Multimodal Frame Identification with Multilingual Evaluation

Teresa Botschen[§†], **Iryna Gurevych**[*§†], **Jan-Christoph Klie**[*†],
Hatem Mousselly-Sergieh[*†], **Stefan Roth**[*§‡]
§ Research Training Group AIPHES
† Ubiquitous Knowledge Processing (UKP) Lab
‡ Visual Inference Lab
Department of Computer Science, Technische Universität Darmstadt
www.{aiphes, ukp, visinf}.tu-darmstadt.de

Abstract

An essential step in FrameNet Semantic Role Labeling is the Frame Identification (FrameId) task, which aims at disambiguating a situation around a predicate. Whilst current FrameId methods rely on textual representations only, we hypothesize that FrameId can profit from a richer understanding of the situational context. Such contextual information can be obtained from common sense knowledge, which is more present in images than in text. In this paper, we extend a state-of-the-art FrameId system in order to effectively leverage multimodal representations. We conduct a comprehensive evaluation on the English FrameNet and its German counterpart SALSA. Our analysis shows that for the German data, textual representations are still competitive with multimodal ones. However on the English data, our multimodal FrameId approach outperforms its unimodal counterpart, setting a new state of the art. Its benefits are particularly apparent in dealing with ambiguous and rare instances, the main source of errors of current systems. For research purposes, we release *(a)* the implementation of our system, *(b)* our evaluation splits for SALSA 2.0, and *(c)* the embeddings for synsets and IMAGINED words.[1]

1 Introduction

FrameNet Semantic Role Labeling analyzes sentences with respect to frame-semantic structures based on FrameNet (Fillmore et al., 2003). Typically, this involves two steps: First, Frame Identification (FrameId), capturing the context around a predicate (*frame evoking element*) and assigning a frame, basically a word sense label for a prototypical situation, to it. Second, Role Labeling, i.e. identifying the participants (*fillers*) of the predicate and connecting them with predefined frame-specific role labels. FrameId is crucial to the success of Semantic Role Labeling as FrameId errors account for most wrong predictions in current systems (Hartmann et al., 2017). Consequently, improving FrameId is of major interest.

The main challenge and source of prediction errors of FrameId systems are ambiguous predicates, which can evoke several frames, e.g., the verb *sit* evokes the frame *Change_posture* in a context like *'a person is sitting back on a bench'*, while it evokes *Being_located* when *'a company is sitting in a city'*. Understanding the predicate context, and thereby the context of the situation (here, *'Who / what is sitting where?'*), is crucial to identifying the correct frame for ambiguous cases.

State-of-the-art FrameId systems model the situational context using pretrained distributed word embeddings (see Hermann et al., 2014). Hence, it is assumed that the context of the situation is explicitly expressed in words. However, language understanding involves implicit knowledge, which is not mentioned but still seems obvious to humans, e.g., *'people can sit back on a bench, but companies cannot'*, *'companies are in cities'*. Such implicit common sense knowledge is obvious enough to be rarely expressed in sentences, but is more likely to be present in images. Figure 1 takes the ambiguous predicate *sit* to illustrate

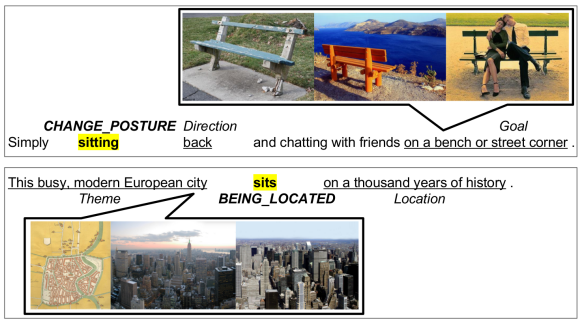

Figure 1: Example sentences demonstrating the potential benefit of images for ambiguous predicates.

[*] named alphabetically
[1] https://github.com/UKPLab/naacl18-multimodal-frame-identification

Proceedings of NAACL-HLT 2018, pages 1481–1491
New Orleans, Louisiana, June 1 - 6, 2018.

how images can provide access to implicit common sense knowledge crucial to FrameId.

When looking at the semantics of events, FrameId has commonalities with event prediction tasks. These aim at linking events and their participants to script knowledge and at predicting events in narrative chains. Ahrendt and Demberg (2016) argue that knowing about the participants helps to identify the event, which suggests the need for implicit context knowledge also for FrameId. This specifically applies to images, which can reflect properties of the participants of a situation in a inherently different way, see Fig. 1.

We analyze whether multimodal representations grounded in images can encode common sense knowledge to improve FrameId. To that end, we extend SimpleFrameId (Hartmann et al., 2017), a recent FrameId model based on distributed word embeddings, to the multimodal case and evaluate for English and German. Note that there is a general lack of evaluation of FrameId systems for languages other than English. This is problematic as they yield different challenges; German, for example, due to long distance dependencies. Also, word embeddings trained on different languages have different strengths in ambiguous words. We elaborate on insights from using different datasets by language.

