Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022)

Improved Opinion Role Labelling in Parliamentary Debates

Laura Bamberg University of Mannheim
Ines Rehbein University of Mannheim
Simone Paolo Ponzetto University of Mannheim
{rehbein,ponzetto}@uni-mannheim.de

Abstract

This paper presents a model for German Opinion Role Labelling (ORL), using the data from the IGGSA-STEPS 2014 and 2016 shared tasks. We frame the problem as a token classification task and employ a simple transformer-based model that achieves new state-of-the-art results on the data. Then we investigate whether we can further improve our model by transferring knowledge from a related task, i.e., Semantic Role Labelling. Our results show that, despite the small size of our data, this transfer learning step yields further improvements for ORL, mostly regarding recall for target prediction. Finally, we present an error analysis, showing where knowledge transfer from SRL can help and what is still difficult for German ORL.

1 Introduction

The extraction of subjective expressions together with their opinion holders and targets is not only an important processing step for the analysis of argumentation mining but is also relevant for political text analysis. For English, the seminal work of Stoyanov et al. (2004) and Wiebe et al. (2005) has provided resources for training and evaluation of opinion mining models for newswire. However, resources for other languages, domains and text types are still scarce.

Previous work on German has focussed on the political domain where Ruppenhofer et al. (2014, 2016) have presented a corpus of Swiss-German parliamentary debates annotated with subjective expressions, their opinion holders (or sources) and targets (Figure 1). The data set has been used in two shared tasks.[1] However, compared to the MPQA 2.0 corpus (Wiebe et al., 2005; Wilson, 2008) which includes more than 8,500 sentences,

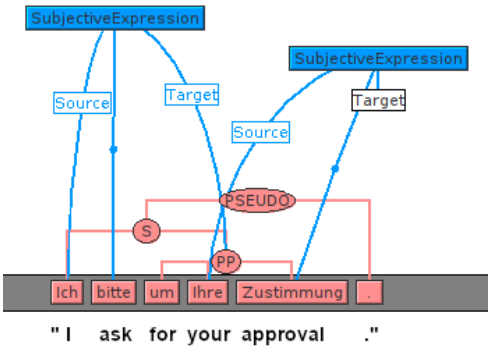

Figure 1: Screenshot of example annotations from the IGGSA-STEPS shared task data for the verb "bitten" (*ask*) and the noun "Zustimmung" (*approval*), visualised in Salto (Burchardt et al., 2006a).

the data is rather small with less than 1,200 sentences. This is reflected in the low results for opinon holder and target extraction, where scores for the best systems from the 2016 shared task were in the range of 46% F1 (micro) for holders and 40% F1 for targets. Follow-up work by Wiegand et al. (2019a) has improved the extraction of opinion holders by around 4 percentage points but failed to increase results for target extraction. The low results imply that, at this stage, the models are not yet good enough to be used in downstream applications.

Since then, transformer-based models (Vaswani et al., 2017; Devlin et al., 2019) and transfer learning approaches have brought huge improvements to the field of Natural Language Understanding (NLU) and are particularly well suited for task settings where only small data are available. Therefore, in our work we exploit the expressive power of transformers and transfer learning and present a simple transformer-based system for German opinion holder and target extraction.

As expected, our baseline system already beats previous work by far, yielding improvements in the range of 10-15 percentage points. We then explore

[1] See the IGGSA-STEPS 2014 shared task: https://sites.google.com/site/iggsasharedtask/task-1 and for 2016: https://iggsasharedtask2016.github.io.

whether we can further improve results by transferring knowledge from a related task, i.e., Semantic Role Labelling (SRL). Transfer from SRL to ORL has been successful for improving results for English Opinion Role Labelling (ORL) (Marasovic and Frank, 2018). However, it is unclear whether a similar approach will work for German where the size of the training data is only a fraction of the English ORL data. To answer this question, we exploit a German newspaper corpus with frame-semantic annotations (Burchardt et al., 2006b) and introduce an intermediate training step where we fine-tune our model on the SRL data, showing that this intermediate training step can further improve results, mostly in terms of recall.

The contributions of this work are as follows. We present a neural system for German opinion holder and target extraction, based on transformer-based transfer learning, and report new state-of-the-art results. We replicate previous results obtained for English, using SRL data for transfer learning, and show that this approach also works when substantially less data is available. Our final system outperforms previous best results by more than 15 percentage points.[2]

2 Related Work

Opinion mining, the "computational study of opinions, sentiments, and emotions expressed in text" (Liu, 2010), has become a vivid field of research in the last 20 years. Among the main goals of opinion mining is the extraction of the source or opinion holder (the one who has the opinion) and its topic or target (what the opinion is about).

