGOA on MPQA and to gain a better understanding of what is solved and what is next.

We suspect that one of the fundamental obstacles for neural models trained on MPQA is its small size. One way to address scarcity of labeled data is to use multi-task learning (MTL) with appropriate auxiliary tasks. A promising auxiliary task candidate for ORL is Semantic Role Labeling (SRL), the task of predicting predicate-argument structure of a sentence, which answers the question *Who did what to whom, where and when?*.

| | Australia | said | it | feared | violence | if | voters | thought | the | election | had | been | stolen | . |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| say.01 | A0 | - | A1 | A1 | A1 | A1 | A1 | A1 | A1 | A1 | A1 | A1 | A1 | - |
| fear.01 | - | - | **A0** | - | **A1** | AM-ADV | AM-ADV | AM-ADV | AM-ADV | AM-ADV | AM-ADV | AM-ADV | AM-ADV | - |
| think.01 | - | - | - | - | - | - | A0 | - | A1 | A1 | A1 | A1 | A1 | - |
| steal.01 | - | - | - | - | - | - | - | - | A1 | A1 | - | - | - | - |

Table 1: Output of the SRL demo.

Table 1 illustrates the output of the SRL demo[2] for example (1), following the PropBank SRL scheme (Palmer et al., 2005)[3].

**SRL4ORL.** The semantic roles of the predicate *fear* (marked blue bold) correspond to the opinion roles H and T, according to MPQA. For this reason, the output of SRL systems has been commonly used for feature-based FGOA models (Kim and Hovy, 2006; Johansson and Moschitti, 2013; Choi et al., 2006; Yang and Cardie, 2013). Additionally, a considerable amount of training data is available for training SRL models (Table 2 in Sec. 3), which made neural SRL models successful (Zhou and Xu, 2015; Yang and Mitchell, 2017).

**Obstacles.** Although SRL is similar in nature to ORL, it cannot solve ORL for all cases (Ruppenhofer et al., 2008). In example (2) holder and target of the predicate *please* correspond to A1, A0 semantic roles respectively, wheres for the predicate *fear* in (1) holder and target correspond to A0, A1 respectively. We took into account this observation when deciding on an appropriate MTL model by splitting its parameters into shared and task-specific ones (i.e. hard-parameter sharing).

(2) $[I]_H^{A1}$ am very [**pleased**]$_{O_{pos}}$ that [the Council has now approved the Kyoto Protocol thus enabling the EU to proceed with its ratification]$_T^{A0}$.

A further obstacle for properly exploiting SRL training data with MTL could be specificities, inconsistency and incompleteness of the MPQA annotations. In example (3), *Rice* expressed his negative sentiment towards *the three countries in question* by *setting the criteria* which states something negative about those countries: they *are repressive and grave human rights violators [...]*. In this case, the model should not pick any local semantic role for the target.

(3) The criteria [**set by**]$_{O_{neg}}$ [Rice]$_H$ are the following: [the three countries in question]$_T$ are

repressive and grave human rights violators, and aggressively seeking weapons [...].

In examples (4–5), the same opinion expression *concerned* realizes different scopes for the target. A model which exploits SRL knowledge could be biased to always label targets as complete SRL role constituents, as in example (5).

(4) Rice told us [the administration]$_H$ was [**concerned**]$_{O_{neg}}$ that [Iraq]$_T$ would take advantage of the 9/11 attacks.

(5) [The Chinese government]$_H$ is deeply [**concerned**]$_{O_{neg}}$ about [the sudden deterioration in the Middle East situation]$_T$, Tang said.

Regarding incompleteness, prior work (Katiyar and Cardie, 2016) has shown that their model makes reasonable predictions in sentences which do not have annotations at all, e.g. [mothers]$_H$ [**care**]$_O$ for [their young]$_T$, in: *From the fact that mothers care for their young, we can not deduce that they ought to do so, Hume argued.*

The examples above show that incorporating SRL knowledge via multi-task learning is a reasonable way to improve ORL, but at the same time they alert us that given the specificities of MPQA and ORL annotations in general, it is not obvious whether MTL can overcome divergences in the annotation schemes of opinion and semantic role labeling. We investigate this research question by adopting one of the recent successful architectures for SRL (Zhou and Xu, 2015) and experiment with different multi-task learning frameworks.

Our contributions are: (i) we adapt a recently proposed neural SRL model for ORL, (ii) we enhance the model using different MTL techniques with SRL to tackle the problem of scarcity of labeled data for ORL, (iii) we show that most of the MTL models improve the single-task model for labeling of both holders and targets on development and test sets, and two of them make yield significant improvements, (iv) by deeper analysis we provide a better understanding of what is solved and where to head next for neural ORL.

---

[2] http://barbar.cs.lth.se:8081
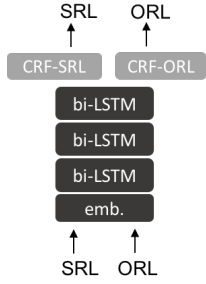[3] Henceforth we use the PropBank SRL framework.
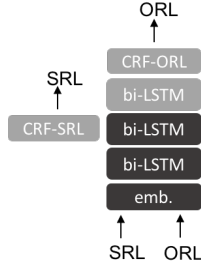
Figure 1: Fully-shared (FS) MTL.
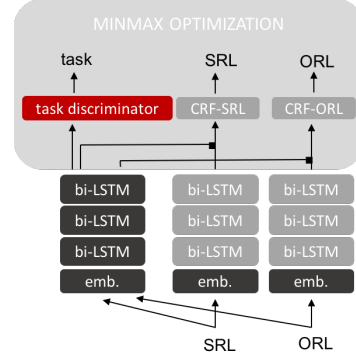


Figure 2: Hierarchical-MTL (H-MTL).



Figure 3: (Adversarial) state-private ((A)SP) MTL.

## 2 Neural MTL for SRL and ORL

Neural multi-task learning (MTL) receives a lot of attention and new MTL architectures emerge regularly. Yet there is no clear consensus which MTL architecture to use in which conditions. We experiment with well-received architectures that could adapt to different cases of ORL from Section 1.

As a general neural architecture for single- and multi-task learning we use the recently proposed SRL model (Zhou and Xu, 2015) (Z&X-STL) which successfully labels semantic roles without any syntactic guidance. This model consists of a stack of bi-directional LSTMs and a CRF which makes the final prediction. The inputs to the first LSTM are not only token embeddings but three additional features: embedding of the predicate, embedding of the context of the predicate and an indicator feature (1 if the current token is in the predicate context, 0 otherwise). Thus, every sentence is processed as many times as there are predicates in it. Adapting this model for labeling of opinion roles is straightforward, the only difference being that opinion expressions can be multi-words and only two opinion roles are assigned.

MTL techniques aim to learn several tasks jointly by leveraging knowledge from all tasks. In the context of neural networks, MTL is commonly used in such a way that it is predefined which layers have tied parameters and which are task-specific (i.e. hard-parameter sharing). There are various ways of defining which parameters should be shared and how to train them.

**Fully-shared (FS) MTL model.** A fully-shared model (Fig. 1) shares all parameters of the general model except the output layer. Each task has a task-specific output layer which makes the prediction based on the representation produced by the final LSTM. When training on a mini-batch of a certain task, parameters of the output layer of the other tasks are not updated. This model should be effective for constructions with a clear mapping between opinion and semantic roles such as {H ↦ A0, T ↦ A1} as in example (1) (Sec. 1).