Contributions. *(1)* We propose a pipeline and architecture of a FrameId system, extending state-of-the-art methods with the option of using implicit multimodal knowledge. It is flexible toward modality and language, reaches state-of-the-art accuracy on English FrameId data, clearly outperforming several baselines, and sets a new state of the art on German FrameId data. *(2)* We discuss properties of language and meaning with respect to implicit knowledge, as well as the potential of multimodal representations for FrameId. *(3)* We perform a detailed analysis of FrameId systems. First, we develop a new strong baseline. Second, we suggest novel evaluation metrics that are essential for assessing ambiguous and rare frame instances. We show our system's advantage over the strong baseline in this regard and by this improve upon the main source of errors. Third, we analyze gold annotated datasets for English and German showing their different strengths. Finally, we release the implementation of our system, our evaluation splits for SALSA 2.0, and the embeddings for synsets and IMAGINED words.

2 Related Work

2.1 Frame identification

State-of-the-art FrameId systems rely on pretrained word embeddings as input (Hermann et al., 2014). This proved to be helpful: those systems consistently outperform the previously leading FrameId system SEMAFOR (Das et al., 2014), which is based on a handcrafted set of features. The open source neural network-based FrameId system SimpleFrameId (Hartmann et al., 2017) is conceptually simple, yet yields competitive accuracy. Its input representation is a concatenation of the predicate's pretrained embedding and an embedding of the predicate context. The dimension-wise mean of the pretrained embeddings of all words in the sentence is taken as the context. In this work, we first aim at improving the representation of the predicate context using multimodal embeddings, and second at assessing the applicability to another language, namely German.

Common sense knowledge for language understanding. Situational background knowledge can be described in terms of frames (Fillmore, 1985) and scripts (Schank and Abelson, 2013). Ahrendt and Demberg (2016) report that knowing about a script's participants aids in predicting events linked to script knowledge. Transferring this insight to FrameId, we assume that a rich context representation helps to identify the sense of ambiguous predicates. Addressing ambiguous predicates where participants have different properties depending on the context, Feizabadi and Padó (2012) give some examples where the location plays a discriminating role as participant: motion verbs that have both a concrete motion sense and a more abstract sense in the cognitive domain, e.g., *struggle*, *lean*, *follow*.

Frame identification in German. Shalmaneser (Erk and Pado, 2006) is a toolbox for semantic role assignment on FrameNet schemata of English and German (integrated into the SALSA project for German). Shalmaneser uses a Naive Bayes classifier to identify frames, together with features for a bag-of-word context with a window over sentences, bigrams, and trigrams of the target word and dependency annotations. They report an F1 of 75.1 % on FrameNet 1.2 and 60 % on SALSA 1.0. These scores are difficult to compare against more recent work as the evaluation uses older versions of datasets and custom splits. Shalmaneser

requires software dependencies that are not available anymore, hindering application to new data. To the best of our knowledge, there is no FrameId system evaluated on SALSA 2.0.

Johannsen et al. (2015) present a simple, but weak translation baseline for cross-lingual FrameId. A SEMAFOR-based system is trained on English FrameNet and tested on German Wikipedia sentences, translated word-by-word to English. This translation baseline reaches an F1 score of 8.5 % on the German sentences when translated to English. The performance of this weak translation baseline is worse than that of another simple baseline: a 'most frequent sense baseline' – computing majority votes for German (and many other languages) – reaches an F1 score of 53.0 % on the German sentences. This shows that pure translation does not help with FrameId and, furthermore, indicates a large room for improvement for FrameId in languages other than English.

2.2 Multimodal representation learning

There is a growing interest in Natural Language Processing for enriching traditional approaches with knowledge from the visual domain, as images capture qualitatively different information compared to text. Regarding FrameId, to the best of our knowledge, multimodal approaches have not yet been investigated. For other tasks, multimodal approaches based on pretrained embeddings are reported to be superior to unimodal approaches. Textual embeddings have been enriched with information from the visual domain, e.g., for Metaphor Identification (Shutova et al., 2016), Question Answering (Wu et al., 2017), and Word Pair Similarity (Collell et al., 2017). The latter presents a simple, but effective way of extending textual embeddings with so-called multimodal IMAGINED embeddings by a learned mapping from language to vision. We apply the IMAGINED method to our problem.

In this work, we aim to uncover whether representations that are grounded in images can help to improve the accuracy of FrameId. Our application case of FrameId is more complex than a comparison on the word-pair level as it considers a whole sentence in order to identify the predicate's frame. However, we see a potential for multimodal IMAGINED embeddings to help: their mapping from text to multimodal representations is learned from images for nouns. Such nouns, in turn, are candidates for role fillers of predicates. In order to identify the correct sense of an ambiguous predicate, it could help to enrich the representation of the context situation with multimodal embeddings for the entities that are linked by the predicate.

3 Our Multimodal FrameId Model

Our system builds upon the SimpleFrameId (Hartmann et al., 2017) system for English FrameId based on textual word embeddings. We extend it to multimodal and multilingual use cases; see Fig. 2 for a sketch of the system pipeline. Same as SimpleFrameId, our system is based on pretrained embeddings to build the input representation out of the predicate context and the predicate itself.

