Automatic Assignment of Semantic Frames in Disaster Response Team Communication Dialogues

Natalia Skachkova and Ivana Kruijff-Korbayová
DFKI, Saarland Informatics Campus, 66123 Saarbrücken Germany
natalia.skachkova@dfki.de;ivana.kruijff@dfki.de

Abstract

We investigate frame semantics as a meaning representation framework for team communication in a disaster response scenario. We focus on the automatic frame assignment and retrain PAFIBERT, which is one of the state-of-the-art frame classifiers, on English and German disaster response team communication data, obtaining accuracy around 90%. We examine the performance of both models and discuss their adjustments, such as sampling of additional training instances from an unrelated domain and adding extra lexical and discourse features to input token representations. We show that sampling has some positive effect on the German frame classifier, discuss an unexpected impact of extra features on the models' behaviour and perform a careful error analysis.

1 Introduction

In this paper we employ the theory of frame semantics as a meaning representation framework for dialogues from the domain of disaster response. Our work is part of a larger research project developing methods to capture and interpret verbal team communication in disaster response scenarios and use the extracted run-time mission knowledge for mission process assistance, as described in (Willms et al., 2019). Team communication interpretation encompasses several aspects, some of which have been addressed in earlier publications of our team: (Anikina and Kruijff-Korbayova, 2019) present dialogue act classification results; (Skachkova and Kruijff-Korbayova, 2020) provide an analysis of contextual reference phenomena. The present paper complements this by results on semantic frame assignment. To our knowledge, our work is the first to use semantic frames in the domain of disaster response, and one of the few attempts implementing a frame classifier for dialogue.

Frame semantics is a paradigm defining the meaning of words through the context they are used in (Fillmore, 1976). This assumes that, depending on context, a word (or an expression) is able to evoke in our minds a certain event or situation together with a set of slots called frame elements associated with it, even if some of these slots were not explicitly filled in the sentence.

Using frame semantics as a meaning representation requires frame semantic parsing, namely identifying frame-evoking elements (targets) and the corresponding frames, as well as recognizing certain spans as frame elements and classifying them. In this paper we address the task of automatic semantic frame assignment, given a target. The novelty is that we work on English and German dialogues in robot-assisted disaster response teams.

We use the TRADR corpus (Kruijff-Korbayová et al., 2015), which contains transcribed communication in teams of firefighters using robots for incident site reconnaissance during a series of exercises that simulated situations after a disaster, such as a fire, explosion, etc. Towards the aim of creating structured representations of the events and activities during a first response mission by means of semantic frames, we experiment with some existing models. We start with a simple sequence classification approach that assumes fine-tuning of a pretrained BERT model (Devlin et al., 2019) on the TRADR corpus. Next, we use one of the existing state-of-the-art frame classifiers called PAFIBERT (Tan and Na, 2019). We re-implement and train it on the English FrameNet (Baker et al., 1998) data, and evaluate the model on English TRADR dialogues. We also experiment with re-training PAFIBERT on the TRADR data, despite the small corpus size. In addition, we investigate a possibility of training a frame classifier on mixed data - FrameNet or SALSA (Burchardt et al., 2006) plus TRADR - and consider three sampling approaches. Finally, we examine whether enriching the input

Proceedings of the 14th International Conference on Computational Semantics, pages 93–109
June 17–18, 2021.

with lexical and discourse features has an effect on the classifier performance. In contrast to many papers that report standard accuracy or F-score to measure the performance of a frame classifier, we use the index of balanced accuracy metric (García et al., 2009) designed specifically for imbalanced data.

In Section 2 we give a brief overview of the theory of frame semantics. In Section 3 we introduce the most notable frameworks designed to perform automatic frame assignment or frame-semantic parsing. In Section 4 we examine the distribution of semantic frames and the role of ambiguous targets in the TRADR corpus, compare our data with FrameNet and SALSA, and explain how we prepared and split all the data into training, validation and test sets. Section 5 describes the experiments and their results. In Section 6 we make a conclusion and indicate possible further steps.

2 Frame Semantics

According to Petruck (2019), frame semantics is a research program in empirical semantics which emphasizes the continuities between language and experience, and provides a framework for presenting the results of that research.

The theory of frame semantics goes back to the 1970s. One of the pioneers in this area was Charles J. Fillmore. He suggested that a language description should include not only lexicon and grammar, but also a set of 'frames' that incorporate the semantics of the language elements (Fillmore, 1976). Fillmore (1982) uses the word 'frame' as a general cover term for such concepts as 'schema', 'script', 'scenario', or 'cognitive model'. He defines a frame as a system of concepts which are related to each other, and states that one cannot understand a concept without understanding the whole structure it is a part of. Frame semantics tries to describe and formalize such structures.

The FrameNet project (Baker et al., 1998) is considered one of the first practical realizations of the theory of frame semantics for English. One of its achievements was the creation of a lexical database that covers more than 13,000 word senses, is both human- and machine-readable and available online. Besides, more than 200,000 sentences were annotated with about 1,200 semantic frames, and are now known as the FrameNet corpus.

Examples 2.1 and 2.2 present a definition of the *Inspecting* frame and its frame elements (FEs), annotated with respect to the target *inspected*. Note that FEs can be 'core' (i.e. essential to the meaning of a frame) and 'non-core' (i.e. not uniquely characterizing). Usually, core FEs are part of the frame definition, like INSPECTOR and GROUND in Example 2.2.

Example 2.1 *'Inspecting' Frame Definition*
An INSPECTOR *directs his/her perceptual attention to a* GROUND *to ascertain whether the* GROUND *is intact or whether an* UNWANTED_ENTITY *is present. Alternatively, the desired outcome of the inspection may be presented as a* PURPOSE.

Example 2.2 *'Inspecting' Frame's FEs*
[$_{INSPECTOR}$ *He] moved toward the control panel and [*$_{TARGET}$ ***inspected***] *[*$_{GROUND}$ *it] [*$_{LOCATION\_OF\_PROTAGONIST}$ *from a distance], [*$_{MEANS}$ *without touching it].*

Databases similar to FrameNet were also created for other languages. In Section 4 we compare the FrameNet corpus and its German counterpart SALSA with the TRADR data.

3 Related Work

Frame semantics is not one of the most common meaning representation frameworks. However, research in the area of frame-semantic parsing has increased since frame-semantic structure extraction was included as a task in SemEval'07 (Baker et al., 2007). Most of the existing works present models trained on text data. Some of the projects deal only with automatic frame assignment, others have a bigger goal, namely, recognizing targets, frames and frame elements. In what follows we will focus on automatic frame assignment.

Most of the early frameworks are based on the idea of learning the frame labels from frame-evoking targets represented as rather elaborated sets of features, which include the target's lemma, its part of speech, etc. Many features rely on dependency syntax. For non-ambiguous targets a frame can be retrieved using a simple mapping. If the target is ambiguous, the correct label is learned using a Naive Bayes classifier, e.g., as shown by Erk (2005), or an SVM classifier like in the framework called LTH (Johansson and Nugues, 2007), or a discriminative probabilistic (log-linear) model like in SEMAFOR by Das et al. (2010).