**Hierarchical MTL (H-MTL) model.** For NLP applications, often some given (high-level) task is supposed to benefit from another (low-level) task more than the other way around, e.g. parsing from POS tagging. This intuition lead to designing hierarchical MTL models (Søgaard and Goldberg, 2016; Hashimoto et al., 2017) in which predictions for low-level tasks are not made on the basis of the representation produced at the final LSTM, but on the representation produced by a lower-layer LSTM (Fig. 2). Task-specific layers atop shared layers could potentially give the model more power to distinguish or ignore certain semantic roles. If so, this MTL model is more suitable for examples like (2) and (3) (Sec. 1).

**Shared-private (SP) MTL model.** In the state-private model, in addition to the stack of shared LSTMs, each task has a stack of task-specific LSTMs (Liu et al., 2017) (Fig. 3). Representations at the outermost shared LSTM and the task-specific LSTM are concatenated and passed to the task-specific output layer. The ORL representation produced independently from SRL gives the model the ability to utilize the shared and entirely task-specific information. For labeling of targets, it is expected that for examples (1) & (5) the model relies mostly on the shared representation, for examples (2) & (4) on both shared and ORL-specific representations, and for example (3) solely on the ORL-specific representation.

**Adversarial shared-private (ASP) model.** The limitation of the SP model is that it does not prevent the shared layers from capturing task-specific features. To ensure this, ASP extends the

| | task | train size | dev size | test size | $|\mathcal{Y}|$ |
|---|---|---|---|---|---|
| CoNLL'05 | SRL | 90750 | 3248 | 6071 | 106 |
| MPQA (4-CV) | ORL | 3141.25 | 1055 | 1036.75 | 7 |
| MPQA (10-CV) | ORL | 3516.3 | 1326 | 349.3 | 7 |

Table 2: Datasets w/ nb. of SRL predicates/ORL opinions in train, dev & test set, size of label inventory.

SP model with a *task discriminator* (Liu et al., 2017). The task discriminator (Fig. 3, marked red) predicts to which task the current batch of data belongs, based on the representation produced by the shared LSTMs. If the shared LSTMs are task-invariant, the discriminator should perform badly. Thus, we update the shared parameters to maximize the discriminator's cross-entropy loss. At the same time we want the discriminator to challenge the shared LSTMs, so we update the discriminator's parameters to minimize its cross-entropy loss. This minmax optimization is known as *adversarial training* and recently it gained a lot of attention for NLP applications (Liu et al., 2017; Chen et al., 2017; Kim et al., 2017; Qin et al., 2017; Wu et al., 2017; Gui et al., 2017; Li et al., 2017; Zhang et al., 2017; Joty et al., 2017).

## 3 Experimental setup

### 3.1 Datasets

For SRL we use the newswire CoNLL-2005 shared task dataset (Carreras and Màrquez, 2005), annotated with PropBank predicate-argument structures. Sections 2-21 of the WSJ corpus (Charniak et al., 2000) are used for training and section 24 as dev set. The test set consists of section 23 of WSJ and 3 sections of the Brown corpus.

For ORL we use the manually annotated MPQA 2.0. corpus (Wiebe et al., 2005; Wilson, 2008). It mostly contains news documents, but also travel guides, transcripts of spoken conversations, e-mails, fundraising letters, textbook chapters and translations of Arabic source texts.

We report detailed pre-processing of MPQA[4] and data statistics in the Appendix.

### 3.2 Evaluation metrics

For both tasks we adopt evaluation metrics from prior work. For SRL, precision is defined as the proportion of semantic roles predicted by a system

---

[4]Examples how to use our scripts can be found at https://github.com/amarasovic/naacl-mpqa-srl4orl/blob/master/generate_mpqa_jsons.py.

which are correct, recall is the proportion of gold roles which are predicted by a system, F1 score is the harmonic mean of precision and recall.

In case of ORL, we report 10-fold CV[5] and repeated 4-fold CV with *binary F1 score* and *proportional F1 score*, for holders and targets separately. *Binary precision* is defined as the proportion of predicted holders (targets) that overlap with the gold holder (target), *binary recall* is the proportion of gold holders (targets) for which the model predicts an overlapping holder (target). *Proportional recall* measures the proportion of the overlap between a gold holder (target) and an overlapping predicted holder (target), *proportional precision* measures the proportion of the overlap between a predicted holder (target) and an overlapping gold holder (target). F1 scores are the harmonic means of (the corresponding) precision and recall.

### 3.3 Training details

We evaluate our models using two evaluation setting. First, we follow Katiyar and Cardie (2016) which set aside 132 documents for development and used the remaining 350 documents for 10-fold CV. However, in the 10-fold CV setting, the test sets are more than 3 times smaller than the dev set (Table 2, row 3), and, consequently, results in high-variance estimates on the test sets. Therefore we additionally evaluate our models with 4-fold CV. We set aside 100 documents for development and use 25% of the remaining documents for testing. The resulting test sets are comparable in size to the dev set (Table 2, row 2). We run 4-fold CV twice with two different random seeds. We do not tune hyperparameters (HPs), but follow suggestions proposed in the comprehensive HPs study for sequence labeling tasks in Reimers and Gurevych (2017). All HPs can be found in the Appendix.

## 4 Results

We evaluate all models after every $\left\lceil \frac{\text{train size}}{\text{batch size}} \right\rceil$ iteration on the ORL dev set and save them if they achieve a higher arithmetic mean of proportional F1 scores of holders and targets on the ORL dev set. The saved models are used for testing.

We report the mean of F1 scores over 10 folds and the standard deviation (appears as a subscript) of all models in Table 3. We report the mean of F1 scores over 4 folds and 2 different seed (8 evalua-

---

[5]We used the same splits as the prior work (Katiyar and Cardie, 2016). We thank the authors for providing the splits.

|  | dev (MPQA) | | | | |  | test (MPQA) | | | |
|---|---|---|---|---|---|---|---|---|---|---|
|  | holder | | target | | |  | holder | | target | |
|  | binary F1 | prop. F1 | binary F1 | prop. F1 |  |  | binary F1 | prop. F1 | binary F1 | prop. F1 |
| Z&X-STL | $80.15_{1.10}$ | $76.87_{1.26}$ | $74.62_{0.67}$ | $70.23_{1.04}$ |  | Z&X-STL | $80.24_{2.91}$ | $77.98_{2.90}$ | $76.30_{2.55}$ | $71.18_{2.55}$ |
| FS-MTL | $83.68^{\bullet}_{0.44}$ | $81.45^{\bullet}_{0.58}$ | $\mathbf{76.23}^{\bullet}_{0.75}$ | $\mathbf{73.01}^{\bullet}_{0.93}$ |  | FS-MTL | $83.47^{\bullet}_{2.26}$ | $81.80^{\bullet}_{2.26}$ | $\mathbf{77.60}^{\bullet}_{2.52}$ | $\mathbf{73.77}^{\bullet}_{2.28}$ |
| H-MTL | $\mathbf{84.14}^{\bullet}_{0.72}$ | $\mathbf{81.86}^{\bullet}_{0.48}$ | $76.11^{\bullet}_{0.61}$ | $72.55^{\bullet}_{0.73}$ |  | H-MTL | $\mathbf{84.03}^{\bullet}_{2.65}$ | $\mathbf{82.34}^{\bullet}_{2.51}$ | $77.41_{2.14}$ | $73.10_{1.96}$ |
| SP-MTL | $82.18^{\bullet\diamond}_{0.89}$ | $79.66^{\bullet\diamond}_{0.72}$ | $74.99^{\diamond}_{1.17}$ | $71.32^{\diamond}_{1.81}$ |  | SP-MTL | $82.19^{\bullet}_{2.49}$ | $80.11^{\bullet\diamond}_{2.36}$ | $76.01_{3.03}$ | $71.51_{3.34}$ |
| ASP-MTL | $82.63^{\bullet\diamond}_{0.84}$ | $80.20^{\bullet\diamond}_{0.99}$ | $74.24^{\bullet\diamond}_{0.58}$ | $70.16^{\diamond}_{1.29}$ |  | ASP-MTL | $83.15^{\bullet}_{2.92}$ | $81.12^{\bullet}_{2.66}$ | $75.89_{2.66}$ | $71.21_{2.78}$ |