However, different to SimpleFrameId, our representation of the predicate context is multimodal: beyond textual embeddings we also use IMAGINED and visual embeddings. More precisely, we concatenate all unimodal representations of the predicate context, which in turn are the unimodal mean embeddings of all words in the sentence. We use concatenation for fusing the different embeddings as it is the simplest yet successful fusion approach (Bruni et al., 2014; Kiela and Bottou, 2014). The input representation is processed by a two-layer Multilayer Perceptron (MLP, Rosenblatt, 1958), where we adapt the number of hidden nodes to the increased input size and apply dropout to all hidden layers to prevent overfitting (Srivastava et al., 2014). Each node in the output layer corresponds to one frame-label class. We use rectified linear units (Nair and Hinton, 2010) as activation function for the hidden layers, and a soft-

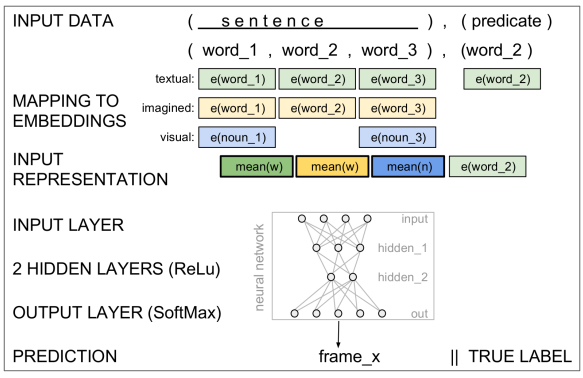

Figure 2: Sketch of the pipeline. *(1)* Data: sentence with predicate. *(2)* Mapping: words to embeddings. *(3)* Representation: concatenation of modality-specific means. *(4)* Classifier: neural network predicting frame.

max for the output layer yielding a multinomial distribution over frames. We take its $\arg\max$ as the final prediction at test time. Optionally, filtering based on the lexicon can be performed on the predicted probabilities for each frame label. The development set was used to determine the architecture and hyperparameters, see Sec. 6.

Majority baselines. We propose a new strong baseline based on a combination of two existing ones. These are: first, the most-frequent-sense baseline using the data majority (Data Baseline) to determine the most frequent frame for a predicate; second, the baseline introduced by Hartmann et al. (2017) using a lexicon (Lexicon Baseline) to consider the data counts of the Data Baseline only for those frames available for a predicate. We propose to combine them into a Data-Lexicon Baseline, which uses the lexicon for unambiguous predicates and for ambiguous ones it uses the data majority. This way, we trust the lexicon for unambiguous predicates but not for ambiguous ones, there we rather consider the data majority. Comparing a system to these baselines helps to see whether it just memorizes the data majority or the lexicon, or actually captures more.

All majority baselines strongly outperform the weak translation baseline of Johannsen et al. (2015) when training the system on English data and evaluating it on German data.

4 Preparation of Input Embeddings

Textual embeddings for words. We use the 300-dimensional GloVe embeddings (Pennington et al., 2014) for English, and the 100-dimensional embeddings of Reimers et al. (2014) for German. GloVe and Reimers have been trained on the Wikipedia of their targeted language and on additional newswire text to cover more domains, resulting in similarly low out-of-vocabulary scores.

Visual embeddings for synsets. We obtain visual embeddings for WordNet synsets (Fellbaum, 1998; , Ed.): we apply the pretrained VGG-m-128 Convolutional Neural Network model (Chatfield et al., 2014) to images for synsets from ImageNet (Deng et al., 2009), we extract the 128-dimensional activation of the last layer (before the softmax) and then we L_2-normalize it. We use the images of the WN9-IMG dataset (Xie et al., 2017), which links WordNet synsets to a collection of ten ImageNet images. We average the embeddings of all images corresponding to a synset, leading to a vocabulary size of 6555 synsets. All synsets in WN9-IMG are part of triples of the form entity-relation-entity, i.e. synset-relation-synset. Such synset entities that are participants of relations with other synset entities are candidates for incorporating the role fillers for predicates and, therefore, may help to find the correct frame for a predicate (see Sec. 5 for details about sense-disambiguation.)

Linguistic embeddings for synsets. We obtain 300-dimensional linguistic synset embeddings: we apply the AutoExtend approach (Rothe and Schütze, 2015) to GloVe embeddings and produce synset embeddings for all synsets having at least one synset lemma in the GloVe embeddings. This leads to a synset vocabulary size of 79 141. Linguistic synset embeddings are based on textual word embeddings and the synset information known by the knowledge base WordNet, thus they complement the visual synset embeddings.

IMAGINED embeddings for words. We use the IMAGINED method (Collell et al., 2017) for learning a mapping function: it maps from the word embedding space to the visual embedding space given those words that occur in both pretrained embedding spaces (7220 for English and 7739 for German). To obtain the English synset lemmas, we extract all lemmas of a synset and keep those that are nouns. We automatically translate English nouns to German nouns using the Google Translate API to obtain the corresponding German synset lemmas. The IMAGINED method is promising for cases where one embedding space (here, the textual one) has many instances without correspondence in the other embeddings space (here, the visual one), but the user still aims at obtaining instances of the first in the second space. We aim to obtain visual correspondences for the textual embeddings in order to incorporate regularities from images into our system. The mapping is a nonlinear transformation using a simple neural network. The objective is to minimize the cosine distance between each mapped representation of a word and the corresponding visual representation. Finally, a multimodal representation for any word can be obtained by applying this mapping to the word embedding.