Opinion Role Labelling (ORL) for English Most work on ORL has been conducted for English. Initially, the task has been modelled in a pipeline approach where the models first identify the opinion (or subjective expression) and then, given the opinion, in a second step predict the roles of *opinion holder* and *target*. There is, of course, a close link to semantic role labelling, and many works have exploited that link.

Kim and Hovy (2006), for example, have augmented the frame-semantic annotations in FrameNet (Baker et al., 1998) with opinion holder and target roles and used clustering techniques to predict semantic frames for subjective expressions not known by FrameNet. They then decompose the task into three phases where they first identify all opinion-bearing predicates in a sentence, then use SRL to label the semantic roles for the predicate and, finally, identify the holder and topic of the opinion-bearing expression among the labeled semantic roles.

Other work has tried to jointly learn the opinion-bearing expressions and their roles (Choi et al., 2006; Yang and Cardie, 2013; Katiyar and Cardie, 2016). The most recent one of those works, Katiyar and Cardie (2016), use deep bidirectional LSTMs to jointly extract opinion expressions and their holders and targets. The neural model does not outperform previous work that uses CRFs in combination with Integer Linear Programming (ILP) (Yang and Cardie, 2013). However, one advantage of the neural approach is that, unlike other work (Kim and Hovy, 2006; Johansson and Moschitti, 2013; Yang and Cardie, 2013; Wiegand and Ruppenhofer, 2015), it does not depend on external resources such as opinion lexicons, dependency parsers or SRL systems.

Marasovic and Frank (2018) present a neural approach, based on BiLSTMs and CRFs, that exploits external knowledge from SRL in a multi-task learning (MTL) setup. They focus on holder and target prediction and show that the MTL approach results in substantial improvements over a single-task baseline.

Quan et al. (2019) are the first to apply a transformer-based architecture (Vaswani et al., 2017; Devlin et al., 2019) for ORL. Their approach is similar to the one of Katiyar and Cardie (2016) and jointly learns the opinion expressions, their holders and targets. Their end-to-end model integrates BERT with a BiLSTM and CRF component and improves over a simple BiLSTM baseline. However, it fails to outperform the previous state-of-the-art of Katiyar and Cardie (2016) by far. The authors ascribe this to the limited size of the training data and the resource hunger of neural approaches. If that is true, then we cannot expect improvements for German where the size of the training data is even smaller than for English ORL and SRL. We thus want to explore whether is is possible to transfer knowledge from SRL to ORL for German in a low(er)-resource setting.

Our work is similar to Marasovic and Frank (2018) in that we also use Semantic Role Labelling data to address the problem of data sparsity for Opinion Role Labelling, which is much more se-

[2]Our models are available for download from https://github.com/umanlp/ORLde.

DE	Die Kantone	können,	wenn sie wollen,	also	eine Regelung treffen.	dummy-token
EN	The cantons	can,	if they wish,	therefore	make a regulation.	
TRANS	"The cantons can therefore, if they wish, make a regulation."					
instance 1	*Die Kantone*	können,	wenn sie wollen,	*also eine Regelung treffen .*		**inferred**
instance 2	Die Kantone	können,	wenn **sie** wollen	*also eine Regelung treffen .*		–
instance 3	**Die Kantone**	können,	wenn sie wollen	also eine Regelung treffen .		–

Table 1: Three example subjective expressions (underlined) within the same sentence, with their opinion holders (red) and targets (blue); example taken from the IGGSA-STEPS 2016 shared task test set.

vere for German than for English. We do not use a multi-task learning setup, as the size of the SRL data is around 8 times as large as the ORL data and we expect this imbalance to be a challenge for the MTL approach. Instead, we apply transfer learning through intermediate training where we first fine-tune a pretrained BERT model on the SRL data and then use the learned model to initialise the weights for our final ORL model that we fine-tune on the downstream task, i.e., Opinion Role Labelling.

ORL for German Most work on Opinion Role Labelling for German has been conducted in the context of two shared tasks, the IGGSA-STEPS 2014 and 2016 Shared Task on Source and Target Extraction from Political Speeches (Ruppenhofer et al., 2014, 2016). The data for the shared task includes debates from the Swiss parliament, annotated with subjective expressions, their opinon holders and targets. The data set is fairly small with 605 sentences for training and 581 sentences for testing. The number of annotated instances in the data, however, is substantially higher and amounts to 1,115 subjective expressions, 997 opinion holders (excluding *inferred opinion holders*, see §3.1 below) and 1,608 targets for training (see Table 2).