Table 3: ORL 10-fold CV results.

|  | dev (MPQA) | | | | |  | test (MPQA) | | | |
|---|---|---|---|---|---|---|---|---|---|---|
|  | holder | | target | | |  | holder | | target | |
|  | binary F1 | prop. F1 | binary F1 | prop. F1 |  |  | binary F1 | prop. F1 | binary F1 | prop. F1 |
| Z&X-STL | $79.73_{1.19}$ | $77.06_{1.14}$ | $76.09_{0.94}$ | $70.45_{1.07}$ |  | Z&X-STL | $80.42_{1.92}$ | $77.48_{2.06}$ | $73.84_{1.17}$ | $67.03_{2.13}$ |
| FS-MTL | $\mathbf{83.58}^{\bullet\diamond}_{0.69}$ | $\mathbf{82.16}^{\bullet}_{0.59}$ | $\mathbf{78.32}^{\bullet}_{1.57}$ | $\mathbf{75.09}^{\bullet}_{2.27}$ |  | FS-MTL | $\mathbf{83.67}^{\bullet}_{1.52}$ | $\mathbf{81.59}^{\bullet}_{1.50}$ | $77.04^{\bullet}_{1.45}$ | $73.01^{\bullet}_{2.53}$ |
| H-MTL | $82.36^{\bullet\diamond}_{0.81}$ | $80.84^{\bullet\diamond}_{0.98}$ | $78.11^{\bullet}_{0.82}$ | $74.89^{\bullet}_{1.33}$ |  | H-MTL | $82.80^{\bullet}_{1.87}$ | $80.40^{\bullet}_{1.91}$ | $\mathbf{77.12}^{\bullet}_{1.34}$ | $\mathbf{73.16}^{\bullet}_{1.78}$ |
| SP-MTL | $82.21^{\bullet\diamond}_{0.79}$ | $80.23^{\bullet\diamond}_{0.88}$ | $76.14^{\diamond}_{1.18}$ | $71.14^{\diamond}_{0.97}$ |  | SP-MTL | $82.51^{\bullet}_{2.17}$ | $80.03^{\bullet}_{2.00}$ | $74.61^{\bullet}_{1.32}$ | $68.70^{\diamond}_{2.32}$ |
| ASP-MTL | $81.41^{\diamond}_{1.27}$ | $79.39^{\bullet\diamond}_{1.45}$ | $76.49_{1.39}$ | $72.13^{\bullet}_{1.87}$ |  | ASP-MTL | $81.77^{\diamond}_{1.74}$ | $79.32^{\diamond}_{1.62}$ | $74.92^{\bullet\diamond}_{0.84}$ | $69.89^{\bullet}_{1.80}$ |

Table 4: ORL repeated 4-fold CV results.

tions) and the standard deviation of all models in Table 4. Evaluation metrics follow Section 3.2.

We mark significant difference between MTL models and the single-task (Z&X-STL) model, observed using a Kolmogorov-Smirnov significance test ($p < 0.05$) (Massey Jr, 1951), with $\bullet$ in superscript and between the FS-MTL model and other MTL models with $\diamond$.

**STL vs. MTL.** In the 10-fold CV evaluation setting (Table 3), the ==FS-MTL and the H-MTL models improve over the Z&X-STL model in all evaluation measures,== for both holders and targets. When evaluated in the repeated 4-fold CV setting (Table 4), all MTL models improve over the Z&X-STL model in all evaluation measures, for both holders and targets.

The FS-MTL and the H-MTL models improve *significantly* in all evaluation measures, for both holders and targets, on both dev and test sets, when evaluated with repeated 4-fold CV. With 10-fold CV the improvements are also significant, except for targets on the test set. This is probably due to the small size of the test sets (Table 2, row 3), which results in a high-variance estimate. Indeed, standard deviations on the 10-fold CV test sets are always much higher compared to the dev set or to the test sets of 4-fold CV.

It is not surprising that larger improvements are visible in the labeling of holders. They are usually short, less ambiguous and often presented with the

A0 semantic role, whereas annotating targets is a challenging task even for humans.[6]

Larger improvements are visible for proportional F1 score than for binary F1 score. That is, more data and SRL knowledge helps the model to better annotate the scope of opinion roles.

**Comparing MTL models.** In Section 2 we introduced MTL models with task-specific LSTM layers hypothesizing that these layers should give MTL models more power to adapt to a variety of potentially problematic cases that we illustrated in the Introduction. ==However, our results show that the FS-MTL model performs significantly better or comparable to MTL models that include task-specific layers.== Reimers and Gurevych (2017) show that MTL is especially sensitive to the selection of HPs. Thus, a firm and solid comparison of the different MTL models requires thorough HP optimization, to properly control the number of parameters and regularization of the models. ==We leave HP optimization for future work.==

## 5 Analysis

Our aim in this section is to ==analyze what the proposed models are good at, in which ways MTL improves over the single-task ORL model and what could be done to achieve further progress.==

We evaluate the FS-MTL and the Z&X-STL

---

[6]Wilson (2008) reports annotator agreement for target labeling of 86.00 binary F1 score.

| | |
|---|---|
| 1 | *Malinga* $_{FS,ZX}$ **said** according to the guidelines in the booklet, the election had been legitimate . |
| 2 | movie um-hum that 's interesting so that was a good movie too well do $you$ $_{FS,ZX}$ **think** we've covered baseball i think so okay well have a good night |
| 3 | *The nation* $_{FS,ZX}$ should certainly **be concerned** about the plans to build a rocket launch pad , work on the infrastructure for which is due to start in 2002 , with launches beginning from 2004 . |
| 4 | Bam on Sunday said $she$ $_{FS,ZX}$ **believed** Zimbabwe's election was not free and fair , adding they were not in line with international standards as well as those of her organisation . |
| 5 | The majority report , endorsed only by the ANC , said *the observer mission* $_{FS,ZX}$ **had noted** that over three million Zimbabweans had cast their votes and this substantially represented the will of the people . |

Table 5: The dev examples for which both models (FS-MTL, Z&X-STL) *correctly* predict the *holder* in 6/8 trials.