5 Data and Preparation of Splits

English FrameId: Berkeley FrameNet. The Berkeley FrameNet (Baker et al., 1998; Ruppenhofer et al., 2016) is an ongoing project for building a large lexical resource for English with expert annotations based on frame semantics (Fillmore, 1976). It consists of two parts, a manually created lexicon that maps predicates to the frames they can evoke, and fully annotated texts (fulltext). The mapping can be used to facilitate the frame identification for a predicate in a sentence, e.g., a sentence in the fulltext corpus. Table 1 contains the lexicon statistics, Table 2 (top left) the dataset statistics. In this work, we use FrameNet 1.5 to ensure comparability with the previous state of the art, with the common evaluation split for FrameId systems introduced by Das and Smith (2011) (with the development split of Hermann et al., 2014). Due to having a single annotation as consent of experts, it is hard to estimate a performance bound of a single human for the fulltext annotation.

German FrameId: SALSA. The SALSA project (Burchardt et al., 2006; Rehbein et al., 2012) is a completed annotation project, which serves as the German counterpart to FrameNet. Its annotations are based on FrameNet up to version 1.2. SALSA adds proto-frames to properly annotate senses that are not covered by the English FrameNet. For a more detailed description of differences between FrameNet and SALSA, see Ellsworth et al. (2004); Burchardt et al. (2009). SALSA also provides a lexicon (see Table 1 for statistics) and fully annotated texts. There are two releases of SALSA: 1.0 (Burchardt et al., 2006) used for Shalmaneser (Erk and Pado, 2006) (cf. Sec. 2.1), and the final release 2.0 (Rehbein et al., 2012), which contains more annotations and adds nouns as predicates. We use the final release.

SALSA has no standard evaluation split; Erk and Pado (2006) used an undocumented random split. Also, it is not possible to follow the splitting method of Das and Smith (2011), as SALSA project distributions do not map to documents. We suggest splitting based on sentences, i.e. all annotations of a sentence are in the same set to avoid mixing training and test sets. We assign sentences to 100 buckets based on their IDs and create a 70/15/15 split for training, development, and test sets based on the bucket order. This procedure allows future work to be evaluated on the same data. Table 2 (bottom left) shows the dataset statistics.

Synsets in FrameNet and SALSA. To prepare the datasets for working with the synset embeddings, we sense-disambiguate all sentences using the API of BabelNet (Navigli and Ponzetto, 2010), which returns multilingual synsets. We thus depend on the state-of-the-art accuracy of BabelNet (Navigli and Ponzetto, 2012) when using synset embeddings on sense-disambiguated sentences. However, this dependence does not hold when applying IMAGINED embeddings to sentences, as the mapping from words to IMAGINED embeddings does not need any synsets labeled in the sentences. After sense-disambiguation some sentences do not contain any synset available in our synset embeddings. The statistics of those sentences that have at least one synset embedding (visual or linguistic AutoExtend) is given in Table 2 (right).

6 Experimental Setup

We contrast our system's performance for context representations based on unimodal (textual) versus multimodal (textual and visual) embeddings. Also, we compare English against German data. We run the prediction ten times to reduce noise in

lexicon	**frames**	**LUs**	**avg**(fr/pred)	**%amb.pred.**
FrameNet	1020	11 942	1.26	17.32
SALSA	1023	1827	2.82	57.56

Table 1: Lexicon statistics for FrameNet 1.5 and for SALSA 2.0: the total number of distinct **frames** and lexical units **LUs** (distinct predicate-frame combinations), the number of frames a predicate can evoke on average **avg**, and the % of **amb**iguous **pred**icates.

		sentences	**frames**	reduced sentences	
				syns-Vis	syns-AutoExt
FrameN	train	2819	15 406	1310	2714
	dev	707	4593	320	701
	test	2420	4546	913	2318
SALSA	train	16 852	26 081	4707	16 736
	dev	3561	5533	1063	3540
	test	3605	5660	1032	3570

Table 2: Dataset statistics for FrameNet 1.5 fulltext with Das split and for SALSA 2.0 with our split: number of **sentences** and **frames** (as used in our experiments). Right half (only used in further investigations): number of sentences when reduced to only those having synsets in the visual and in the linguistic AutoExtend embeddings.

the evaluation (cf. Reimers and Gurevych, 2017) and report the mean for each metric.

Use of lexicon. We evaluate our system in two settings: with and without lexicon, as suggested by Hartmann et al. (2017). In the with-lexicon setting, the lexicon is used to reduce the choice of frames for a predicate to only those listed in the lexicon. If the predicate is not in the lexicon, it corresponds to the without-lexicon setting, where the choice has to be done amongst all frames.

Evaluation metrics. FrameId systems are usually compared in terms of *accuracy*, which we adopt for comparability. As a multiclass classification problem, FrameId has to cope with a strong variation in the annotation frequency of frame classes. Minority classes are frames that occur only rarely; majority classes occur frequently. Note that the accuracy is biased toward majority classes, explaining the success of majority baselines on imbalanced datasets such as FrameNet.

Alternatively, the *F1 score* is sometimes reported as it takes a complementary perspective. The F-measure is the harmonic mean of precision and recall, measuring exactness and completeness of a model, respectively. In previous work, micro-averaging is used to compute F1 scores. Yet, similar to the accuracy, micro-averaging introduces a bias toward majority classes. We compute *F1-macro* instead, for which precision and recall are computed for each class and averaged afterwards, giving equal weight to all classes.