The success of neural networks for many NLP tasks resulted in a gradual switch from the feature-based approaches to embeddings and a broader

usage of neural networks for the task of automatic frame assignment. One of the first semantic parsers to use embeddings was developed by Hermann et al. (2014). They represent targets as vectors, certain parts of which are reserved for certain argument representations. All frame labels are also vectors, and the classifier learns to minimize the distance between the targets and the correct labels. Other frameworks based on embeddings and various types of neural networks include SimpleFrameId (Hartmann et al., 2017) - a two-layer network which also allows to perform frame filtering using mappings of certain lexical units to certain frames from the FrameNet database; a framework by Yang and Mitchell (2017) that performs frame identification using a simple multi-layer network; TSABCNN (Zhao et al., 2018), which uses *word2vec* embeddings and convolutional neural networks.

Recently, there appeared frameworks that rely on BERT embeddings and pretrained models. E.g., PAFIBERT (Tan and Na, 2019) fine-tunes the pretrained BERT model using an attention mechanism to give weights to words that make up the context of the target. An interesting alternative approach was presented by Kalyanpur et al. (2020). They interpret frame-semantic parsing as a sequence-to-sequence generation problem. Their approach is based on the encoder-decoder architecture, namely on the T5 model, which is available via the *HuggingFace* library (Wolf et al., 2020).

Ribeiro et al. (2020) treat automatic frame assignment as a clustering problem. They focus on verbal frame-evoking targets and represent them using contextualized ELMo embeddings. The targets are treated as nodes in a graph, and clustered using the *Chinese Whispers* algorithm (Biemann, 2006). A new instance is classified by determining the closest cluster.

All the above frameworks were trained on text data. We found only two frame-semantic parsers designed specifically for dialogue. One of them was created in the course of the LUNA project (Raymond et al., 2008) and focuses mostly on frame element classification (Coppola et al., 2008). The other was presented by Trione et al. (2015). Its main goal is actually to speed up the manual annotation process, not pure frame-semantic parsing. Frames are detected with the help of a hand-crafted set of lexical triggers, which includes 200 most frequent words from 7 domains.

A comparison of the frameworks mentioned above, as well as the results of their evaluation on the test data can be found in Appendix D. We do not place them here for space reasons.

For the experiments on the TRADR data presented in this paper we have chosen the PAFIBERT approach (Tan and Na, 2019). PAFIBERT is one of the state-of-the-art frame classifiers, it showed about 89% accuracy when evaluated on the FrameNet test set, and it is easy to re-implement.

4 Data for experiments

The TRADR corpus consists of 15 files with dialogues, six files contain dialogues in English, and nine - in German. Six German dialogues were translated into English in order to get more English training data. TRADR dialogues comprise the communication in first responder teams using robots for disaster site reconnaissance. Each team consists of several operators (OP) who control ground and airborne robots, a team leader (TL) and sometimes also a mission commander (MC).

Table 1 shows the distribution of dialogue turns, utterances and tokens between the mission participants in both English and German TRADR dialogues. Also, average numbers of utterances per turn and tokens per utterance are given. We see that both English and German parts of the data contain approximately the same number of dialogue turns, however the turns in the English dialogues are slightly longer, and as a result the English part of the corpus is 1.5 times larger. The utterances are usually rather short - 7-9 tokens on average, as the team participants try to be brief and precise.

	MC	TL	OP	Total
German data				
# Dialogue turns	60	984	1,020	2,064
# Utterances	61	997	1,027	2,085
# Tokens	526	6,165	7,875	14,566
Avg. # utt. per DT	1.02	1.01	1.01	1.01
Avg. # tokens per utt.	8.62	6.18	7.67	6.99
English data (including translations)				
# Dialogue turns	60	1,013	1,021	2,094
# Utterances	61	1,306	1,186	2,553
# Tokens	820	9,983	11,353	22,156
Avg. # utt. per DT	1.02	1.29	1.16	1.23
Avg. # tokens per utt.	13.44	7.64	9.57	8.68

Table 1: TRADR corpus overview

We annotated the utterances in the English TRADR dialogues with frame-evoking targets, corresponding lexical units (LUs), frames and parent frames. Frame elements were not annotated. The German TRADR data was annotated similarly, except that we replaced targets and LUs with 'tar-

get related elements', which represent the whole phrase that the target is a part of. We assumed that each utterance can potentially have several targets or groups of frame related elements. As a result, the number of frame instances in the TRADR corpus is larger than the number of utterances given in Table 1. While annotating our data with semantic frames we tried to follow the FrameNet annotation guidelines (Ruppenhofer et al., 2006). Due to the specifics of our domain, many FrameNet frame definitions had to be adapted. Also, ten new frames were introduced. The English and German parts of the corpus were annotated by two different annotators. To check the reliability of the annotation, one dialogue in German (534 frame instances) was also annotated by the person responsible for the annotation of the English dialogues. Inter-annotator agreement measured using Cohen's Kappa (Carletta, 1996) reached 0.73, which is considered reliable. A team communication example annotated with semantic frames, as well as the definitions of the new frames are available in Appendix C. We are making the annotated data available online.[1]

In total, the English and German parts of the TRADR corpus contain 4,191 and 3,519 frame instances, respectively. These instances are distributed between 190 (English) and 152 (German) different frame labels. The distribution of the frame labels is not uniform.Thus, in English TRADR almost 60% of all the instances belong to the top ten most frequent frames, and 137 out of 190 frames have only ten or less samples, which all together make up about 10% of the data. In German TRADR the instances of the top ten most frequent frames make up approximately 58%, and instances of 105 infrequent frames - almost 11% of the data. The fact that the TRADR data is highly imbalanced motivates the choice of performance metrics for the evaluation of the frame classifiers that will be discussed in the next section.

The English TRADR data counts 434 different LUs. Their distribution is also not uniform: the top ten most common LUs occur in about 40% of all the utterances and at the same time make only slightly more than 2% of the total of different LUs. All LUs are distributed between seven different POS tags. 75% of the utterances contain verbal targets. The second frequent POS tag is an interjection - almost 8% of all the targets.

[1]The TRADR data and the semantic frame annotations can be obtained at http://talkingrobots.dfki.de/.

Only about 15% of all LUs in English TRADR are ambiguous. However, they are realized in nearly 53% of utterances containing targets. Simple calculations show that on average a single LU evokes 1.24 frames. So, while the ambiguous LUs are not very frequent in comparison to non-ambiguous ones, the frames that they evoke are frequent, and this may become a problem for the frame classifier, as it is not always possible to perform frame disambiguation using the utterance context.

Besides TRADR we also use the FrameNet and SALSA datasets for our experiments, so it is necessary to compare them with our data. The differences between the corpora are summarized in Table 14, presented in Appendix. Note that for the experiments all duplicate sentences/utterances (i.e. equal strings with equal labels), as well as elliptical utterances and communication fragments (in TRADR) were removed. The numbers in Table 14 are based on the cleaned versions of the corpora. The only exception is the average utterance length in the TRADR corpus, that was calculated based on the original data in Table 1.

The FrameNet and SALSA data are very different from TRADR, cf. Table 14. First, they are much larger and come from other domains (note that the domains of FrameNet and SALSA are quite close to each other). Both FrameNet and SALSA include many more frames than TRADR, and despite the fact that many frames are common for all the corpora (e.g., about 93% of frame labels in English TRADR also occur in FrameNet), the frame distributions are very different. The fact that less than 65% of TRADR LUs are common with FrameNet LUs, which are much more numerous, supports this. Both FrameNet and SALSA are also imbalanced and FrameNet contains ambiguous targets.