| | |
|---|---|
| 1 | Indonesia **has come under pressure** from several quarters to *take tougher action against alleged terrorist leaders* but has played down the threat $_{ZXFS}$ . |
| 2 | Mugabe even talked about his **desire** to *keep safeguarding Zimbabwe 's sovereignty and land* $_{ZX}$ in spirit $_{FS}$ when he dies , a dream which the veteran leader said forced him to sacrifice a bright teaching career in the 1950s to lead [...]. |
| 3 | Under his blueprint , the government **hopes** to *stabilize the economy* through curtailing state expenditure , reforming public enterprises and expanding agriculture $_{FS,ZX}$ . |
| 4 | He said those who thought the election process would be rigged were supporters of the MDC party , adding that they were prejudging and **wanted** to *direct the process* $_{FS,ZX}$ . |
| 5 | People in the rural areas support the ruling party because our party has been genuine on **its policy** on *land reform* $_{FS,ZX}$ . |

Table 6: The dev examples for which both models (FS-MTL, Z&X-STL) *correctly* predict the *target* in 6/8 trials.

| | easy | hard |
|---|---|---|
| % opinions that are predicates | 91.32 | 93.33 |
| % holders that are subjects | 77.84 | 38.79 |
| % holders that are A0 roles | 74.10 | 33.33 |
| avg. distance between holders & opinions | 1.54 | 7.56 |

Table 7: Statistics of *holder* prediction.

| | easy | hard |
|---|---|---|
| % opinions that are predicates | 92.58 | 89.20 |
| % target's heads that are objects | 22.12 | 14.77 |
| % targets that are A1 roles | 70.62 | 42.61 |
| % targets that are A2 roles | 9.00 | 0.57 |
| avg. distance between targets & opinions | 2.29 | 8.46 |

Table 8: Statistics of *target* prediction.

models on the ORL dev set using 4-fold CV repeated twice with different seeds (8 evaluation trials). We say that a model predicts a role of a given opinion expression *correctly* if the model predicts a role that overlaps with the correct role in at least 6 out of 8 evaluation trials. If a model predicts a role that overlaps with the correct role in at most 2 out of 8 trials, we say that the model predicts the role *incorrectly*. The requirement on 6-8 (in)correct predictions reduces the risk of analyzing inconsistent predictions and enables us to draw firmer conclusions. We analyze the following scenarios:

(i) both the FS-MTL model and the Z&X-STL model make correct predictions (Tables 5–6)

(ii) the FS-MTL model makes a correct prediction, while the Z&X-STL makes an incorrect prediction (Tables 11–12)

(iii) both models make wrong predictions (Tables 9–10)

In the following, we categorize predictions in case (i) as *easy cases,* and predictions in case (iii) as *hard cases*.

In Tables 5-6 and 9–12, the opinion expression is bolded, the correct role is italicized, predictions of the FS-MTL model are colored blue (subscript *FS*), predictions of the Z&X-STL model are colored yellow (subscript *ZX*) and green marks predictions where both models agree. For simplicity, we show only holders or targets, although the models predict both roles jointly.

**What works well?** There are 668/1055 instances in the dev set for which both models predict holders correctly, and 663/1055 for targets.

Examples 1–5 in Table 5 suggest that holders that can be properly labeled by both models (*easy*

| | |
|---|---|
| 1 | It would be entirely improper if , in *its* **defense of** Israel _FS_ , the United States continues to exert pressure on [...] . |
| 2 | Indonesia _FS,ZX_ **has come under pressure** from *several quarters* to take tougher action against alleged terrorist leaders but has played down the threat . |
| 3 | Australia should adhere to the *Cardinal Principle of International Law* , which **states** that all nations in the world must first respect and promote the humanitarian interests and progress of all humankind . |
| 4 | *The department* said that it will cost $ 600 for an HIV/AIDS patient per year at this time , and the following years this cost is **expected** to stand at just $ 400/year for one patient as the production of such drugs becomes stable . |
| 5 | The Organisation of African Unity OAU _ZX_ also backed Zimbabwean President Robert Mugabe 's re-election , with its observer team _FS,ZX_ **describing** the poll as " transparent , credible , free and fair " . |
| 6 | Regarding the American proposed Anti-Missile Defense System too , neither Russia , China , Japan , nor even the European Union , had shown any enthusiasm ; rather *they* _FS_ had all _FS,ZX_ **expressed** their reserves on the project . |
| 7 | The president renewed his pledge to thwart terrorist groups _FS,ZX_ *who* **want** to " mate up " with regimes hoping to acquire weapons of mass destruction and said " nations will come with us " if the US-led war on terrorism is extended . |

Table 9: The dev examples for which both models (FS-MTL, Z&X-STL) *incorrectly* predict the *holder* in 6/8 trials.

| | |
|---|---|
| 1 | *State-sanctioned land invasions* , several times **declared** illegal by Zimbabwe 's courts , as well as a drought have disrupted Zimbabwe 's food production and famine is already looming in much of the country . |
| 2 | But he **told** the nation _FS,ZX_ that in spite of *stiff opposition to the agrarian reforms from powerful Western countries , especially the country 's former colonial power of Britain* , he would press ahead to seize farms from whites and [...] . |
| 3 | If the Europeans wish to influence Israel in the political arena – in *a direction* that many in Israel **would support whole-heartedly** – they will not be able to promote their positions in such a manner . |
| 4 | They _FS,ZX_ are fully aware that these are dangerous *individuals* , he **said** during a press conference [...] . |
| 5 | And her little girl **just complained** , " I don't want to *wash the dishes* " . |
| 6 | During *President Bush's speech* , I **thought** of heckling _ZX_ ; ' What are you going to do with the Kyoto Protocol ? _FS_ ' |
| 7 | At first I **didn't want** to apply for it _FS,ZX_ , but the principal called me during the summer months and said , " Sandra the time is running out , you need to apply " . |

Table 10: The dev examples for which both models (FS-MTL, Z&X-STL) *incorrectly* predict the *target* in 6/8 trials.

cases) are subjects of their governing heads or A0 roles. The statistics in Table 7 (col. 1, rows 2–3) supports this observation.[7] In contrast, holders that both models predict incorrectly (*hard* cases) are less frequently subjects or A0 roles (col. 2, rows 2–3). Also, *easy* holders are close to the corresponding opinion expression: the average distance is 1.54 tokens (Table 7, row 4), contrary to the *hard* holders with the average distance of 7.56.

Examples 1–5 in Table 6 suggest that targets that can be properly labeled by both models are objects of their governing heads or A1 roles. Table 8, row 3, shows that the majority of the *easy* targets are indeed A1 roles, in contrast to the *hard* targets. Similar to holders, the *easy* targets are in average 7 tokens closer to the opinion expression.

**What to do for further improvement?** There are 165/1055 instances in the dev set for which

both models predict holders incorrectly, and 176 for targets.

As we have seen so far, many holders that are subjects or A0 roles, and targets that are A1 roles, are properly labeled by both models. However, a considerable amount of such holders and targets are not correctly predicted (Table 7–8, col. 2, rows 2–3). Thus our models do not work flawlessly for all such cases. A distinguishing property of the *hard* cases is the distance of the role from the opinion. Thus, future work should advance the model's ability to capture long-range dependencies.