Taken together, this yields scores that underestimate (F1-macro) and overestimate (average accuracy) on imbalanced datasets. Previous work just used the overestimate such that a comparison is possible in terms of accuracy in the with-lexicon setting. We suggest to use F1-macro additionally to analyze rare, but interesting classes. Thus, a comparison within our work is possible for both aspects, giving a more detailed picture. Note that previous work reports one score whilst we report the mean score of ten runs.

Hyperparameters. We identified the best hyperparameters for the English and German data based on the respective development sets.[2] The Multilayer Perceptron architecture performed consistently better than a more complex Gated Recurrent Unit model (Cho et al., 2014). We found that more than two hidden layers did not bring any improvement over two layers; using dropout on the hidden layers helped to increase the accuracy. Among the various input representations, a concatenation of the representations of context and predicate was the best amongst others, including dependencies, lexicon indicators, and part-of-speech tags. Training is done using Nesterov-accelerated Adam (Nadam, Dozat, 2016) with default parameters. A batch size of 128 is used. Learning stops if the development accuracy has not improved for four epochs, and the learning rate is reduced by factor of two if there has not been any improvement for two epochs.

[2]Differences in hyperparameters to SimpleFrameId: 'nadam' as optimizer instead of 'adagrad', dropout on hidden layers and early stopping to regularize training. Different number of hidden units, optimized by grid search.

7 Results

First, we report our results on English data (see Table 3, top) and then, we compare against German data (see Table 3, bottom).

7.1 English FrameNet data

Baseline. Our new strong Data-Lexicon Baseline reaches a considerable accuracy of 86.32 %, which is hard to beat by trained models. Even the most recent state of the art only beats it by about two points: 88.41 % (Hermann et al., 2014). However, the accuracy of the baseline drops for ambiguous predicates (69.73 %) and the F1-macro score reveals its weakness toward minority classes (drop from 64.54 % to 37.42 %).

Unimodal. Our unimodal system trained and evaluated on English data slightly exceeds the accuracy of the previous state of the art (88.66 % on average versus 88.41 % for Hermann et al., 2014); our best run's accuracy is 89.35 %. Especially on ambiguous predicates, i.e. the difficult and therefore interesting cases, our average accuracy surpasses that of previous work by more than one point (the best run by almost three points). Considering the proposed F1-macro score for an assessment of the performance on minority classes and ambiguous predicates reveals our main improvement: Our system substantially outperforms the strong Data-Lexicon Baseline, demonstrating that our system differs from memorizing majorities and actually improves minority cases.

Multimodal. From a range of multimodal context representations as extensions to our system,

	model	with lexicon acc	with lexicon acc_amb	with lexicon F1-m	with lexicon F1-m_amb	without lexicon acc	without lexicon acc_amb	without lexicon F1-m	without lexicon F1-m_amb
FrameNet	Data Baseline	79.06	69.73	33.00	37.42	79.06	69.73	33.00	37.42
	Lexicon Baseline	79.89	55.52	65.61	30.95	–	–	–	–
	Data-Lexicon Baseline	86.32	69.73	64.54	37.42	–	–	–	–
	Hermann et al. (2014)	88.41	73.10	–	–	–	–	–	–
	Hartmann et al. (2017)	87.63	73.80	–	–	77.49	–	–	–
	our_uni	88.66	74.92	76.65	53.86	79.96	71.70	57.07	47.40
	our_mm (im, synsV)	**88.82**	**75.28**	**76.77**	**54.80**	**81.21**	**72.51**	**57.81**	**49.38**
SALSA	Data Baseline	77.00	70.51	37.40	28.87	77.00	70.51	37.40	28.87
	Lexicon Baseline	61.57	52.5	19.36	15.68	–	–	–	–
	Data-Lexicon Baseline	77.16	70.51	38.48	28.87	–	–	–	–
	our_uni	**80.76**	**75.59**	**48.42**	**41.38**	**80.59**	**75.52**	**47.64**	**41.17**
	our_mm (im)	80.71	75.58	48.29	41.19	80.51	75.51	47.36	40.93

Table 3: FrameId results (in %) on English (upper) and German (lower) with and without using the lexicon. Reported are **acc**uracy and **F1-m**acro, both also for **amb**iguous predicates (mean scores over ten runs). Models: *(a)* Data, Lexicon, and Data-Lexicon Baselines. *(b)* Previous models for English. *(c)* Ours: unimodal **our**-uni, multimodal on top of our-uni – **our**-mm – with IMAGINED embeddings (and synset visual embeddings for English). Best results highlighted in bold. The best run's results for English were:
our_uni: **acc**: 89.35 ; **acc_amb**: 76.45 ; **F1-m**: 76.95 ; **F1-m_amb**: 54.02 (with lexicon)
our_mm (im, synsV): **acc**: 89.09 ; **acc_amb**: 75.86 ; **F1-m**: 78.17 ; **F1-m_amb**: 57.48 (with lexicon)

the most helpful one is the concatenation of IMAGINED embeddings and visual synset embeddings: it outperforms the unimodal approach slightly in all measurements. We observe that the improvements are more pronounced for difficult cases, such as for rare and ambiguous cases (one point improvement in F1-macro), as well as in the absence of a lexicon (up to two points improvement).