Data	TRADR				Frame-Net	# cls
	Eng	# cls	Ger	# cls		
Training	1,955	81	1,902	72	143,509	931
Validation	489	81	476	72	35,877	931
Test	268	81	259	72	19,923	931
Test (subs.)	234	50	-	-	-	-

Table 2: Training, validation & test data sizes

All the datasets were shuffled and randomly split into training, validation and test data as shown in Table 2. Note that the number of classes (frame labels) is smaller than given in Table 14, as all the frames that have less than five instances were removed. This was necessary to perform 5-fold cross-

validation. Note that we have two English TRADR test sets. The second one is a subset of the first one, and contains the instances of 50 frames common to both FrameNet and TRADR. It is needed to test the PAFIBERT model trained on FrameNet.

5 Experiments and Discussion

In this section we will present semantic frame classifiers for both English and German TRADR dialogues. Our main focus is on the English data. We introduce several models, split into basic and adjusted, and discuss their performance.

As all our datasets have hundreds of classes and are highly imbalanced, many typical performance metrics, e.g., accuracy, precision, F-score, are not reliable (Tharwat, 2020). Instead, we use the index of balanced accuracy (IBA) metric as our main performance measure, calculated using the *Python imbalanced-learn* package (Lemaître et al., 2017). The package also outputs the scores of the common metrics, such as recall, precision and F-score, and we show them for the sake of comparison, as most papers on automatic frame assignment report either accuracy, or these metrics. All the metrics are calculated using macro-averaging.

5.1 Basic models

The first group includes four models. The first one is a naive baseline, represented by the *BertForSequenceClassification* model from the *Transformers* library (Wolf et al., 2020) fine-tuned on English TRADR. *BertForSequenceClassification* was chosen as the most straightforward way to perform sequence classification. It is a pretrained BERT model with an additional linear layer on top of the pooled output. The other three models reproduce the architecture of PAFIBERT. The implementation details can be found in the original paper by Tan and Na (2019). One of the models was trained on the FrameNet data, another - purely on English TRADR data, the last one - on German TRADR data.

All four models were trained with 5-fold cross-validation. As both English and German TRADR datasets are small, different splits into training and test parts may result in noticeable performance variance. We used cross-validation to get a more reliable estimation of the performance of the models, not for hyper-parameter search. All hyper-parameters were taken from the original paper. Following Tan and Na (2019), training was performed for 8 epochs per fold using an adaptive learning rate that starts with 3e-5 and an *AdamW* optimizer. In the course of cross-validation we always saved the model with the best IBA validation score. Next, the model was evaluated on the test data.

The performance of the basic models is summarized in Table 3. We see that *BertForSequenceClassification* demonstrates rather unsatisfactory performance - IBA only 32% - 37%. The reason for this is the fact that simple fine-tuning does not integrate information about the frame-evoking targets and their contexts, so that it is impossible for the model to guess what tokens in the sequence it has to focus on. It is obvious that in order to improve the performance, we need to tell the model which tokens in each utterance it should pay attention to, and PAFIBERT provides a convenient way to do so.

Classifier	Test set	PRE	REC	F1	IBA
BertForSequence-Classification (EN)	TR (EN)	0.33	0.39	0.35	0.37
	TR (subs.)	0.30	0.35	0.31	0.32
PAFIBERT trained on FrameNet (EN)	FN	0.92	0.92	0.92	0.91
	TR (subs.)	0.71	0.53	0.58	0.51
Basic model (EN)	TR (EN)	0.90	0.89	0.89	0.88
	TR (subs.)	0.91	0.88	0.88	0.86
Basic model (DE)	TR test (DE)	0.84	0.84	0.83	0.83

Table 3: Basic models: results; "TR" stands for TRADR test set, "FN" for FrameNet test.

As Table 3 shows, PAFIBERT trained on the FrameNet data has IBA of 91% when evaluated on the test set coming from the same distribution. This score is actually even slightly better than the standard accuracy of 89% reported by Tan and Na (2019). However, when tested on TRADR data, the model shows much worse results, namely, only 51% IBA, despite the fact that the majority of the 50 frames from the given test set have enough instances in the training set.

The main reasons why this classifier fails on the TRADR data are as follows. First of all, due to the fact that FrameNet is very fine-grained, many TRADR instances got classified as belonging to very specific frames which we did not use when annotating the TRADR data, like *'Interior_profile_relation'* and *'Non_gradable_proximity'* (we used their parent frame *'Locative_relation'* instead). Another reason is that TRADR instances of certain frames have targets that, due to domain differences, are not typical for these frames in FrameNet. For instance, all TRADR samples of *'Create_representation'* frame were misclassified, because the model expected *'draw'*, *'carve'* or *'sketch'* as targets, but got

'take/make a picture' and labeled the input utterances as *'Physical_artwork'* instead. Finally, there is also a problem of ambiguity. For example, the target *'change'* can evoke both *'Replacing'* and *'Cause_change'* frames, and the target *'lie'* - *'Posture'* and *'Being_located'*.

So, the error analysis shows that the PAFIBERT model trained on FrameNet is domain-specific, it does not generalize well, and we cannot simply reuse it for TRADR data without special modifications or further fine-tuning.

Now let us have a look at the performance of the PAFIBERT models trained on English and German TRADR. Despite the relatively small size of the training data, the models manage to achieve IBA scores of about 88% (English) and 83% (German). The English model also demonstrates quite good performance (86% IBA) on the subset of the main English TRADR test set, used to evaluate PAFIBERT trained on the FrameNet data. Notice that the IBA metric is fairer than standard accuracy: despite the fact that the subset of the TRADR test set does not contain the instances of the most frequent domain-specific *'Communication_by_protocol'* and *'Communication_response_message'* frames, which are easier to recognize due to their shortness and typical structure, the IBA score for this test set is only 2% lower.

Classifier model	Basic (EN)	Basic (DE)
# errors	30/268	42/259
Target ambiguity	22/30 (73%)	26/42 (62%)
Silly mistakes	5/30 (17%)	14/42 (33%)
Incorrect parsing	2/30 (7%)	2/42 (5%)
Incorrect translations	1/30 (3%)	-

Table 4: PAFIBERT trained on TRADR: error analysis

In order to understand why we have 5% difference in performance between the English and German frame classifiers trained on TRADR, we performed error analysis. The results are summarized in Table 4. We see that the majority of errors happens because of ambiguous targets, and the proportion of such errors is about 10% higher among the errors made by the English frame classifier. At the same time the German frame classifier makes much more the so-called silly mistakes, which encompass the cases when the assigned frame has nothing to do with the given target. We attribute the worse performance of the German classifier mostly to the fact that instead of targets we used 'frame related elements', which sometimes contain several tokens and can be confusing for the classifier. Differences between the languages (i.e. in morphology, syntax, semantics) may also be important. E.g., verbs with separable prefixes, like *'zurückkehren'* or *'vorbeikommen'*, as targets may lead to errors, as the prefixes often get disregarded. Finally, because of small test sizes, the role of chance (in)correct assignments may get exaggerated.

5.2 Adjustments of PAFIBERT

Aiming at performance improvement, we experimented with several adjustments of the PAFIBERT model trained on TRADR. Below we discuss the results and analyse the errors.

Sampling We performed a series of experiments with sampling additional training examples from the subsets of the FrameNet and SALSA corpora, which contain only instances of those frames that occur in TRADR. The FrameNet subset for sampling has 21,492 instances (about 12% of the whole FrameNet corpus), the SALSA subset - 2,486 (about 7% of the corpus). The experiments can be split into two groups. The first group includes training models with different portions of blindly sampled data. The second part involves experiments with informed sampling. Each model is trained on a mixture of TRADR and sampled data, and validated solely on TRADR data.