Examples in Table 9 demonstrate that holders, harder to label with our models, occur with the corresponding opinions in more complicated syntactic constructions. In the first example, the FS-MTL model does not recognize the possessive and is possibly biased towards picking the country (*Isreal*), which occurs immediately after the opinion. In the second example, the opinion expression is

---

[7]The statistics is calculated using the output of mate-tools (Björkelund et al., 2010).

| | |
|---|---|
| 1 | Yoshihisa Murasawa , a management consultant for Booz-Allen & Hamilton Japan Inc. , said his firm $_{FS,ZX}$ will likely be recommending acquisitions of Japanese companies more $_{ZX}$ often to foreign clients in the future . |
| 2 | *The source* $_{FS}$ , interviewed by Interfax in Grozny , **expressed confidence** that that the command of the Russian forces in Chechnya would soon " be able to obtain documentary confirmation " that Khattab was dead . |
| 3 | *The Commonwealth team earlier* this week $_{FS}$ **said** that " the conditions in Zimbabwe did not adequately allow the free and fair expression of will by the electorate ". |
| 4 | Publishing such biased reports will only create **mistrust** among nations $_{FS}$ regarding the objectives and independence of the UN Commission on Human Rights . |
| 5 | The Inkatha Freedom Party , Democratic Alliance , New National Party , African Christian Democratic Party , the Pan Africanist Congress and the United Christian Democratic Party $_{ZX}$ had disagreed with the ANC $_{FS}$ **conclusion** . |
| 6 | The Nigerian leader , President Olusegun Obasanjo $_{ZX}$ , had urged the *minister* $_{FS,ZX}$ not to attack Blair frontally over Britain 's negative position regarding Zimbabwe , but to deal [...] . |
| 7 | US diplomats $_{ZX}$ say *Bush* $_{FS,ZX}$ **will seek to support** Kim 's Nobel Prize winning policy by offering new talks with the North , while remaining firm about North Korea 's missile sales and its feared chemical and biological weapons programmes. |

Table 11: The dev examples for which the FS-MTL model *correctly* predicts the *holder* in 6/8 trials, whereas the Z&X-STL model predicts *incorrectly* in 6/8 trials.

| | |
|---|---|
| 1 | In most cases he **described** *the legal punishments* $_{FS}$ like floggings and executions of murderers and major drug traffickers that are applied based on the Shria , or Islamic law as human rights violations . |
| 2 | In another **verbal attack** Kharazi accused *the United States* $_{FS}$ of wanting to exercise " world dictatorship " since the " horrible attacks " of September 11 . |
| 3 | He said those who thought the election process would be rigged were supporters of the MDC party , adding that they **were prejudging** and wanted to direct the process $_{ZX}$ . |
| 4 | However , the fact that certain countries have a more balanced view of the conflict $_{ZX}$ is not the only reason to doubt that *anti-Israeli decisions* $_{FS}$ will , in fact , **be adopted** . |
| 5 | But *his tough stand on P'yongyang* $_{FS}$ **has provoked concern** in Seoul $_{ZX}$ , where President Kim Tae-chung , who is in the last year of his five-year term , has been trying to prise the hermit state out of isolation . |

Table 12: The dev examples for which the FS-MTL model *correctly* predicts the *target* in 6/8 trials, whereas the Z&X-STL model predicts *incorrectly* in 6/8 trials.

a nominal predicate and the holder is its object. The sentence is in passive voice but the models probably interpret it in the active voice and thus make the wrong prediction. In the third example, the opinion expression is the head of the relative clause that modifies the holder. These examples raise the following questions: would improved consistency with syntax lead to improvements for ORL and could we train a dependency parsing model with SRL and ORL to help the models handle syntactically harder cases?

Example 4 shows that holders specific to the MPQA annotation schema are hard to label as they require inference skills: from *the department said*, we can defeasibly infer that it is *the department* who *expects [this cost] to stand at just $400/year [...]*. To handle such cases, it would be worth trying training our models jointly with models for recognizing textual entailment.

Examples 6–7 illustrate that some gap in performance stems from difficulties in processing MPQA. Example 5 has no gold holder, but the models make plausible predictions. For example 6, FS-MTL predicts the discontinuous holder *they ... all*, while MPQA allows only contiguous entities. Therefor our evaluation scripts interpret *they* and *all* as two separate holders and deem *all* as incorrect, resulting with lower precision. Finally, for example 7 our models make plausible predictions. However, the gold holder is always the entity from the coreference cluster that is the closest to the opinion.[8] The evaluation scripts needs to be extended such that predicting any entity from the coreference cluster is considered to be correct. To conclude, to better evaluate future developments, it would be worth curating MPQA instances with missing roles and extending evaluation to account for coreferent holders and discontinuous roles.

The examples in Table 10 demonstrate that dif-

---

[8]We followed the prior work (Katiyar and Cardie, 2016).

ficulties in labeling targets originate from similar reasons as for holders. Examples 1–3 demonstrate complex syntactic constructions, examples 4–6 MPQA-specific annotations that require inference and example 7 exemplifies a missing target.

**How does MTL help?** There are 18/1055 instances in the dev set for which the FS model predicts the holder correctly and the Z&X-STL model does not, and 19/1055 for targets.

For holders, for 9 out 18 of such examples, the Z&X-STL model does not predict anything (as in Examples 2–5 in Table 11). From Examples 1–5 we notice that SRL data helps to handle more complex syntactic constructions. From Examples 5–7 we observed that using MTL with SRL helps to handle cases when more than one person or organization is present in the close neighborhood of the opinion. For targets, for 11 out of 18 cases the Z&X-STL model does not predict anything as in Examples 1–2 in Table 11. We conclude that the greatest improvements from the FS-MTL model comes from having far fewer missing roles.

## 6 Related work

**FGOA.** Closest to our work are Yang and Cardie (2013) (Y&C) and Katiyar and Cardie (2016) (K&C). They as well label both holders and targets in MPQA. By contrast, our focus is on the task of ORL. We thus refrain from predicting opinion expressions first, to ensure a reproducible evaluation setup on a fixed set of gold opinion expressions. The MTL models we develop in this work will, however, be the basis for the full task in a later stage. Because of these differences, direct comparison to Y&C and K&C is not possible. However, if we compare our results we notice a big gap that demonstrates that opinion expression extraction is the import step in FGOA. Similar to K&C, Liu et al. (2015) jointly labels opinion expressions and their targets in reviews.

Some work focuses entirely on labeling of opinion expressions (Yang and Cardie, 2014; Irsoy and Cardie, 2014). Other work looks into specific subcategories of ORL: opinion role induction for verbal predicates (Wiegand and Ruppenhofer, 2015), categorization of opinion words into actor and speaker view (Wiegand et al., 2016b), opinion roles extraction on opinion compounds (Wiegand et al., 2016a). Wiegand and Ruppenhofer (2015) report 72.54 binary F1 score for labeling of holders in MPQA (results for targets are not reported).

**Neural SRL.** New neural SRL models have emerged (He et al., 2017; Yang and Mitchell, 2017; Marcheggiani and Titov, 2017) since we started this work. In future work we can improve our models with such new proposals.