Significance tests. We conduct a single sample t-test to judge the difference between previous state-of-the-art accuracy (Hermann et al., 2014) and our unimodal approach. The null hypothesis (expected value of our sample of ten accuracy scores equals previous state-of-the-art accuracy) is rejected at a significance level of $\alpha = 0.05$ ($p = 0.0318$). In conclusion, even our unimodal approach outperforms prior state of the art in terms of accuracy.

To judge the difference between our unimodal and our multimodal approach, we conduct a t-test for the means of the two independent samples. The null hypothesis states identical expected values for our two samples of ten accuracy scores. Regarding the setting with lexicon, the null hypothesis cannot be rejected at a significance level of $\alpha = 0.05$ ($p = 0.2181$). However, concerning accuracy scores without using the lexicon, the null hypothesis is rejected at a significance level of $\alpha = 0.05$ ($p < 0.0001$). In conclusion, the multimodal approach has a slight overall advantage and, interestingly, has a considerable advantage over the unimodal one when confronted with a more difficult setting of not using the lexicon.

7.2 German SALSA versus English data

German results. Our system evaluated on German data sets a new state of the art on this corpus with 80.76 % accuracy, outperforming the baselines (77.16 %; no other system evaluated on this dataset). The difference in F1-macro between the majority baselines and our system is smaller than for the English FrameNet. This indicates that the majorities learned from data are more powerful in the German case with SALSA than in the English case, when comparing against our system. Multimodal context representations cannot show an improvement for SALSA with this general dataset.

Lexicon. We report results achieved without the lexicon to evaluate independently of its quality (Hartmann et al., 2017). On English data, our systems outperforms Hartmann et al. (2017) by more than two points in accuracy and we achieve a large improvement over the Data Baseline. Comparing the F1-macro with and without lexicon, it can be seen that the additional information stored in the lexicon strongly increases the score by about 20 points for English data. For German data, the increase of F1-macro with lexicon versus without is small (one point).

8 Discussion

8.1 English data

Insights from the baseline. Many indicators point to our approach not just learning the data majority: our trained models have better F1-macro and especially much higher ambiguous F1-macro scores with lexicon. This clearly suggests that our system is capable of acquiring more expressiveness than the baselines do by counting majorities.

Impact of multimodal representations. Multimodal context representations improve results compared to unimodal ones. It helps to incorporate visual common sense knowledge about the situation's participants. Referring back to our example of the ambiguous predicate *sit*, the multimodal approach is able to transfer the knowledge to the test sentence *'Al-Anbar in general, and Ramadi in particular, are sat with the Americans in Jordan.'* by correctly identifying the frame *Being_located* whilst the unimodal approach fails with predicting *Change_posture*. The increase in performance when adding information from visual synset embeddings is not simply due to higher dimensionality of the embedding space. To verify, we further investigate extending the unimodal system with random word embeddings. This leads to a drop in performance compared to using just the unimodal representations or using these in combination with the proposed multimodal embeddings, especially in the setting without lexicon. Interestingly, replacing visual synset embeddings with linguistic synset embeddings (AutoExtend by Rothe and Schütze (2015), see Sec. 4) in further investigations also showed that visual embeddings yield better performance. This points out the potential for incorporating even more image evidence to extend our approach.

8.2 German versus English data

Difficulties for German data. The impact of multimodal context representations is more difficult to interpret for the German dataset. The fact that they have not helped here may be due to mismatches when translating the English nouns of a synset to German in order to train the IMAGINED embeddings. Here, we see room for future work to improve on simple translation by sense-based translations. In SALSA, a smaller portion of sentences has at least one synset embedding, see Table 2. For further investigations, we reduced the dataset to only sentences actually containing a synset embedding. Then, minor improvements of the multimodal approach were visible for SALSA. This points out that a dataset containing more words linking to implicit knowledge in images (visual synset embeddings) can profit more from visual and IMAGINED embeddings.

Impact of lexicon: English versus German. Even if both lexica approximately define the same number of frames (see Table 1), the number of defined lexical units (distinct predicate-frame combinations) in SALSA is smaller. This leads to a lexicon that is a magnitude smaller than the FrameNet lexicon. Thus, the initial situation for the German case is more difficult. The impact of the lexicon for SALSA is smaller than for FrameNet (best visible in the increase of F1-macro with using the lexicon compared to without), which can be explained by the larger percentage of ambiguous predicates (especially evoking proto-frames) and the smaller size of the lexicon. The evaluation on two different languages highlights the impact of an elaborate, manually created lexicon: it boosts the performance on frame classes that are less present in the training data. English FrameId benefits from the large high-quality lexicon, whereas German FrameId currently lacks a high-quality lexicon that is large enough to benefit the FrameId task.

Dataset properties: English versus German. To better understand the influence of the dataset on the prediction errors, we further analyze the errors of our approach (see Table 4) following Palmer

		with lexicon				without lexicon			
	model	**correct**	**e**_uns	**e**_unsLab	**e**_n	**correct**	**e**_uns	**e**_unsLab	**e**_n
FrameN	**our**_uni	89.35	0.40	3.04	7.22	80.36	1.32	7.68	10.65
	our_mm (im, synsV)	89.79	0.58	3.55	6.08	80.63	1.91	8.50	8.96
SALSA	**our**_uni	80.99	0.49	0.97	17.54	80.80	0.49	1.10	17.61
	our_mm (im)	81.24	1.94	1.88	14.94	80.96	1.94	2.05	15.05

Table 4: Error analysis of best uni- and multimodal systems. **correct**, **e**rrors: unseen, unseen label and normal.

and Sporleder (2010). A wrong prediction can either be a normal classification error, or it can be the result of an instance that was unseen at training time, which means that the error is due to the training set. The instance can either be completely unseen or unseen with the target label. We observe that FrameNet has larger issues with unseen data compared to SALSA, especially data that was unseen with one specific label but seen with another label. This is due to the uneven split of the documents in FrameNet, leading to data from different source documents and domains in the training and test split. SALSA does not suffer from this problem as much since the split was performed differently. It would be worth considering the same splitting method for FrameNet.