In the blind sampling scenario we train ten models gradually increasing the amount of additional training examples randomly chosen from the FrameNet or SALSA subsets.

# sampled inst.	PRE	REC	F1	$IBA_{0.1}$
2,149 inst. (10%)	0.92	0.90	0.90	**0.89**
4,298 inst. (20%)	0.91	0.87	0.88	0.86
6,447 inst. (30%)	0.91	0.89	0.89	0.88
8,596 inst. (40%)	0.92	0.90	0.90	**0.89**
10,746 inst. (50%)	0.92	0.89	0.89	0.88
12,895 inst. (60%)	0.91	0.89	0.89	0.88
15,044 inst. (70%)	0.92	0.89	0.89	0.88
17,193 inst. (80%)	0.91	0.88	0.88	0.87
19,342 inst. (90%)	0.92	0.88	0.89	0.87
21,492 inst. (100%)	0.91	0.90	0.90	**0.89**
Basic model (EN)	0.90	0.89	0.89	0.88

Table 5: Blind random sampling from FrameNet

The results for the English frame classifier are in Table 5. We see that there is no clear correlation between the sampled data size and performance. Three models demonstrate an improvement by 1% in comparison with the basic model, however, this difference is insignificant according to the McNe-

mar's test. A lack of positive influence of the blind sampling can be caused by the fact that the subset of the FrameNet data used for sampling only contains a small amount of really useful instances. If only a part of the subset is sampled, these instances have high chances to be left out due the randomization of the sampling procedure. In case the whole subset is sampled, the additional instances may dominate the original ones, as the FrameNet subset for sampling is much larger than TRADR.

In contrast to this, the effect of the blind random sampling on the German frame classifier is clearly positive. As Table 6 shows, having more training data leads to the IBA score increase by 4%.

# sampled inst.	PRE	REC	F1	$IBA_{0.1}$
248 inst. (10%)	0.86	0.85	0.85	0.84
497 inst. (20%)	0.88	0.88	0.87	**0.87**
745 inst. (30%)	0.87	0.87	0.86	0.86
994 inst. (40%)	0.87	0.85	0.85	0.84
1,243 inst. (50%)	0.89	0.88	0.88	**0.87**
1,491 inst. (60%)	0.89	0.88	0.87	**0.87**
1,740 inst. (70%)	0.89	0.88	0.87	**0.87**
1,988 inst. (80%)	0.89	0.88	0.88	**0.87**
2,237 inst. (90%)	0.89	0.88	0.87	**0.87**
2,486 inst. (100%)	0.90	0.88	0.88	**0.87**
Basic model (DE)	0.84	0.84	0.83	0.83

Table 6: Blind random sampling from SALSA

To get an explanation why blind sampling has a different impact on the two classifiers, we plot the learning curves that show how training and validation losses depend on the proportion of sampled data. As Figure 1 shows, adding training instances from FrameNet and SALSA does not lead to validation loss decrease and better generalization ability of the models. Notice that even without sampling the gap between the two curves in each plot is large, with training losses being close to zero, which is usually interpreted as overfitting. This finding lead us to check the learning curves of PAFIBERT trained on the much larger FrameNet data. The overfitting problem occurs in that case, too (see Appendix A). To tackle the overfitting issue, we tried out several experiments with increased dropout rate and fewer training epochs, but they only led to the IBA score decrease. We conclude that some fundamental changes in PAFIBERT's architecture would be needed to avoid overfitting.

In both plots in Figure 1 the validation loss grows together with the number of sampled examples. This means that even if the models continue making correct predictions, their confidence sinks. In case of the German frame classifier this growth is not so rapid, which can probably be explained by the fact that the SALSA subset for sampling is much smaller than the corresponding FrameNet subset. Knowing that IBA is actually improving, we hypothesize that sampled data from an unrelated domain can be helpful, but the right amount of these instances and their quality criteria are rather difficult to determine.

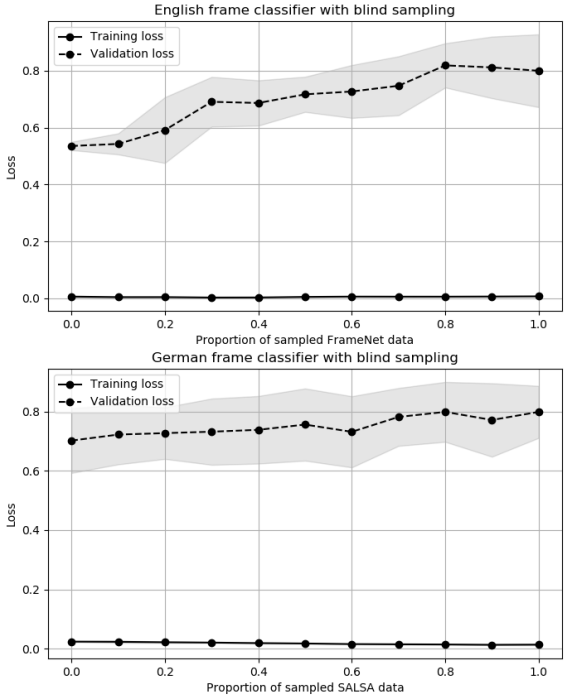

Figure 1: Learning curves of English and German frame classifies with blind sampling

The main disadvantage of blind sampling is that the instances are picked out regardless of their distribution in both original training data and data held out for sampling, which may aggravate the imbalance problem.

To overcome this, we tried two approaches using informed sampling. One is balancing sampling. It assumes sampling for each class no more than the maximum number of instances of the most common frame in the TRADR training data. The approach is supposed to deal with the class imbalance problem. However, this method also has a potential disadvantage. In case the number of original TRADR utterances is small, and the number of sampled instances is much larger, with their targets being different from those in the original utterances, the model will be biased towards the dominating training samples and thus prone to misclassification of the TRADR test examples. To avoid this we introduce equal sampling, which has an additional

constraint that the number of sampled examples cannot exceed the number of the original ones.

The scores in Table 7 show that the informed sampling does not produce the expected positive effect on the English frame classifier.

Sampling type	PRE	REC	F1	$IBA_{0.1}$
Balancing: 10,902 inst. ($\approx$ 51%)	0.91	0.88	0.88	0.87
Equal: 1,622 inst. ($\approx$ 7.5%)	0.92	0.89	0.89	0.88
Basic model (EN)	0.90	0.89	0.89	0.88

Table 7: Informed random sampling from FrameNet

However, as Table 8 demonstrates, in case of the German frame classifier the informed sampling clearly has a positive influence. Its significance was confirmed by the corresponding McNemar's tests. Balancing sampling helps to reduce the number of silly errors, as well as errors caused by ambiguous target expressions. The latter reduction is mostly due to a better recognition of ambiguous target expressions represented by single tokens, reflexive verbs and verbs with separable prefixes. The difference in performance between the balancing and equal sampling approaches was insignificant according to McNemar's test.

Sampling type	PRE	REC	F1	$IBA_{0.1}$
Balancing: 2,375 inst. ($\approx$ 96%)	0.89	0.89	0.88	0.88
Equal: 611 inst. ($\approx$ 25%)	0.91	0.91	0.90	**0.90**
Basic model (DE)	0.84	0.84	0.83	0.83

Table 8: Informed random sampling from SALSA

So, we see that sampling failed to produce any positive effect on the English frame classifier, but worked for the German one. We hypothesize that this happens to a large extent because sampling mostly helps to resolve simple mistakes, but is less effective in cases where disambiguation is necessary. More complex morphology of German may also be a reason why additional training examples proved to be more useful.