**Auxiliary tasks for MTL.** Other work investigates under which conditions MTL is effective. Martínez Alonso and Plank (2017) show that the best auxiliary tasks have low kurtosis of labels (usually a small label set) and high entropy (labels occur uniformly). We show that the best MTL model for ORL is the model which uses shared layers only. Thus it seems reasonable to consider only a small and uniform SRL label set {A0, A1}.

Bingel and Søgaard (2017) show that MTL works when the main task has a flattening learning curve, but the auxiliary task curve is still steep. We notice such behavior in our learning curves.

## 7 Conclusions

We address the problem of scarcity of annotated training data for labeling of opinion holders and targets (ORL) using multi-task learning (MTL) with Semantic Role Labeling (SRL). We adapted a recently proposed neural SRL model for ORL and enhanced it with different MTL techniques. Two MTL models achieve significant improvements with all evaluation measures, for both holders and targets, on both dev and test set, when evaluated with repeated 4-fold CV. We recommend evaluation with comparable dev and test set sizes for future work, as this enables more reliable evaluation.

With deeper analysis we show that future developments should improve the ability of the models to capture long-range dependencies, investigate if consistency with syntax can improve ORL, and consider other auxiliary tasks such as dependency parsing or recognizing textual entailment. We emphasize that future improvements can be measured more reliably if the evaluation covers opinion expressions with missing roles and considers all mentions in opinion role coreference chains as well as discontinuous roles.

### Acknowledgments

# References

Joachim Bingel and Anders Søgaard. 2017. Identifying beneficial task relations for multi-task learning in deep neural networks. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Association for Computational Linguistics, Valencia, Spain, pages 164–169. http://www.aclweb.org/anthology/E17-2026.

Anders Björkelund, Bernd Bohnet, Love Hafdell, and Pierre Nugues. 2010. A high-performance syntactic and semantic dependency parser. In *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations*. Association for Computational Linguistics, pages 33–36.

Xavier Carreras and Lluís Màrquez. 2005. Introduction to the CoNLL-2005 Shared Task: Semantic Role Labeling. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*. Association for Computational Linguistics, Ann Arbor, Michigan, pages 152–164. http://www.aclweb.org/anthology/W/W05/W05-0620.

Eugene Charniak, Don Blaheta, Niyu Ge, Keith Hall, John Hale, and Mark Johnson. 2000. Bllip 1987-89 wsj corpus release 1. *Linguistic Data Consortium, Philadelphia* 36.

Xinchi Chen, Zhan Shi, Xipeng Qiu, and Xuanjing Huang. 2017. Adversarial multi-criteria learning for chinese word segmentation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vancouver, Canada, pages 1193–1203. http://aclweb.org/anthology/P17-1110.

Yejin Choi, Eric Breck, and Claire Cardie. 2006. Joint extraction of entities and relations for opinion recognition. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Sydney, Australia, pages 431–439. http://www.aclweb.org/anthology/W/W06/W06-1651.

Lingjia Deng, Yoonjung Choi, and Janyce Wiebe. 2013. Benefactive/malefactive event and writer attitude annotation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Sofia, Bulgaria, pages 120–125. http://www.aclweb.org/anthology/P13-2022.

Lingjia Deng and Janyce Wiebe. 2014. Sentiment propagation via implicature constraints. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Gothenburg, Sweden, pages 377–385. http://www.aclweb.org/anthology/E14-1040.

Yarin Gal and Zoubin J. C. Ghahramani. 2016. A theoretically grounded application of dropout in recurrent neural networks. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems (NIPS)*.

Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning (ICML)*. pages 1180–1189.

Tao Gui, Qi Zhang, Haoran Huang, Minlong Peng, and Xuanjing Huang. 2017. Part-of-speech tagging for twitter with adversarial neural networks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, pages 2401–2410. https://www.aclweb.org/anthology/D17-1255.

Kazuma Hashimoto, Caiming Xiong, Yoshimasa Tsuruoka, and Richard Socher. 2017. A joint many-task model: Growing a neural network for multiple nlp tasks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, pages 446–456. https://www.aclweb.org/anthology/D17-1046.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*. pages 1026–1034.

Luheng He, Kenton Lee, Mike Lewis, and Luke Zettlemoyer. 2017. Deep semantic role labeling: What works and what's next. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vancouver, Canada, pages 473–483. http://aclweb.org/anthology/P17-1044.

Mikael Henaff, Arthur Szlam, and Yann LeCun. 2016. Recurrent orthogonal networks and long-memory tasks. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*. pages 2034–2042.

Ozan Irsoy and Claire Cardie. 2014. Opinion mining with deep recurrent neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, pages 720–728. http://www.aclweb.org/anthology/D14-1080.

Richard Johansson and Alessandro Moschitti. 2013. Relational Features in Fine-Grained Opinion Analysis. *Computational Linguistics* 39:473–509.

Shafiq Joty, Preslav Nakov, Lluís Màrquez, and Israa Jaradat. 2017. Cross-language Learning with

Adversarial Neural Networks. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*. Association for Computational Linguistics, Vancouver, Canada, pages 226–237. http://aclweb.org/anthology/K17-1024.

Rafal Jozefowicz, Wojciech Zaremba, and Ilya Sutskever. 2015. An empirical exploration of recurrent network architectures. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*. pages 2342–2350.

Arzoo Katiyar and Claire Cardie. 2016. Investigating LSTMs for Joint Extraction of Opinion Entities and Relations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany, pages 919–929. http://www.aclweb.org/anthology/P16-1087.

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*. Sydney, Australia, pages 1–8. http://www.aclweb.org/anthology/W/W06/W06-0301.

Young-Bum Kim, Karl Stratos, and Dongchan Kim. 2017. Adversarial adaptation of synthetic or stale data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vancouver, Canada, pages 1297–1307. http://aclweb.org/anthology/P17-1119.

Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*. San Diego.

Jiwei Li, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, and Dan Jurafsky. 2017. Adversarial learning for neural dialogue generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, pages 2147–2159. https://www.aclweb.org/anthology/D17-1229.

Pengfei Liu, Shafiq Joty, and Helen Meng. 2015. Fine-grained opinion mining with recurrent neural networks and word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, pages 1433–1443. http://aclweb.org/anthology/D15-1168.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2017. Adversarial multi-task learning for text classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguis-

tics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vancouver, Canada, pages 1–10. http://aclweb.org/anthology/P17-1001.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*. pages 55–60. http://www.aclweb.org/anthology/P/P14/P14-5010.

Diego Marcheggiani and Ivan Titov. 2017. Encoding sentences with graph convolutional networks for semantic role labeling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, pages 1506–1515. https://www.aclweb.org/anthology/D17-1159.

Héctor Martínez Alonso and Barbara Plank. 2017. When is multitask learning effective? semantic sequence prediction under varying data conditions. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Association for Computational Linguistics, Valencia, Spain, pages 44–53. http://www.aclweb.org/anthology/E17-1005.

Frank J. Massey Jr. 1951. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American statistical Association* 46(253):68–78.

Martha Palmer, Paul Kingsbury, and Daniel Gildea. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics* 31:71–106.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. 2013. On the difficulty of training recurrent neural networks. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*. pages 1310–1318.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, pages 1532–1543. http://www.aclweb.org/anthology/D14-1162.