8.3 Future work

As stated previously, FrameId has commonalities with event prediction. Since identifying frames is only one way of capturing events, our approach is transferable to other schemes of event prediction and visual knowledge about participants of situations should be beneficial there, too. It would be interesting to evaluate the multimodal architecture on other predicate-argument frameworks, e.g., script knowledge or VerbNet style Semantic Role Labeling. In particular the exploration our findings on visual contributions to FrameId in the context of further event prediction tasks forms an interesting next step.

More precisely, future work should consider using implicit knowledge not only from images of the participants of the situation, but also from the entire scene in order to directly capture relations between the participants. This could provide access to a more holistic understanding of the scene. The following visual tasks with accompanying datasets could serve as a starting point: (a) visual Verb Sense Disambiguation with the VerSe dataset (Gella et al., 2016) and (b) visual SRL with several datasets, e.g., imSitu (Yatskar et al., 2016) (linked to FrameNet), V-COCO (Gupta and Malik, 2015) (verbs linked to COCO), VVN (Ronchi and Perona, 2015) (visual VerbNet) or even SRL grounded in video clips for the cooking-domain (Yang et al., 2016) and visual Situation Recognition (Mallya and Lazebnik, 2017). Such datasets could be used for extracting visual embeddings for verbs or even complex situations in order to improve the visual component in the embeddings for our FrameId system. Vice versa: visual tasks could profit from multimodal approaches (Baltrušaitis et al., 2017) in a similar sense as our textual task, FrameId, profits from additional information encoded in further modalities. Moreover, visual SRL might profit from our multimodal FrameId system to a similar extend as any FrameNet SRL task profits from correctly identified frames (Hartmann et al., 2017).

Regarding the combination of embeddings from different modalities, we suggest to experiment with different fusion strategies complementing the middle fusion (concatenation) and the mapping (IMAGINED method). This could be a late fusion at decision level operating like an ensemble.

9 Conclusion

In this work, we investigated multimodal representations for Frame Identification (FrameId) by incorporating implicit knowledge, which is better reflected in the visual domain. We presented a flexible FrameId system that is independent of modality and language in its architecture. With this flexibility it is possible to include textual and visual knowledge and to evaluate on gold data in different languages. We created multimodal representations from textual and visual domains and showed that for English FrameNet data, enriching the textual representations with multimodal ones improves the accuracy toward a new state of the art. For German SALSA data, we set a new state of the art with textual representations only and discuss why incorporating multimodal information is more difficult. For both datasets, our system is particularly strong with respect to ambiguous and rare classes, considerably outperforming our new Data-Lexicon Baseline and thus addressing a key challenge in FrameId.

Acknowledgments

This work has been supported by the DFG-funded research training group "Adaptive Preparation of Information form Heterogeneous Sources" (AIPHES, GRK 1994/1). We also acknowledge the useful comments of the anonymous reviewers.

References

Simon Ahrendt and Vera Demberg. 2016. Improving event prediction by representing script participants. In *Proceedings of NAACL-HLT*, pages 546–551, San Diego, USA.

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*, pages 86–90, Stroudsburg, PA, USA.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2017. Multimodal machine learning: A survey and taxonomy. *arXiv preprint arXiv:1705.09406*.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49(2014):1–47.

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 *Proceedings of LREC*, Genoa, Italy.

Aljoscha Burchardt, Katrin Erk, Anette Frank, Andrea Kowalski, Sebastian Padó, and Manfred Pinkal. 2009. Using FrameNet for the semantic analysis of German: Annotation, representation, and automation. In *Multilingual FrameNets in Computational Lexicography*, pages 209–244. Mouton de Gruyter, New York City, USA.

Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Return of the devil in the details: Delving deep into convolutional nets. In *Proceedings of the British Machine Vision Conference (BMVC 2015)*, Nottingham, Great Britain.

Kyunghyun Cho, Bart van Merriënboer, Gülçehre Çağlar, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar.

Guillem Collell, Ted Zhang, and Marie-Francine Moens. 2017. Imagined visual representations as multimodal embeddings. In *Thirty-First AAAI Conference on Artificial Intelligence*, pages 4378–4384, San Francisco, USA.

Dipanjan Das, Desai Chen, André F. T. Martins, Nathan Schneider, and Noah A. Smith. 2014. Frame-semantic parsing. *Computational Linguistics*, 40:1:9–56.

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 - Volume 1*, pages 1435–1444, Stroudsburg, PA, USA.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, Miami, USA.

Timothy Dozat. 2016. Incorporating Nesterov momentum into Adam. In *International Conference on Representation Learning (ICLR): Posters*, pages 1–7, San Juan, Puerto Rico.