As reported in Wiegand et al. (2019b), 845 (850) subjective expression frames in the training (test) data include both, holder *and* target, while 152 (214) subjective expressions include only the holder. More frequent are subjective frames that include only the target, with a count of 763 (920). Subjective frames with neither holder nor target amount to 468 (433) in the training (test) set.

This is a typical low-resource scenario, and we thus want to investigate whether (and by how much) we are able to improve results over previous work that employs linguistic features, information from external knowedge bases and linguistic modelling. Our work addresses the following research questions:

RQ1: Can transformer-based transfer learning improve results for German ORL over previous best work, despite the small size of the training data?

RQ2: Can we replicate previous work on English and further improve results by harvesting information from German SRL?

We address RQ1 by fine-tuning a pretrained transformer-based language model on the ORL task and compare results to previous work on the same data. To answer our second RQ, we use the German SRL data from the CoNLL 2009 shared task "Syntactic and Semantic Dependencies in Multiple Languages" (Hajič et al., 2009) for transfer learning and investigate whether we will find similar improvements as have been reported for English.

3 A BERT model for German ORL

3.1 Task description and data

The task of opinion role labelling consists in identifying all opinion holders and targets for a given subjective expression. For illustration, see the example in Table 1 where three subjective expressions are given (können (*can*), wollen (*want*), Regelung treffen (*make regulation*)). The task then is to predict the opinion holder and target for each of these expressions.

In the first instance extracted from the example, only the target is expressed overtly while the opinion holder of können (*can*) has to be inferred as the speaker of the utterance. Those *inferred holders* are quite frequent and amount to 26% of all holders in the data (Wiegand et al., 2019b). In the

	#sent	SE (toks)	SE (types)	Holder	Target
train	605	2,105	1,115	997	1,608
test	581	2,166	1,110	1,064	1,770
Total	1,186	4,271		2,061	3,378

Table 2: Some statistics on the IGGSA shared task data.

second instance where the subjective expression is wollen (*want*), both holder and target are realised as arguments of the subjective predicate. Finally, the subjective expression Regelung treffen (*make regulation*) in the third instance is a support verb construction with an explicitly stated holder but the target role remains unfilled.

As in Marasovic and Frank (2018), we assume that the subjective expressions are given and focus on the ORL task. Given an input sentence, the task then consists in detecting the respective token spans for holder and target and assigning the correct label to each role.

Preprocessing We preprocess the data so that we extract one training (or test) instance for each subjective expression and its opinon roles, i.e., its opinion holder and target (including *inferred holders*). Please note that not each sentence includes a subjective expression (SE), and not every SE has an opinion holder and target.

Experimental setup In our first set of experiments, we train an ORL classifier for German, using the data from the IGGSA-STEPS 2014 and 2016 shared tasks (Ruppenhofer et al., 2014, 2016). To make our results comparable, we follow the setup of the 2016 shared task setup, using the data from the 2014 shared task for training and development (605 sentences) and evaluate our models on the same test portion used in the 2016 shared task, including 581 sentences. Table 2 shows some statistics for the data.

We model the task as a token classification task and use the BIO schema to distinguish the first token of each span from the tokens inside a span. We use the "O" label for all tokens that are not part of either holder or target. In the shared task data, the inferred holders are annotated by means of a flag and have to be predicted. We follow Wiegand et al. (2019a) and add a dummy token at the end of each instance which is assigned the label "Inferred" for all instances with implicit opinion holders. For instances with explicitly expressed holders and those without a holder, the dummy token is assigned the label "O" instead.

3.2 Baseline model

Our baseline model for ORL uses a simple token classification setup, similar to the argument detection and labelling step in the BERT-based SRL model of Shi and Lin (2019). There are, however, two differences between their model and ours. The first one concerns the model architecture, the second the representation of the input. The model of Shi and Lin (2019) integrates a BiLSTM layer on top of the BERT encoder, followed by a Multi-Layer Perceptron (MLP). To encode the information about the predicate (for SRL) or subjective expression (for ORL), they concatenate the classification [CLS] token, the input sentence, a separator token [SEP] and the predicate and input the whole sequence into the BERT encoder.

Instead of concatenating the input sentence and the predicate (or subjective expression), we use BERT's token-type-ids to encode this information. Specifically, we set the token type ids of all tokens that are part of the subjective expression to 1 and all other token ids to 0. Our model does not use an additional BiLSTM on top of BERT but, following the NER model presented in Devlin et al. (2019), inputs the encoded sequence directly into the MLP layer.

