Supplement to Paper:
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

1 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 CoreNLP (Manning et al., 2014) which tokenizes text and gives the byte span of every token.[1] 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.

(1) 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 ↦ w, nhs), (2) removing 'w' from the beginning (e.g. w, ip ↦ ip), (3) removing duplicates (e.g. w, mug, mug ↦ 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.

(2) 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 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

[1] We used python wrapper: https://github.com/brendano/stanford_corenlp_pywrapper

	# 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 1: Statistics of the ORL (MPQA) data for 4-fold CV.

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.

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

(4) [...] 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 believing that they would have been discarded by the corpus creators if they are really incorrect.

For reproducibility we report detailed data statistics in Tables 1 and 2: 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`.

2 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

	# 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 2: Statistics of the ORL (MPQA) data for 10-fold CV.

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 $\lceil \frac{\text{train size}}{\text{batch size}} \rceil$ 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 $k_p \in 0.7$.

References

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.

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.

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.

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.

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

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.

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.

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

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.