Christiane Fellbaum (Ed.). 1998. *WordNet: An Electronic Lexical Database*. The MIT Press, Cambridge, MA ; London.

Michael Ellsworth, Katrin Erk, Paul Kingsbury, and Sebastian Padó. 2004. PropBank, SALSA, and FrameNet: How design determines product. In *Proceedings of the LREC 2004 Workshop on Building Lexical Resources from Semantically Annotated Corpora*, Lisbon, Portugal.

Katrin Erk and Sebastian Pado. 2006. Shalmaneser - A flexible toolbox for semantic role assignment. In *Proceedings of LREC*, Genoa, Italy.

Parvin Sadat Feizabadi and Sebastian Padó. 2012. Automatic identification of motion verbs in WordNet and FrameNet. In *KONVENS*, pages 70–79, Vienna, Austria.

Christiane Fellbaum, editor. 1998. *WordNet: An electronic lexical database*. MIT Press, Cambridge, USA.

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

Charles J Fillmore. 1985. Frames and the semantics of understanding. *Quaderni di semantica*, 6(2):222–254.

Charles J. Fillmore, Christopher R. Johnson, and Miriam R.L. Petruck. 2003. Background to FrameNet. *International Journal of Lexicography*, 16(3):235–250.

Spandana Gella, Mirella Lapata, and Frank Keller. 2016. Unsupervised visual sense disambiguation for verbs using multimodal embeddings. In *Proceedings of NAACL-HLT*, pages 182–192, San Diego, USA.

Saurabh Gupta and Jitendra Malik. 2015. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*.

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*, pages 471–482, Valencia, Spain.

Karl Moritz Hermann, Dipanjan Das, Jason Weston, and Kuzman Ganchev. 2014. Semantic frame identification with distributed word representations. In *Proceedings of the ACL*, pages 1448–1458, Baltimore, USA.

Anders Johannsen, Héctor Martínez Alonso, and Anders Søgaard. 2015. Any-language frame-semantic parsing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, page 20622066, Lisbon, Portugal.

Douwe Kiela and Léon Bottou. 2014. Learning image embeddings using convolutional neural networks for improved multi-modal semantics. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 36–45, Doha, Qatar.

Arun Mallya and Svetlana Lazebnik. 2017. Recurrent models for situation recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 455–463, Venice, Italy.

Vinod Nair and Geoffrey E. Hinton. 2010. Rectified linear units improve restricted Boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning*, pages 807–814, Haifa, Israel.

Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, Uppsala, Sweden.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Alexis Palmer and Caroline Sporleder. 2010. Evaluating FrameNet-style semantic parsing: The role of coverage gaps in FrameNet. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 928–936, Beijing, China.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Empirical Methods in Natural Language*, pages 1532–1543, Doha, Qatar.

Ines Rehbein, Josef Ruppenhofer, Caroline Sporleder, and Manfred Pinkal. 2012. Adding nominal spice to SALSA - Frame-semantic annotation of German nouns and verbs. In *Proceedings of KONVENS 2012*, pages 89–97, Vienna, Austria.

Nils Reimers, Judith Eckle-Kohler, Carsten Schnober, Jungi Kim, and Iryna Gurevych. 2014. GermEval-2014: Nested named entity recognition with neural networks. In *Workshop Proceedings of the 12th Edition of the KONVENS Conference*, pages 117–120, Hildesheim, Germany.

Nils Reimers and Iryna Gurevych. 2017. Reporting score distributions makes a difference: Performance study of LSTM-networks for sequence tagging. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 338–348, Copenhagen, Denmark.

Matteo Ruggero Ronchi and Pietro Perona. 2015. Describing common human visual actions in images. In *Proceedings of the British Machine Vision Conference (BMVC 2015)*, Swansea, Wales.

Frank Rosenblatt. 1958. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, pages 65–386.

Sascha Rothe and Hinrich Schütze. 2015. AutoExtend: Extending word embeddings to embeddings for synsets and lexemes. In *Proceedings of the ACL*, Beijing, China.

Josef Ruppenhofer, Michael Ellsworth, Miriam R.L. Petruck, Christopher R. Johnson, and Jan Scheffczyk. 2016. *FrameNet II: Extended Theory and Practice*, revised november 1, 2016 edition. International Computer Science Institute, Berkeley, USA.

Roger C Schank and Robert P Abelson. 2013. *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Psychology Press.

Ekaterina Shutova, Douwe Kiela, and Jean Maillard. 2016. Black holes and white rabbits: Metaphor identification with visual features. In *Proceedings of NAACL-HLT*, pages 160–170, San Diego, USA.

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

Qi Wu, Damien Teney, Peng Wang, Chunhua Shen, Anthony Dick, and Anton van den Hengel. 2017. Visual question answering: A survey of methods and datasets. *Computer Vision and Image Understanding*, 163:21–40.

Ruobing Xie, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2017. Image-embodied knowledge representation learning. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pages 3140–3146, Melbourne, Australia.

Shaohua Yang, Qiaozi Gao, Changsong Liu, Caiming Xiong, Song-Chun Zhu, and Joyce Y Chai. 2016. Grounded semantic role labeling. In *Proceedings of NAACL-HLT*, pages 149–159, San Diego, USA.

Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. 2016. Situation recognition: Visual semantic role labeling for image understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5534–5542, Las Vegas, USA.