	ORL	SRL
optimizer	AdamW	AdamW
learning rate	2.693154582157772e-05	0.00003808
batch size	16	8
weight decay	0.019840937077311938	0.055
epsilon	5.45374378277376e-07	0.000001194

Table 3: Hyperparameters used for the ORL/SRL tasks.

Training details We implement our models with the huggingface transformers library (Wolf et al., 2020) and pytorch (Paszke et al., 2017) and do hyperparameter tuning with Weights & Biases (Biewald, 2020). We limit the input sequence length to 120 subword tokens and train in batches of 16 instances, using the AdamW optimizer with random search to determinine the optimal learning rate α, weight decay and epsilon ϵ (sampled from a uniform distribution with $min = 0.02$ and $max = 0.00001$ for α, $min = 0$ and $max = 0.1$ for weight decay and $min = 5e-9$ and $max = 0.000002$ for ϵ), with the objective to minimize the training loss.

Then we use the same tuned (hyper)parameters to train three independent versions of our model with different initialisations, each for 25 epochs. We select the best performing model on the development set and report results for each indiviual run and averaged results and standard deviation over all three runs.[3] Table 3 shows the (hyper)parameter

[3]Given that standard deviation between the different initialisations was quite low (see Table 4), we decided to report

System		*Holder*			*Target*		
		Prec	**Rec**	**F1**	**Prec**	**Rec**	**F1**
UDS-supervised		59.4	38.3	46.6	42.6	31.7	36.3
UDS-rulebased		59.9	28.6	38.7	**69.2**	28.9	40.8
WCR19		58.0	44.0	50.3	48.1	35.0	40.5
ORL-ST	avg.	**67.8**$_{\pm 0.6}$	**63.5**$_{\pm 0.3}$	**65.6**$_{\pm 0.2}$	54.2$_{\pm 0.7}$	**53.2**$_{\pm 0.3}$	**53.9**$_{\pm 0.2}$

Table 4: Results for ORL on the STEPS-2016 test set (UDS-sup: supervised UDS system, UDS-rule: rule-based UDS system; WCR19: Wiegand et al. (2019a); ORL-ST: BERT-based single-task ORL system; results averaged over 3 runs; stdev reports standard deviation over 3 runs.).

settings for our experiments.

3.3 Baseline results

We now report results for our BERT single-task model, ORL-ST, and compare them to previous work (Table 4). For evaluation, we use the scorer from the IGGSA-STEPS shared tasks, kindly provided by the organisers, to ensure the comparability of the results.[4] We report the strict measure for (micro) precision, recall and F1 for opinion holders and targets that only considers a predicted holder or target as correct if *all tokens that belong to this entity have been predicted correctly*. Please note that the results for opinion holders also include predictions for inferred holders (see Table 1, instance 1).

We compare against the University of Saarland (UDS) contributions from the IGGSA-STEPS 2016 shared task (UDS-supervised and UDS-rulebased) (Wiegand et al., 2016) and the supervised feature-based approach of Wiegand et al. (2019a). The authors refer to the moderate results reported for deep learning approaches for ORL (Katiyar and Cardie, 2016) as motivation for not using deep learning in their work, and highlight the importance of linguistic information and, in particular, syntactic dependency relations for resolving opinion holders and targets. Finally, the small size of the German data questions the benefits to be expected from neural approaches, which is why Wiegand et al. (2019a) decided to employ SVMs in their work.

Table 4 shows that the baseline BERT model outperforms previous work by a large margin, with improvements in the range of 15-22% for opinion holders and 13-17% for the identification of targets. The rule-based approach (UDS-rulebased), however, beats the BERT system wrt. precision, but at the cost of a very low recall. For all other models, results increase for both, precision and recall.

This answers our first research question, **RQ1**: Transfer learning approaches are well suited to increase results for German ORL over previous feature-based approaches even in low-resource scenarios.

4 SRL for German ORL

We now turn to our second research question and investigate whether it is possible to further improve results for German ORL by means of an additional knowledge transfer from the semantic role labelling (SRL) task. As training data for SRL, we use the German part of the CoNLL 2009 shared task data (Hajič et al., 2009) and train a BERT-based classifier, using the same model architecture and setup as for the ORL task. The data comes originally from the SALSA corpus (Burchardt et al., 2006b), a corpus of newspaper text from a German daily newspaper (*Frankfurter Rundschau*). SALSA includes verbal predicates and their frame elements, with annotations in the flavor of Berkeley FrameNet (Baker et al., 1998). The semantic frames and roles have been automatically converted from FrameNet-style annotations to PropBank (Palmer et al., 2005) style for the shared task.