Extra features Our next goal is to check if extending BERT embeddings with extra features has any positive impact on the performance of PAFIBERT. We divide the features into two groups: lexical features that include POS tags and subword masks, and discourse features represented by speaker tags and dialogue acts. Our modifications of the original architecture by Tan and Na (2019) are given in Appendix, Figures 3 and 4.

The introduction of lexical features is motivated by the following reasons. First, we have cases, when the POS tag of a target may be important to differentiate one frame from another. E.g., in the utterance *"Can you position yourself onto the track?"* the target *'position'* is a verb and evokes the *'Placing'* frame, while in the utterance *"What's your current position?"* *'position'* is a noun that induces the frame *'Locale_by_collocation'*. Second, BERT tokenization splits the tokens that are not included in the tokenizer vocabulary, and sometimes it happens that some parts of a token lie outside of the target's context window.

POS tagging was done with a tagger from the *Python SpaCy* library (Honnibal and Montani, 2017). There are 19 coarse-grained tags that follow the Universal Dependencies scheme. We add two more tags to this set: SPECIAL to mark special tokens used by BERT and separate them from 'normal' ones, and PAD for padded tokens. If a token gets split by the tokenizer, each sub-token is assigned the POS tag of the original word. Our subword masks are bit vectors where all sub-tokens are marked with ones, and intact tokens - with zeros.

Embeddings for lexical features are trained together with the model. They are concatenated with the BERT model output, namely with (sub)token vectors, and used as input for the position-based attention layer of PAFIBERT. As (sub)token representations get longer, we have to increase the size of the first linear layer of PAFIBERT accordingly.

The second group of additional features includes discourse features, namely the speaker tag and dialogue act type, which also can be useful for frame disambiguation. E.g., given a short utterance *"Try it"* with the target *'try'*, the classifier may have difficulties labeling it, because to assign the correct frame it needs to know the perspective, i.e. the speaker. If the speaker is the team leader, then the correct frame is *'Attempt_suasion'*, if it is an operator, then it should be the *'Attempt'* frame. The information about the dialogue act type can be used to strengthen the impact of the speaker tag, because there exist a strong correlation between the speaker and the dialogue act in the tradr dialogues (Anikina and Kruijff-Korbayová, 2019).

Following Anikina and Kruijff-Korbayová (2019), we use three labels to encode the speakers:

MC for the mission commander, TL for the team leader and OPERATOR for the rest of the team.

As for dialogue acts, we use 12 labels based on the ISO-24617-2 guidelines Bunt (2019), with a few modifications. Eight tags correspond to those used in Anikina and Kruijff-Korbayová (2019): *'Affirmative'*, *'Confirm'*, *'Contact'*, *'Disconfirm'*, *'Inform'*, *'Negative'*, *'Question'* and *'Request'*. The other four labels are *'Communication Management'*, *'Time Management'*, *'Discourse Structuring'* and *'Social Obligations'*.

Embeddings for discourse features are trained jointly with the model. Since they characterize the whole utterance and not separate (sub)tokens, we concatenate them with the output of the PAFIBERT position-based attention layer. We increase the size of the first linear layer in the model accordingly.

The performance of the English frame classifier trained on the data enriched with lexical and dialogue features is given in Table 9. We test the features separately and in combinations. We see that taken separately, the features do not bring any improvement, and sometimes the scores are actually slightly worse than the score achieved by the basic classifier. The combination of POS tags and subword masks seems to increase the performance by 1%, but the difference is insignificant according to the McNemar's test.

Feature	PRE	REC	F1	$IBA_{0.1}$
POS tag	0.89	0.88	0.88	0.87
Subword mask	0.89	0.88	0.87	0.87
POS tag + Subw. mask	0.91	0.90	0.90	**0.89**
Speaker	0.89	0.88	0.88	0.88
Dialogue act	0.89	0.87	0.87	0.86
Speaker + Dialogue act	0.90	0.88	0.88	0.88
POS tag + Subw. mask + Sp.	0.88	0.88	0.87	0.87
Basic model (EN)	0.90	0.89	0.89	0.88

Table 9: Extra features: English frame classifier

As for the German frame classifier, we tested only the impact of extra lexical features. Dialogue features were not used, as the current data does not include speaker and dialogue act annotations. The results were similar to those demonstrated by the English frame classifier with extra lexical features. We do not include them here due to space constraints. They are available in Appendix B.

It is difficult to say why neither lexical nor discourse features lead to performance improvement. One of possible reason is that our learned feature embeddings are rather short (2-4 neurons) in comparison with input embeddings (768 neurons) or context-target embeddings (1536 neurons), so their impact on the whole (sub)token/utterance representations is actually negligible or even confusing. We think that in order to get a better estimation of the role of additional features, some further experiments with more data are necessary.

6 Conclusion and Future Work

We investigated the potential of frame semantics as a meaning representation framework for English and German dialogues in the domain of robot-assisted disaster response team communication. We found semantic frames convenient for capturing the meaning of an utterance depending on the target - the approach is span-based and does not require complex data annotation or pre-processing.

We reused the PAFIBERT model on the TRADR data and achieved an IBA score of 88%–90% on the test sets. Our results are comparable with those reported by Tan and Na (2019), who trained their models on the much larger FrameNet corpus. However, being a powerful model, PAFIBERT memorized the small TRADR training data, leading to overfitting and thus lack of generalization.

We also studied the impact of sampling additional training instances from an unrelated domain on the classifier's performance, and found that it was useful only for the German frame classifier. Error analysis indicates that sampling is beneficial for handling silly errors, but rather ineffective for cases that require disambiguation. We did not perform any experiments with over- and/or undersampling which imply sampling from the original dataset and are often used with imbalanced data. This can be a subject for further research. Especially interesting is an approach that assumes generating synthetic training instances, e.g., embeddings incorporating the targets with their contexts.

In contrast to our expectations, both lexical and discourse features failed to demonstrate a positive influence on the models' performance.

Error analysis showed that the largest group of errors is due to ambiguous targets, many of which evoke semantically close frames. The problem of disambiguation requires more research in order to improve the performance of the models.

Acknowledgments

This work is part of the project "A-DRZ: Setting up the German Rescue Robotics Center", funded by the German Ministry of Education and Research

(BMBF), grant No. I3N14856.[2] We would like to thank our A-DRZ colleagues for discussions, Daria Fedorova for data annotation and the IWCS 2021 reviewers for valuable comments which we hope helped us to improve the paper.

References

Tatiana Anikina and Ivana Kruijff-Korbayova. 2019. Dialogue act classification in team communication for robot assisted disaster response. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 399–410, Stockholm, Sweden. Association for Computational Linguistics.

Tatiana Anikina and Ivana Kruijff-Korbayová. 2019. Dialogue act classification in team communication for robot assisted disaster response. In *Proceedings of SIGDIAL 2019*.

Collin F. Baker, Michael Ellsworth, and Katrin Erk. 2007. SemEval-2007 Task 19: Frame semantic structure extraction. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 99–104.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *COLING-ACL '98: Proceedings of the Conference*, pages 86–90, Montreal, Canada.