Lianhui Qin, Zhisong Zhang, Hai Zhao, Zhiting Hu, and Eric Xing. 2017. Adversarial connective-exploiting networks for implicit discourse relation classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vancouver, Canada, pages 1006–1017. http://aclweb.org/anthology/P17-1093.

Nils Reimers and Iryna Gurevych. 2017. Optimal hyperparameters for deep lstm-networks for sequence labeling tasks. *arXiv preprint arXiv:1707.06799* .

Josef Ruppenhofer and Jasper Brandes. 2016. Effect functors for opinion inference. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*.

Josef Ruppenhofer, Swapna Somasundaran, and Janyce Wiebe. 2008. Finding the Sources and Targets of Subjective Expressions. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC)*. Marrakech, Morocco, pages 2781–2788.

Anders Søgaard and Yoav Goldberg. 2016. Deep multi-task learning with low level tasks supervised at lower layers. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 231–235. http://anthology.aclweb.org/P16-2038.

Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *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. *Language Resources and Evaluation* 39:165–210.

Michael Wiegand, Christine Bocionek, and Josef Ruppenhofer. 2016a. Opinion holder and target extraction on opinion compounds – a linguistic approach. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, San Diego, California, pages 800–810. http://www.aclweb.org/anthology/N16-1094.

Michael Wiegand and Josef Ruppenhofer. 2015. Opinion holder and target extraction based on the induction of verbal categories. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*. Association for Computational Linguistics, Beijing, China, pages 215–225. http://www.aclweb.org/anthology/K15-1022.

Michael Wiegand, Marc Schulder, and Josef Ruppenhofer. 2016b. Separating actor-view from speaker-view opinion expressions using linguistic features. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, San Diego, California, pages 778–788. http://www.aclweb.org/anthology/N16-1092.

Theresa Ann Wilson. 2008. *Fine-Grained Subjectivity And Sentiment Analysis: Recognizing The Intensity, Polarity, And Attitudes Of Private States*. Ph.D. thesis, University of Pittsburgh.

Yi Wu, David Bamman, and Stuart Russell. 2017. Adversarial training for relation extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, pages 1779–1784. https://www.aclweb.org/anthology/D17-1187.

Bishan Yang and Claire Cardie. 2013. Joint Inference for Fine-grained Opinion Extraction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Sofia, Bulgaria, pages 1640–1649. http://www.aclweb.org/anthology/P13-1161.

Bishan Yang and Claire Cardie. 2014. Joint modeling of opinion expression extraction and attribute classification. *Transactions of the Association for Computational Linguistics* 2:505–516.

Bishan Yang and Tom Mitchell. 2017. A joint sequential and relational model for frame-semantic parsing. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark, pages 1258–1267. https://www.aclweb.org/anthology/D17-1129.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vancouver, Canada, pages 1959–1970. http://aclweb.org/anthology/P17-1179.

Jie Zhou and Wei Xu. 2015. End-to-end learning of semantic role labeling using recurrent neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Beijing, China, pages 1127–1137. http://www.aclweb.org/anthology/P15-1109.

## A  MPQA Pre-processing

MPQA is challenging not only because it captures a variety of phenomena as we have illustrated in the Introduction, but as well because it is hard to process it in such a way that it can be presented to a neural sequence labeling model. Code or sufficient description how the corpus was processed is not available from the prior work.

The first difficulty is that we are designing a model that labels at the token-level, but annotation spans are given in bytes. Thus, we used Stanford

|  | # DSEs (incl. ignored) | # implicit DSEs (ignored) | # inferred DSEs (ignored) | # filtered DSEs | # some uncrt. filt. DSEs |
|---|---|---|---|---|---|
| TRAIN (avg) | 3723.5 | 481 | 101.25 | 3141.25 | 133 |
| TEST (avg) | 1229.5 | 159 | 33.75 | 1036.75 | 44 |
| DEV | 1263 | 168 | 40 | 1055 | 43 |

|  | # very uncrt. filt. DSEs | # no Hs filt. DSEs | # no roles filt. DSEs | # no Ts filt. DSEs | # insubs. filt. DSEs |
|---|---|---|---|---|---|
| TRAIN (avg) | 40.5 | 171.25 | 66.75 | 413.75 | 528.75 |
| TEST (avg) | 13.5 | 56.75 | 22.25 | 136.25 | 174.25 |
| DEV | 15 | 56 | 22 | 146 | 180 |

|  | # Hs of filt. DSEs | # Ts of filt. DSEs | # some uncrt. Hs | # some uncrt. Ts | # overlap. entites |
|---|---|---|---|---|---|
| TRAIN (avg) | 2903.25 | 19528.5 | 17.25 | 27.75 | 961 |
| TEST (avg) | 957.75 | 6424.5 | 5.75 | 9.25 | 318 |
| DEV | 977 | 6073 | 5 | 6 | 305 |

|  | sentiment neg | sentiment pos | arguing pos | other attitude | intention pos |
|---|---|---|---|---|---|
| TRAIN (avg) | 946 | 817.5 | 438.25 | 381 | 238.75 |
| TEST (avg) | 314 | 270.5 | 143.75 | 126 | 79.25 |
| DEV | 299 | 300 | 131 | 126 | 66 |

|  | arguing neg | agree pos | speculation | agree neg | intention neg |
|---|---|---|---|---|---|
| TRAIN (avg) | 110.5 | 99.25 | 64.5 | 68.25 | 19.5 |
| TEST (avg) | 35.5 | 32.75 | 20.5 | 22.75 | 6.5 |
| DEV | 48 | 40 | 25 | 31 | 5 |

Table 13: Statistics of the ORL (MPQA) data for 4-fold CV.

CoreNLP (Manning et al., 2014) which tokenizes text and gives the byte span of every token.[9] However, due to to the absence of punctuation for transcripts of spoken conversations the sentence splitter treats a whole document as if it were one sentence. Therefore, for sentences longer than 150 tokens, we take 15 tokens preceding the opinion expression, the expression itself and 15 tokens after as proxy for a sentence that we present to the model.

Opinion expressions we are interested in are annotated in MPQA as *direct subjectives* (DSEs). We discard implicit DSEs which frequently point to the attitude which covers the whole sentence and reflects the attitude of the author of the document as in example (6). These DSEs are not useful for the task we are looking into. Although such DSEs should be marked with the `implicit` attribute, sometimes they are not. Some of such cases we capture by demanding that a DSE is longer than one byte and that the author is not the only holder. There are few DSEs for which byte spans did not match with any sentence, and we discard those as well.

(6) But there can not be any real [**talk**]$_O$ of success until the broad strategy against terrorism begins to bear fruit.

For every document, we collected from the corresponding annotation file: identifiers and byte spans of all holders marked with GATE_agent ($\mathcal{H}$), attitudes marked with GATE_attitude, and targets marked with GATE_target. Holders and targets can be marked multiple times with the same id, but with different byte spans. If the `nested-source` attribute of a DSE or the `target-link` attribute of its attitude point to identifiers of such holders and targets, we pick the byte spans which are closest to the DSE. In many cases the `nested-source` attribute of a DSE pointed to a holder which is not marked in the annotation file ($\notin \mathcal{H}$). We tried to fix the `nested-source` attribute by doing the following transformations: (1) adding 'w' to the beginning (e.g. nhs $\mapsto$ w, nhs), (2) removing 'w' from the beginning (e.g. w, ip $\mapsto$ ip), (3) removing duplicates (e.g. w, mug, mug $\mapsto$ w, mug). Although these transformations helped a lot, they are a few holders and targets we could not trace.