The data we use for training includes over 36,000 sentences, out of which 14,282 sentences include at least one annotated predicate. The number of training instances (where sentences with more than one annotated predicate result in multiple instances, as described for the ORL preprocessing step) thus amounts to 17,400 instances. The development set includes 2,000 sentences and the test data 400 sentences.

Please note that our goal is not to optimize results for the SRL task but to use SRL as an auxiliary task to transfer knowledge about predicate argument structure to ORL. For this, we compare

results for 3 individual runs only.

[4]We would like to thank the shared task organisers for providing us with the scorer and system outputs from the IGGSA-STEPS shared task.

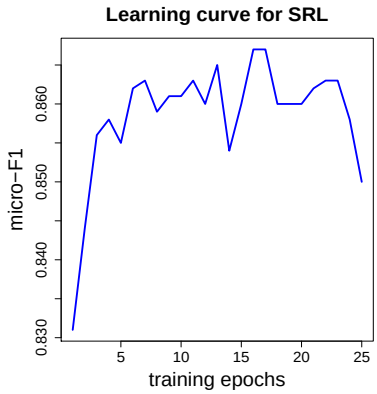

Figure 2: Learning curves for SRL over 25 epochs of training (micro-F1 on the SRL development set).

two different settings. In the first setting, we select the best performing model for SRL, based on the F1 scores on the development set, and use this model to initialise the BERT parameters for subsequent ORL fine-tuning. In the second setting, we do not fully train the model on the SRL data until convergence but stop the training process when the learning curve starts to flatten, which happens after the third training epoch (see Figure 2). Table 5 reports results on the SRL development set for both models (Exp. 1 and 2).

Training details We use this model to initialise the parameters of the ORL model that we then fine-tune on the downstream task (ORL). Model architecture and parameter settings are the same as described in Section 3.2 and Table 3. As before, we train 3 individual models with different initialisations for 25 epochs and select the best performing model for each run on the development set. We report results for each individual run and averaged results and standard deviation over all runs.

4.1 Results for transfer learning from SRL

Table 6 shows results for transfer learning from SRL to ORL. We notice that the intermediate training has a noticable effect on the downstream task. The SRL model that has been trained for 16 epochs and achieved best results on the SRL dev set (Figure 2) fails to further improve results for ORL. Using the parameters from the ORL-3 model that has been trained for 3 epochs only to initialise the BERT ORL model, however, results in another increase in results. This increase is rather small for target prediction with 0.7% but more pronounced for the prediction of opinion holders with 1.6%.

A possible explanation for the better performance of the undertrained SRL model as source

Exp.	Model	Prec	Rec	F1
SRL-1	best-on-dev	86.7	86.7	86.7
SRL-2	3-epochs	86.2	85.1	85.6

Table 5: Results for SRL with BERT (dev set).

of knowledge transfer is that the size of the ORL training data is only a fraction of the SRL data (605 sentences versus 14,282 sentences). Thus, the model has been fitted for a different task (SRL) and has not seen enough data to adapt to the new task (ORL). This suggests that other architectures might be more promising for a low-resource setting like this, such as adapter-based fine-tuning (Rebuffi et al., 2018; Houlsby et al., 2019; Bapna and Firat, 2019; Pfeiffer et al., 2020). We plan to explore this in future work.

As mentioned above, the results in Table 6 come from a strict evaluation where we only count roles as correct if *all* tokens that belong to that role have been identified correctly. This explains why results for targets are substantially lower than the ones for holders, given their average lengths (2.1 tokens for opinion holders vs. 5.5 tokens for targets). To add another perspective, we augment the results reported above by a token-based evaluation (Table 7) where we remove the prefixes from the BIO scheme and compute precision, recall and F1 on the token level. Table 8 illustrates the difference between the two evaluation measures, using a constructed example sentence.

For the *strict* evaluation in Table 8, we count one correctly identified role, i.e., the target. We also count one false positive, as we have predicted a span that does not exist in the gold standard. Additionally, we count one false negative because we failed to identify the correct holder (or source) span. For the *token-based* evaluation, on the other hand, we count 7 true positives (2 for the holder and 5 for the target) and one false negative for the missed token "auch" (*also*).

As expected, results for target prediction are much higher in the token-based evaluation setting in Table 8. While the general trends are the same as for the strict evaluation, with best results being obtained by the ORL-3 system (transfer from SRL to ORL), we note that the single-task model, ORL-1, outperforms the transfer model in terms of precision for all three roles (holder, target, inferred holder) while the transfer step mostly helps to increase recall (Table 7).