Claude Barras, Edouard Geoffrois, Zhibiao Wu, and Mark Liberman. 1998. Transcriber: a free tool for segmenting, labeling and transcribing speech. In *First International Conference on Language Resources and Evaluation (LREC)*, pages 1373–1376.

Chris Biemann. 2006. Chinese whispers - an efficient graph clustering algorithm and its application to natural language processing problems. In *Proceedings of TextGraphs: the First Workshop on Graph Based Methods for Natural Language Processing*, pages 73–80, New York City. Association for Computational Linguistics.

Harry Bunt. 2019. *Guidelines for using ISO standard 24617-2*. [s.n.]. TiCC TR 2019–1.

Aljoscha Burchardt, Katrin Erk, Anette Frank, Andrea Kowalski, Sebastian Padó, and Manfred Pinkal. 2006. The SALSA corpus: a German corpus resource for lexical semantics. In *LREC*, pages 969–974.

Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Comput. Linguist.*, 22(2):249–254.

Bonaventura Coppola, Alessandro Moschitti, Sara Tonelli, and Giuseppe Riccardi. 2008. Automatic Framenet-based annotation of conversational speech. In *2008 IEEE Spoken Language Technology Workshop*, pages 73–76.

[2] https://www.rettungsrobotik.de

Dipanjan Das, Nathan Schneider, Desai Chen, and Noah A. Smith. 2010. Probabilistic frame-semantic parsing. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 948–956, Los Angeles, California. Association for Computational Linguistics.

Dipanjan Das and Noah A Smith. 2011. Semi-supervised frame-semantic parsing for unknown predicates. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1435–1444.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Katrin Erk. 2005. Frame assignment as word sense disambiguation. In *Proceedings of IWCS*, volume 6.

Charles J. Fillmore. 1976. Frame semantics and the nature of language. *Annals of the New York Academy of Sciences*, 280(1):20–32.

Charles J. Fillmore. 1982. *Frame semantics*, pages 111–137. Hanshin Publishing Co., Seoul, South Korea.

FrameNet. 2021. The official website for the FrameNet project. https://framenet.icsi.berkeley.edu/fndrupal/. Accessed: 2021-02-03.

Vicente García, Ramón Alberto Mollineda, and José Salvador Sánchez. 2009. Index of balanced accuracy: A performance measure for skewed class distributions. In *Iberian conference on pattern recognition and image analysis*, pages 441–448. Springer.

Silvana Hartmann, Ilia Kuznetsov, Teresa Martin, and Iryna Gurevych. 2017. Out-of-domain FrameNet semantic role labeling. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 471–482, Valencia, Spain. Association for Computational Linguistics.

Karl Moritz Hermann, Dipanjan Das, Jason Weston, and Kuzman Ganchev. 2014. Semantic frame identification with distributed word representations. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1448–1458, Baltimore, Maryland. Association for Computational Linguistics.

Matthew Honnibal and Ines Montani. 2017. SpaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Richard Johansson and Pierre Nugues. 2007. LTH: Semantic structure extraction using nonprojective dependency trees. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 227–230, Prague, Czech Republic. Association for Computational Linguistics.

Aditya Kalyanpur, Or Biran, Tom Breloff, Jennifer Chu-Carroll, Ariel Diertani, Owen Rambow, and Mark Sammons. 2020. Open-domain frame semantic parsing using transformers.

Ivana Kruijff-Korbayová, Francis Colas, Mario Gianni, Fiora Pirri, Joachim de Greeff, Koen Hindriks, Mark Neerincx, Petter Ögren, Tomáš Svoboda, and Rainer Worst. 2015. TRADR project: Long-term human-robot teaming for robot assisted disaster response. *KI - Künstliche Intelligenz*, 29(2):193–201.

Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. 2017. Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17):1–5.

Miriam R L Petruck. 2019. Meaning representation of null instantiated semantic roles in FrameNet. In *Proceedings of the First International Workshop on Designing Meaning Representations*, pages 121–127, Florence, Italy. Association for Computational Linguistics.

Christian Raymond, Kepa Joseba Rodriguez, and Giuseppe Riccardi. 2008. Active Annotation in the LUNA Italian Corpus of Spontaneous Dialogues. In *LREC*.

Eugénio Ribeiro, Andreia Sofia Teixeira, Ricardo Ribeiro, and David Martins de Matos. 2020. Semantic frame induction as a community detection problem.

Josef Ruppenhofer, Michael Ellsworth, Miriam R.L. Petruck, Christopher R. Johnson, and Jan Scheffczyk. 2006. *FrameNet II: Extended Theory and Practice*. International Computer Science Institute, Berkeley, California. Distributed with the FrameNet data.

Natalia Skachkova and Ivana Kruijff-Korbayova. 2020. Reference in team communication for robot-assisted disaster response: An initial analysis. In *Proceedings of the Third Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 122–132, Barcelona, Spain (online). Association for Computational Linguistics.

Sang-Sang Tan and Jin-Cheon Na. 2019. Positional attention-based frame identification with BERT: A deep learning approach to target disambiguation and semantic frame selection. *arXiv preprint arXiv:1910.14549*.

Alaa Tharwat. 2020. Classification assessment methods. *Applied Computing and Informatics*.

Jeremy Trione, Frederic Bechet, Benoit Favre, and Alexis Nasr. 2015. Rapid FrameNet annotation of spoken conversation transcripts. In *Proceedings of the 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-11)*, London, UK. Association for Computational Linguistics.

Christian Willms, Constantin Houy, Jana-Rebecca Rehse, Peter Fettke, and Ivana Kruijff-Korbayová. 2019. Team communication processing and process analytics for supporting robot-assisted emergency response. In *2019 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*, pages 216–221. IEEE.

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. HuggingFace's Transformers: State-of-the-art natural language processing.

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*, pages 1247–1256.

Hongyan Zhao, Ru Li, Fei Duan, Zepeng Wu, and Shaoru Guo. 2018. TSABCNN: Two-stage attention-based convolutional neural network for frame identification. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, pages 289–301, Cham. Springer International Publishing.

A PAFIBERT: training and validation losses

Figure 2 shows the changes of training and validation losses with each training epoch of our re-implementation of the original PAFIBERT according to Tan and Na (2019). One can see that the model is powerful enough to memorize the training data by the end of the training, but, judging by the gap between the two curves, it has difficulties in generalizing and making confident predictions. Starting from the second epoch, the validation loss almost does not change, and it is also larger than the validation loss of the frame classifiers trained on TRADR, which can probably be attributed to the fact that FrameNet has many more classes than TRADR.

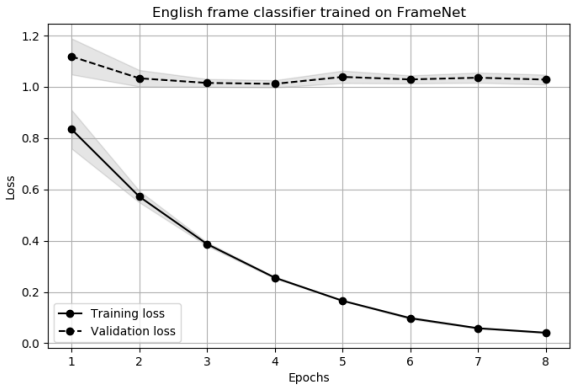

Figure 2: Traning and validation losses of the original PAFIBERT model

B German frame classifier with lexical features

Table 10 shows the performance of the German frame classifier with extra lexical features. Extending token embeddings with the corresponding POS tag embeddings seems to have a small positive effect on the IBA score, however, it is not significant according to the McNemar's test. Adding subword mask embeddings as well as using the combination of two extra features also does not seem to influence the performance of the classifier. Finally, we try extending token embeddings with POS tag embeddings together with the equal sampling from SALSA. However, the equal sampling, which earlier helped us achieve the IBA score of 90%, fails to provide the anticipated positive effect - the current score is only 84%, and the McNemar's test interprets the improvement as insignificant. We conclude that adding lexical features confuses the frame classifier, so that sampling looses its positive effect on the accuracy.