In some cases, as in example (7), an opinion expression and its opinion roles overlap. In average, we discard 74.7 such holders and 16.2 targets, because we train the output CRF to predict only one label by token. Notice that the prior work (Katiyar and Cardie, 2016) had to do the same.

(7) Mugabe said [Zimbabwe]$_T$ needed their continued support against what he called [**hostile [international]$_H$ attention**]$_O$.

We discard inferred attitudes, as labeling of their targets is considered to be another task (Deng et al., 2013; Deng and Wiebe, 2014; Ruppenhofer

| | # DSEs (incl. ignored) | # implicit DSEs (ignored) | # inferred DSEs (ignored) | # filtered DSEs | # some uncrt. filt. DSEs |
|---|---|---|---|---|---|
| TRAIN (avg) | 4173.3 | 537.3 | 119.7 | 3516.3 | 137.7 |
| TEST (avg) | 457.8 | 43.9 | 29.9 | 349.3 | 15.2 |
| DEV | 1579 | 211 | 42 | 1326 | 67 |

| | # very uncrt. filt. DSEs | # no Hs filt. DSEs | # no roles filt. DSEs | # no Ts filt. DSEs | # insubs. filt. DSEs |
|---|---|---|---|---|---|
| TRAIN (avg) | 47.7 | 187.2 | 77.4 | 459.9 | 567.9 |
| TEST (avg) | 7.3 | 19.3 | 11.8 | 82.6 | 150.5 |
| DEV | 16 | 76 | 25 | 185 | 252 |

| | # Hs of filt. DSEs | # Ts of filt. DSEs | # some uncrt. Hs | # some uncrt. Ts | # overlap. entites |
|---|---|---|---|---|---|
| TRAIN (avg) | 3251.7 | 21664.8 | 17.1 | 27.9 | 1064.7 |
| TEST (avg) | 957.4 | 1700 | 19.4 | 37.8 | 84.9 |
| DEV | 1225 | 7978 | 9 | 12 | 401 |

| | sentiment neg | sentiment pos | arguing pos | other attitude | intention pos |
|---|---|---|---|---|---|
| TRAIN (avg) | 1008.9 | 949.5 | 471.6 | 440.1 | 266.4 |
| TEST (avg) | 107.8 | 89.4 | 50.7 | 40.2 | 25.6 |
| DEV | 438 | 333 | 189 | 144 | 88 |

| | arguing neg | agree pos | speculation | agree neg | intention neg |
|---|---|---|---|---|---|
| TRAIN (avg) | 133.2 | 115.2 | 80.1 | 74.7 | 16.2 |
| TEST (avg) | 14.1 | 11.1 | 8.6 | 6.5 | 1.428571429 |
| DEV | 46 | 44 | 21 | 39 | 13 |

Table 14: Statistics of the ORL (MPQA) data for 10-fold CV.

and Brandes, 2016).

Further, a DSE can have multiple attitudes and each attitude can point to different targets. Again, because the model can predict only one label by token, we have to pick one attitude and non-overlapping targets. We chose attitudes according to the following priorities: sentiment, intention, agreement, arguing, other-attitude, speculation.

We kept DSEs with the `insubstantial` attribute which are either not significant (8) or not not real within the discourse (9). Our models should demonstrate the ability of properly labeling roles of insubstantial DSEs. However, note that when FGOA is used for opinion-oriented summarization or QA, opinion roles of insubstantial opinions should not be labeled. A full FGOA system should additionally predict whether an opinion is substantial within the discourse, before labeling its opinion roles.

(8) [...] it completely supports the [U.S.]$_H$ [**stance**]$_O$ [...].

(9) [...] Antonio Martino, meanwhile, said [...] that his country would not support an attack on Iraq without "proven proof" that [Baghdad]$_H$ is [**supporting**]$_O$ [al Qaeda]$_T$.

Finally, DSE, holder and target annotations allow an attribute that indicates whether an annotator was uncertain with possible values: somewhat-and very-uncertain. We did not discard those be-lieving that they would have been discarded by the corpus creators if they are really incorrect.

For reproducibility we report detailed data statistics in Tables 13 and 14: average number (calculated over folds) of all extracted DSEs, implicit DSEs, inferred DSEs, DSEs used in experiments (not implicit or inferred), somewhat uncertain DSEs used in experiments, very uncertain DSEs used in experiments, insubstantial DSEs used in experiments, the average number (calculated over folds) of DSEs used in experiments without a holder, without a target, without the `attitude-link` attribute, without both roles, the average number (calculated over folds) of holders, somewhat uncertain holders, very uncertain holders, targets, somewhat uncertain targets and very uncertain targets, the average number (calculated over folds) of different attitude types used in the experiments.

Examples how to easily use our MPQA pre-processing scripts can be found at `https://github.com/amarasovic/naacl-mpqa-srl4orl/blob/master/mpqa2-pytools.ipynb`.

## B Training details

The code for training and evaluating our models can be found at `https://github.com/amarasovic/naacl-mpqa-srl4orl`.

**Input representation.** We used 100d GloVe word embeddings (Pennington et al., 2014) pre-

trained on Gigaword and Wikipedia and did not fine-tune them. For MTL models vocabulary was built from all the words in the training data of both tasks, and OOV words were replaced with an UNK token. The embedding of the context of a predicate or an opinion is the average of the embeddings of the predicate or the opinion phrase, of 2 preceding words and 2 words after.

**Weights initialization.** The size of all LSTM hidden states was set to 100. The number of the backward and the forward LSTM layers is set to 3, which counts for 6 LSTM layers in Z&X. Z&X achieved circa 2% higher SRL F1 score with 8 LSTM layers, but such a deep model would cause overfitting on the small-sized ORL data. In the H-MTL model, SRL is supervised at the 2nd LSTM layer. We initialized the LSTM weights with random orthogonal matrices (Henaff et al., 2016), all other weight matrices with the *He initialization* (He et al., 2015). LSTM forget biases were initialized with 1s (Jozefowicz et al., 2015), all other biases with 0s.

**Optimization.** We trained our model in mini-batches of size 32 using Adam (Kingma and Ba, 2015) with the learning rate of $10^{-3}$. For MTL we alternate batches from different tasks. We clip gradients by global norm (Pascanu et al., 2013), with a clipping value set to 1. Single-task models were trained for 10K iterations and MTL models for 20K. One epoch counts for $\left\lceil \frac{\text{train size}}{\text{batch size}} \right\rceil$ iterations. We stop training if the arithmetic mean of proportional F1 scores of holders and targets is not improved in 25 epochs. For the minmax optimization we use a gradient reversal layer (Ganin and Lempitsky, 2015). The discriminator's cross-entropy loss is scaled with 0.1.

**Regularization.** Variational dropout (Gal and Ghahramani, 2016) with a keep probability $k_p \in 0.85$ was applied to the outputs and the recurrent connections of the LSTMs. Standard dropout (Srivastava et al., 2014) was applied to the output classifier weights with a keep probability $k_p \in 0.85$ and to the input embeddings with $k_p \in 0.7$.