Exp.	Model	Run	Prec	Rec	F1	Prec	Rec	F1
			Holder			*Target*		
ORL-1	single-task best-on-dev	1	67.1	63.8	65.4	53.4	52.8	53.6
		2	68.2	63.3	65.7	54.5	53.2	53.9
		3	68.2	63.3	65.7	54.8	53.5	54.1
		avg	67.8	63.5	65.6	54.2	53.2	53.9
ORL-2	SRL-to-ORL best-on-dev	1	66.9	63.3	65.0	52.0	51.2	51.6
		2	66.4	65.3	65.8	52.9	52.4	52.6
		3	64.2	64.5	64.3	53.2	52.9	53.1
		avg	65.8	**64.4**	65.0	52.7	52.2	52.4
ORL-3	SRL-to-ORL 3 epochs	1	70.7	64.7	67.5	54.3	55.0	54.6
		2	71.6	63.8	67.5	54.0	53.7	53.8
		3	68.7	64.7	66.6	55.2	55.4	55.3
		avg	**70.3**	**64.4**	**67.2**	**54.5**	**54.7**	**54.6**

Table 6: Results for the single-task ORL baseline (ORL-1) and for the transfer learning experiments (ORL-2, ORL-3) with intermediate training on SRL (best-on-dev: model that gave best results on the development set; 3 epochs: model has been trained for 3 epochs only).

Exp.	Model	Run	Holder			Target			Speaker (inferred)		
			Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
ORL-1	single-task best-on-dev	1	76.1	50.7	60.9	73.5	74.5	74.0	71.2	79.3	75.0
		2	78.8	49.8	61.1	67.6	81.4	73.9	68.6	75.3	71.8
		3	71.0	56.2	62.7	70.9	80.6	75.4	70.5	73.0	71.8
		avg	**75.3**	52.2	61.6	**70.7**	78.8	74.4	**70.1**	75.9	72.9
ORL-2	SRL-to-ORL best-on-dev	1	72.0*	52.1	60.5*	67.3	82.5***	74.1	67.4	78.1	72.4
		2	68.2**	56.0*	61.5***	69.1	80.6	74.4	67.4	76.6	71.7
		3	70.3	54.7	61.5	67.2	81.7	73.7	61.1***	83.1***	70.4***
		avg	70.2	54.3	61.2	67.9	81.6	74.1	65.3	**79.3**	71.5
ORL-3	SRL-to-ORL 3 epochs	1	74.1***	55.3***	63.3***	68.4	85.0***	75.8***	70.7	77.6	74.0
		2	76.0	52.4	62.0	65.2	86.4**	74.3	71.8	76.9	74.3
		3	72.7	55.7	63.1	69.1	83.8*	75.8	67.8**	77.6**	72.4**
		avg	74.3	**54.5**	**62.8**	67.6	**85.1**	**75.3**	**70.1**	77.4	**73.6**

Table 7: Token-based evaluation: precision, recall and F1 (micro) for holders, targets and inferred speakers (asterisks indicate statistical significance for ORL-1 vs. ORL-2 and ORL-1 vs. ORL-3 according to an approximate randomisation test where * $p \leq 0.01$; ** $p \leq 0.001$; *** $p \leq 0.0001$).

	Example sentence					*measure*	TP	FP	FN
DE	Diese Auffassung	wird	auch	in einem Großteil der Lehre	vertreten.				
EN	This view	will	also	in a large part of the doctrine	be held.				
TRANS	"This view is also held by a large part of the doctrine."								
gold	Target			Holder					
auto	Target			Holder		*strict*	1	1	1
						tok-based	7	0	1

Table 8: Example sentence (constructed) illustrating the difference between the *strict* and the *token-based* evaluation (gold: gold annotation; auto: predicted labels; TP: true positives, FP: false positives, FN: false negatives).

Exp.	Frames		Holder			Target			Speaker (inferred)		
		#	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
ORL-1	*holder-only*	214	94.6	38.5	54.7	0	0	0	0	0	0
	target-only	247	0	0	0	86.0	79.4	82.6	0	0	0
	target+inferred	923	0	0	0	67.0	81.1	73.4	91.9	77.7	84.2
	holder+target	847	90.6	52.7	66.6	72.6	81.7	76.9	0	0	0
ORL-3	*holder-only*	214	92.0	45.3	60.7	0	0	0	0	0	0
	target-only	247	0	0	0	82.4	86.6	84.4	0	0	0
	target+inferred	923	0	0	0	64.3	86.1	73.6	94.4	79.7	86.4
	holder+target	847	90.4	54.2	67.8	70.5	86.7	77.8	0	0	0

Table 9: Token-based evaluation for different subsets of the test set.