Model	PRE	REC	F1	$IBA_{0.1}$
POS tag	0.87	0.86	0.86	**0.85**
Subword mask	0.86	0.84	0.84	0.83
POS tag + subword mask	0.84	0.84	0.83	0.82
POS tag + equal sampling	0.87	0.85	0.85	0.84
Basic model (DE)	0.84	0.84	0.83	0.83

Table 10: Extra features: German frame classifier

C Team communication example

Table 11 shows one of the TRADR dialogues. The first column presents the speakers, the second - the utterances that sometimes also contain output from the *Transcriber* tool (Barras et al., 1998) (e.g., [ent=unk.skippable]), the third - the assigned frames depending on the targets (given in bold). According to our annotation approach, each utterance may contain several targets and thus evoke several frames. To make the dependencies between the targets and the corresponding frames clear, we annotated only one target-frame pair per row. This resulted in creating copies of the utterances containing several targets. They are given in *italics*. Most of the targets in the example dialogue are verbs which reflects our focus on various activities performed as part of the rescue mission.

The team communication example also illustrates two out of ten frames that we had to introduce during the annotation, as the FrameNet database (FrameNet, 2021) is not exhaustive, and it was not always possible to adapt the available frames to new phenomena. These two frames are *'Communication_by_protocol'* and *'Communication_response_message'*. They are domain-specific and are actually the most frequent in the whole TRADR corpus. Other eight frames that were introduced are rare. Table 12 contains the definitions and examples of all the new frames that we introduced. Frame elements are given in CAPITAL letters. We have not worked out their definitions yet. This is planned for future work.

The presented dialogue also has instances of the FrameNet frames that we adapted. Assigning the frame labels, sometimes it was impossible to follow the frame definitions given in the FrameNet database strictly. Considering that FrameNet is not exhaustive and that we were cautious to introduce too many new frames, we had to interpret

certain frame definitions in a more relaxed way. E.g., FrameNet defines the frame *'Existence'* as *"An Entity is declared to exist, generally irrespective of its position or even the possibility of its position being specified.* ⟨...⟩ *This frame is to be contrasted with Presence, which describes the existence of an Entity in a particular (and salient) spacio-temporal context, and which also entails the presence of an observer who can detect the existence of the Entity in that context."* We used *Existence* in a more straightforward way, namely with a reference to some news, findings, updates, etc. are present/available at a certain moment. Other adapted frames present in the dialogue are *Presence* and *Identity*. We do not present a full list of the adapted frames here, as there are quite many of them.

Notice that some utterances in the dialogue do not contain targets, as they are elliptical. In such cases we usually try to infer the missing elements, and assign the frame label that corresponds to the 'restored' utterance.

D Approaches to automatic frame assignment: a summary

Table 13 summarizes the characteristics of most of the frameworks mentioned in Section 3. The frameworks are given in chronological order, which helps illustrate the shift from the rule- and/or feature-based approaches to the embeddings-based ones, as well as the replacement of more 'traditional' classifiers with neural networks. The introduction of embeddings allowed to avoid manual feature engineering, and helped achieve better or comparable results with much less effort. However, the embeddings (even contextual ones, like ELMo or BERT) are still not able to deal with sense ambiguity effectively, which is one of the main problems in automatic frame assignment task.

The last row shows the performance of the frameworks. Those that have scores given were trained on the FrameNet corpus (versions may differ) and evaluated on one of the most commonly used Das test set (Das and Smith, 2011), which represents a part of FrameNet 1.5 data. Unfortunately, it is not always possible to compare the frameworks directly, as some researchers report F-score as a performance measure, others - accuracy. Five frameworks were evaluated on different test data, and we therefore omit their scores.

TL	Andreas, Andreas from Markus, **come in**.	Communication_by_protocol
OP	Yes, Andreas **come in**.	Communication_by_protocol
	⟨...⟩	
OP	Yes, for information, I am **ready** [EHM].	Activity_ready_state
	Shall I go ahead with my search command, or begin?	Desirable_event
	*Shall I **go ahead** with my search command, or begin?*	Activity_ongoing
	*Shall I go ahead with my search command, or **begin**?*	Activity_start
TL	Yes, **begin** immediately without possible – least possible time delay, to [EHM] have a higher chance for person rescue.	Activity_start
	*Yes, begin immediately without possible – least possible time delay, to [EHM] have a higher **chance** for person rescue.*	Likelihood
OP	Yes, **understood**, I begin with the search.	Communication_response_message
	*Yes, understood, I **begin** with the search.*	Activity_start
	⟨...⟩	
TL	Andreas from Markus, **come in**. [ent=unk.skippable]	Communication_by_protocol
OP	Yes, Andreas, **come in**.	Communication_by_protocol
TL	[ent=unk.skippable] **Are there** already any noteworthy findings? [ent=unk.skippable]	Existence
OP	**Negative**. No noteworthy findings. [ent=unk.skippable]	Communication_response_message
	Negative. No noteworthy findings. [ent=unk.skippable]	Existence
TL	Yes, **understood**.	Communication_response_message
	[ent=unk.skippable] Daniel, Daniel from Markus, **come in**.	Communication_by_protocol
	[ent=unk.skippable] Andreas from Markus, **come in**.	Communication_by_protocol
	⟨...⟩	
OP	Andreas, Markus from Andreas, **come in**.	Communication_by_protocol
TL	Andreas, **come in**.	Communication_by_protocol
OP	On first floor in the smoke **found** a barrel, green, labeled as environmentally hazardous material.	Locating
TL	Yeah, **can** you [unintelligible] whether anything is leaking?	Capability
	*Yeah, can you [unintelligible] whether anything is **leaking**?*	Fluidic_motion
OP	Yeah. It **is** a 200 liter barrel, whether anything is leaking I cannot currently tell.	Identity
	*Yeah. It is a 200 liter barrel, whether anything is **leaking** I cannot currently tell.*	Fluidic_motion
	*Yeah. It is a 200 liter barrel, whether anything is leaking I **can**not currently tell.*	Capability
	*Yeah. It is a 200 liter barrel, whether anything is leaking I cannot currently **tell**.*	Becoming_aware
TL	[EHM] Any thermal emission?	Presence
OP	No thermal emission.	Presence
TL	Okay. Priority on **continuing** person search.	Activity_ongoing
	Andreas from Markus, priority on **continuing** person search.	Activity_ongoing

Table 11: English TRADR dialogue annotated with semantic frames

Be_piece_of	
Inherits from:	Being_included
Definition:	A PART is considered to be a constituent of some entity described by the WHOLE. The relation is seen from the point of view of the PART.
Examples:	*I can also see [PART fragments] that **belong to** [WHOLE the building] [PART lying around here].*
Being_reasonable	
Inherits from:	Gradable_attributes
Definition:	Certain BEHAVIOR of PROTAGONIST is seen as practical and sensible.
Examples:	*As I can't see anything at the moment, it would definitely **make sense** if [PROTAGONIST you] [BEHAVIOR let the* UAV *guide you to some other points as soon as they've started again].*
Communication_by_protocol	
Inherits from:	Communication
Definition:	A COMMUNICATOR speaks to an ADDRESSEE using the phrases of special form (protocol) to establish/finish the conversation by radio.
Examples:	*[COMMUNICATOR Team leader] [ADDRESSEE for Tango].* *[COMMUNICATOR Team leader], **here** is [ADDRESSEE Tango].* *[COMMUNICATOR* UAV*] [ADDRESSEE to* UGV-1*] please **answer**.* *[COMMUNICATOR* UAV*] **speaking** [ADDRESSEE* IDI*].*
Communication_fragment	
Inherits from:	None
Definition:	An auxiliary frame which serves the purpose of marking conversational fillers and sequences with unclear meaning. The frame is characterized by conflation of target and FRAGMENT itself.
Examples:	*[FRAGMENT Also... I'm with... erm...]* *[FRAGMENT Eeh eeh my my my...]* *[*FRAGMENT *Whether a person or its... below at the bottom edge there's a...]*
Communication_response_message	
Inherits from:	Statement
Definition:	A COMMUNICATOR gives a short usually positive or negative reply to an ADDRESSEE's question or request. Sometimes a TOPIC is also mentioned.
Examples:	***Roger** [TOPIC that], [ADDRESSEE team leader].* ***Okay**.* ***Yes** [COMMUNICATOR by ground operator 1].*
Correction	
Inherits from:	Communication
Definition:	A COMMUNICATOR informs an ADDRESSEE that what the PATIENT has communicated is not right, true or suitable by providing the corrected version of the MESSAGE.
Examples:	*[COMMUNICATOR I] have to **correct** [PATIENT myself]: [MESSAGE* UGV-1*].*
Face_direction	
Inherits from:	State
Definition:	An ENTITY faces a particular DIRECTION.
Examples:	*For your information: [ENTITY it]'s **looking** [DIRECTION towards south].*
Lead	
Inherits from:	Cause_to_perceive
Definition:	An ENTITY leads in a particular DIRECTION or to some GOAL.
Examples:	*[ENTITY The stairwell] **leads** [DIRECTION upwards].* *There's smoke development at [ENTITY the first stairs] that **go** [DIRECTION upwards].*
Level_of_clarity	
Inherits from:	Gradable_attributes
Definition:	A DEGREE to which a REPRESENTATION is clear and detailed.
Examples:	*Yes, [REPRESENTATION the pictures] aren't [DEGREE very] **sharp**.*
Level_of_substance	
Inherits from:	Gradable_attributes
Definition:	A DEGREE of smoke in the air at some LOCATION.
Examples:	*It's actually [DEGREE quite] **smoky** [LOCATION* DNI*].*

Table 12: TRADR: new frames

Characteristic features	Framework											
	Erk (2005)	LTH (2007)	LUNA (2008)	SEMAFOR (2010)	Hermann et al. (2014)	SimpleFrameId (2017)	Open-Sesame (2017)	Yang & Mitchell (2017)	TSABCNN (2018)	PAFIBERT (2019)	Ribeiro et al. (2020)	Kalyanpur et al. (2020)
hand-crafted rules		✓	✓	✓	✓							
hand-crafted features	✓	✓	✓	✓	✓							
kernels			✓									
parsing		✓	✓	✓	✓	✓					✓	
embeddings					✓	✓	✓	✓	✓	✓	✓	✓
Naive Bayes classifier	✓											
SVM		✓	✓									
conditional log-linear model				✓	✓							
neural network						✓	✓	✓	✓	✓		✓
CRF							✓	✓				
clustering											✓	
graph structure											✓	
Frame assignment accuracy	n/a	n/a	n/a	82.97*	88.41	87.63	70.9*†	88.2	89.72	89.57	n/a	n/a

Table 13: Comparison of various frame-semantic parsing frameworks; scores marked with '*' stand for F-score (the authors do not report accuracy); 'n/a' means that the authors used a test set different from Das and Smith (2011); † stands for joint evaluation of frame assignment and argument identification

Corpus	English TRADR	German TRADR	FrameNet	SALSA
Domain	team communication in disaster response	team commumication in disaster response	mostly business, politics, economics related texts	newspaper texts
# inst.	2,930	2,813	199,508	35,236
# tokens	31,211	33,625	4,751,140	838,307
# classes	190 (177 occur in FrameNet)	152 (80 occur in SALSA)	1,014	880
Avg. sent. len.	8.68	6.99	22.92	21.78
# LUs	434 (280 occur in FrameNet)	-	8,333	-
% ambig. LUs wrt. # LUs	14.98	-	15.61	-
% ambig. LUs wrt. all inst.	52.90	-	34.99	-

Table 14: Corpora comparison

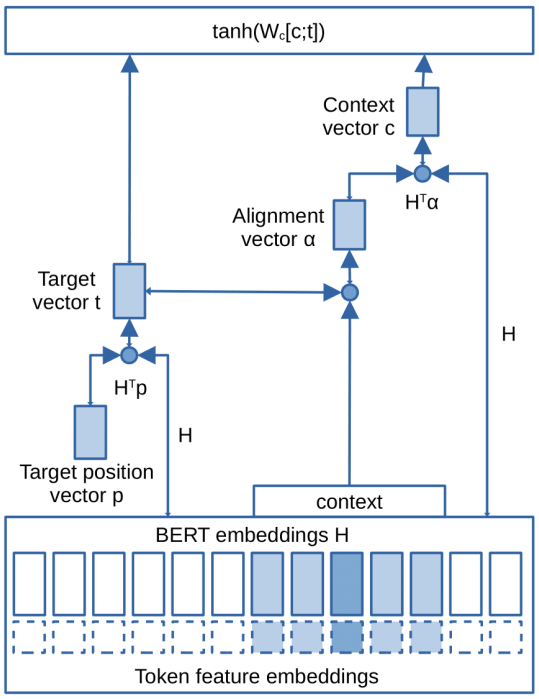

Figure 3: Adding lexical features (dashed borders) to PAFIBERT

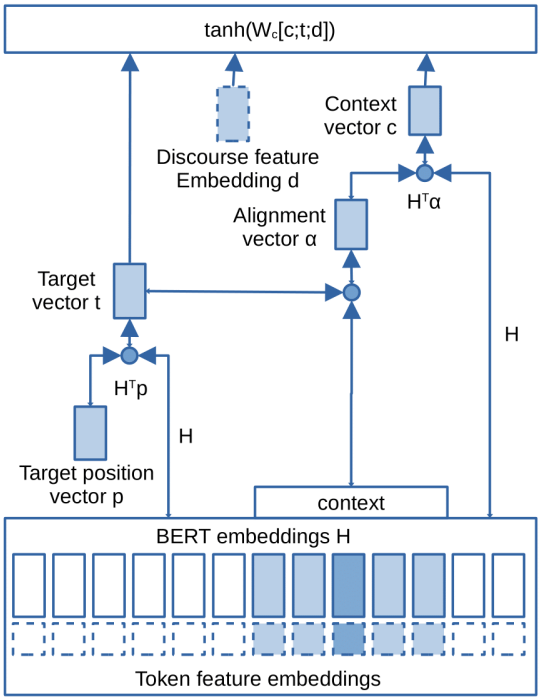

Figure 4: Adding lexical and discourse features (dashed borders) to PAFIBERT