Exp.	subjective expr.		Holder			Target			Speaker (inferred)		
	POS	#	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
ORL-1	V	823	84.0	62.6	71.7	69.2	88.8	77.8	64.9	70.2	67.4
	N	849	76.1	27.2	40.1	72.8	56.4	63.5	47.9	59.6	53.1
	A	404	66.2	36.2	46.8	67.7	78.8	72.8	78.9	95.1	86.3
ORL-3	V	823	82.1	65.0	72.6	67.7	91.7	77.9	68.3	73.6	70.8
	N	849	71.0	31.6	43.7	68.4	67.8	68.1	58.9	55.9	57.4
	A	404	59.2	30.5	40.2	64.6	84.0	73.0	76.5	93.3	84.0

Table 10: Token-based evaluation for verbal, nominal and adjectival subjective expressions (test set), excluding multi-word expressions.

We run an approximate randomisation test with 10,000 iterations on the output of the different models (ORL-1 vs. ORL-2 and ORL-1 vs. ORL-3) (Table 7). We can see that not all improvements are statistically significant. Only recall for target prediction (ORL-3) yields significant improvements for each individual run over the single-task system (ORL-1).

Table 7 also shows that, according to the token-based evaluation, the inferred holders are easier to identify than the explicit opinion holders, with around 10% higher F1. This is in contrast to the findings of Wiegand et al. (2019b, p.26) who state that inferred sources are "more difficult to detect than normal sources".

We can now answer our second research question, **RQ2**, and conclude that despite the small size of the German data set, it is possible to transfer knowledge from SRL to ORL. Improvements, however, are far more modest than the ones reported for English (Marasovic and Frank, 2018) and mostly improve recall.

4.2 Error analysis

We now take a closer look at the results, to find out where transfer learning helps and what is still difficult for our models. For our error analysis, we look at the predictions of the ORL-1 single task model and the ORL-3 (SRL-to-ORL transfer) model.[5] We first compare the output of the two models, focussing on the performance on different subsets of the data, i.e., subjective frames that include only a holder (but no target), a target (but no holder), targets with inferred sources and frames with both, holder and target.

Table 9 shows that the largest improvements for the transfer model (Exp.3) are due to a higher recall for the subjective frames that include holders only. Here we observe an increase in F1 of 6% (from 54.7% to 60.7%) over the single-task model. The results also suggest that holder-only frames are the most difficult category for opinion role prediction, while F1 for holder prediction for frames that include both, holder *and* target, are substantially higher for both, the single-task and the transfer model.

Next, we investigate how our models perform on subjective expressions with different parts of speech (Table 10). Interesting but by no means unexpected is the decrease in results for the SRL-to-ORL model on adjectival triggers for opinion holders and inferred sources (for explicit holders from 46.8% to 40.2% and for inferred holders from 86.3% to 84%). The largest improvements can be observed for nominal subjective expressions. Here the additional knowledge about predicate argument structure helps the most which, on first glance, is a bit surprising, given that the German SRL data includes semantic roles for verbal predicates only. However, keeping in mind that the subjective expressions are already given, what we need to know in order to predict the opinion roles is which token spans are probable arguments. Our transfer model seems to have learned useful information for this task from SRL, as shown by the increase in F1 for nominal subjective expressions in the range of 3.6% (for holders) to 4.6% (for targets).

[5] We use the models for Exp. ORL-1 and ORL-3 from the 2nd run in our analysis.

5 Conclusions

In the paper, we have presented a transformer-based system for German ORL on parliamentary debates, with new state-of-the-art results for the IGGSA-STEPS shared task. We have further shown that we can improve our baseline system through transfer learning, based on knowledge about predicate argument structure learned from SRL. We include this information via intermediate training and show that we mostly obtain improvements for recall and, in particular, for nominal subjective expressions and subjective frames where only the holder is expressed.

One challenge for transfer learning is the imbalance between the SRL and ORL training data. In future work, we would thus like to explore whether adapters might help us to make more efficient use of the data by injecting knowledge about predicate argument structure in our model without outweighing the information learned from the ORL data.

Acknowledgements

This work was supported in part by a Research Seed Capital (RiSC) grant, funded by the "Ministerium für Wissenschaft, Forschung und Kunst Baden-Württemberg" (MWK BW). We would like to thank the reviewers for their thorough and constructive feedback.

References

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet Project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, ACL '98/COLING '98, page 86–90, USA. Association for Computational Linguistics.

Ankur Bapna and Orhan Firat. 2019. Simple, scalable adaptation for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1538–1548, Hong Kong, China. Association for Computational Linguistics.

Lukas Biewald. 2020. Experiment tracking with weights and biases. Software available from wandb.com.

Aljoscha Burchardt, Katrin Erk, Anette Frank, Andrea Kowalski, and Sebastian Pado. 2006a. SALTO: A versatile multi-level annotation tool. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC 2006, Genoa, Italy, May 22-28, 2006*, pages 517–520, Genoa, Italy.

Aljoscha Burchardt, Katrin Erk, Anette Frank, Andrea Kowalski, and Sebastian Padó. 2006b. The SALSA Corpus: a German Corpus Resource for Lexical Semantics. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC 2006, Genoa, Italy, May 22-28, 2006*, pages 969–974. European Language Resources Association (ELRA).

Yejin Choi, Eric Breck, and Claire Cardie. 2006. Joint extraction of entities and relations for opinion recognition. In *EMNLP 2006, Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, 22-23 July 2006, Sydney, Australia*, pages 431–439. ACL.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, Pavel Straňák, Mihai Surdeanu, Nianwen Xue, and Yi Zhang. 2009. The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task*, pages 1–18, Boulder, Colorado. Association for Computational Linguistics.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Richard Johansson and Alessandro Moschitti. 2013. Relational features in fine-grained opinion analysis. *Computational Linguistics*, 39(3):473–509.

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)*, pages 919–929, Berlin, Germany. Association for Computational Linguistics.

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, Sydney, Australia. Association for Computational Linguistics.

Bing Liu. 2010. *Sentiment analysis and subjectivity*, volume 2, pages 627–666. Taylor and Francis Group.

Ana Marasovic and Anette Frank. 2018. SRL4ORL: Improving Opinion Role Labeling Using Multi-Task Learning with Semantic Role Labeling. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 583–594. Association for Computational Linguistics.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71–106.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. In *NIPS Autodiff Workshop*.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020. AdapterHub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54, Online. Association for Computational Linguistics.

Wei Quan, Jinli Zhang, and Xiaohua Tony Hu. 2019. End-to-End Joint Opinion Role Labeling with BERT. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 2438–2446, Los Alamitos, CA, USA. IEEE Computer Society.

Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. 2018. Efficient parametrization of multi-domain deep neural networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8119–8127.

Josef Ruppenhofer, Julia Maria Struß, Jonathan Sonntag, and Stefan Gindl. 2014. IGGSA-STEPS: Shared Task on Source and Target Extraction from Political Speeches. *Journal for Language Technology and Computational Linguistics*, 29(1):33 – 46.

Josef Ruppenhofer, Julia Maria Struß, and Michael Wiegand. 2016. Overview of the IGGSA 2016 Shared Task on Source and Target Extraction from Political Speeches. In *Proceedings of the IGGSA Shared Task 2016 Workshop*, pages 1–9.

Peng Shi and Jimmy J. Lin. 2019. Simple BERT Models for Relation Extraction and Semantic Role Labeling. *CoRR*, abs/1904.05255.

Veselin Stoyanov, Claire Cardie, Diane Litman, and Janyce Wiebe. 2004. Evaluating an opinion annotation scheme using a new multi-perspective question and answer corpus. In *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

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

Michael Wiegand, Nadisha-Marie Aliman, Tatjana Anikina, Patrick Carroll, Margarita Chikobava, Erik Hahn, Marina Haid, Katja König, Leonie Lapp, Artuur Leeuwenberg, Martin Wolf, and Maximilian Wolf. 2016. Saarland University's Participation in the Second Shared Task on Source, Subjective Expression and Target Extraction from Political Speeches. In *Proceedings of the IGGSA Shared Task 2016 Workshop*, pages 14–23.

Michael Wiegand, Margarita Chikobava, and Josef Ruppenhofer. 2019a. A Supervised Learning Approach for the Extraction of Sources and Targets from German Text. In *Proceedings of the 15th Conference on Natural Language Processing, KONVENS 2019, Erlangen, Germany, October 9-11, 2019*.

Michael Wiegand, Leonie Lapp, and Josef Ruppenhofer. 2019b. A descriptive analysis of a German corpus annotated with opinion sources and targets. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)*, pages 20 – 29, München [u.a.]. German Society for Computational Linguistics & Language Technology und Friedrich-Alexander-Universität Erlangen-Nürnberg.

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*, pages 215–225, Beijing, China. Association for Computational Linguistics.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, pages 38–45. Association for Computational Linguistics.

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)*, pages 1640–1649, Sofia, Bulgaria. Association for Computational Linguistics.