Classifying Semantic Clause Types: Modeling Context and Genre Characteristics with Recurrent Neural Networks and Attention

Maria Becker[◇♠], **Michael Staniek**[◇♠], **Vivi Nastase**[◇♠], **Alexis Palmer**[♣], **Anette Frank**[◇♠]

[◇] Leibniz ScienceCampus "Empirical Linguistics and Computational Language Modeling"
[♠]Heidelberg University, Department of Computational Linguistics
[♣]University of North Texas, Department of Linguistics
{mbecker,staniek,nastase,frank}@cl.uni-heidelberg.de
alexis.palmer@unt.edu

Abstract

Detecting aspectual properties of clauses in the form of situation entity types has been shown to depend on a combination of syntactic-semantic and contextual features. We explore this task in a deep-learning framework, where tuned word representations capture lexical, syntactic and semantic features. We introduce an attention mechanism that pinpoints relevant context not only for the current instance, but also for the larger context. Apart from implicitly capturing task relevant features, the advantage of our neural model is that it avoids the need to reproduce linguistic features for other languages and is thus more easily transferable. We present experiments for English and German that achieve competitive performance. We present a novel take on modeling and exploiting genre information and showcase the adaptation of our system from one language to another.

1 Introduction

Semantic clause types, called *Situation Entity (SE)* types (Smith, 2003; Palmer et al., 2007) are linguistic characterizations of aspectual properties shown to be useful for argumentation structure analysis (Becker et al., 2016b), genre characterization (Palmer and Friedrich, 2014), and detection of generic and generalizing sentences (Friedrich and Pinkal, 2015). Recent work on automatic identification of SE types relies on feature-based classifiers for English that have been successfully applied to various textual genres (Friedrich et al., 2016), and also show that a sequence labeling approach that models contextual clause labels yields improved classification performance.

Deep learning provides a powerful framework in which linguistic and semantic regularities can be implicitly captured through word embeddings (Mikolov et al., 2013b). Patterns in larger text fragments can be encoded and exploited by recurrent (RNNs) or convolutional neural networks (CNNs) which have been successfully used for various sentence-based classification tasks, e.g. sentiment (Kim, 2014) or relation classification (Vu et al., 2016; Tai et al., 2015).

We frame the task of classifying clauses with respect to their aspectual properties – i.e., situation entity types – in a recurrent neural network architecture. We adopt a Gated Recurrent Unit (GRU)-based RNN architecture that is well suited to modeling long sequences (Yin et al., 2017). This initial model is enhanced with an attention mechanism shown to be beneficial for sentence classification (Wang et al., 2016) and sequence modeling (Dong and Lapata, 2016). We explore the usefulness of attention in two settings: (i) the individual classification task and (ii) in a setting approximating sequential labeling in which the attention vector provides features that describe the clauses preceding the current instance. Compared to the strong baseline provided by the feature based system of Friedrich et al. (2016), we achieve competitive performance and find that attention as well as context representation using predicted or gold-standard labels of the previous N clauses, and text genre information improve our model.

A strong motivation for developing NN-based systems is that they can be transferred with low cost to other languages without major feature engineering or use of hand-crafted linguistic knowledge resources. Given the highly-engineered feature sets used for SE classification so far (Friedrich et al., 2016), porting such classifiers to other languages is a non-trivial issue. We test the portability of our system by applying it to German.

*Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 230–240,
Vancouver, Canada, August 3-4, 2017.

We present a novel take on modeling and exploiting genre information and test it on the English multi-genre corpus of Friedrich et al. (2016).

Our aims and contributions are: (i) We study the performance of GRU-based models enhanced with attention for modeling local and non-local characteristics of semantic clause types. (ii) We compare the effectiveness of the learned attention weights as features for a sequence labeling system to the explicitly defined syntactic-semantic features in (Friedrich et al., 2016). (iii) We define extensions of our models that integrate external knowledge about genre and show that this can be used to improve classification performance across genres. (iv) We test the portability of our models to other languages by applying them to a smaller, manually annotated German dataset. The performance is comparable to English.

2 Semantic Clause Types

Semantic clause types can be distinguished by the function they have within a text or discourse. We use the inventory of semantic clause types, also known as **situation entity (SE) types**, developed by Smith (2003) and extended in later work (Palmer et al., 2007; Friedrich and Palmer, 2014). SE types describe abstract semantic types of situations evoked in discourse through clauses. As such, they capture the manner of presentation of content, along with the information content itself. The seven SE types we use are described below.

1. STATE (S): *Armin has brown eyes.*
2. EVENT (EV): *Bonnie ate three tacos.*
3. REPORT (R) provides attribution:
   The agency said costs had increased.
4. GENERIC SENTENCE (GEN) predicates over classes or kinds:
   Birds can fly. – Scientists make arguments.
5. GENERALIZING SENTENCE (GS) describes regularly occurring events:
   Fei travels to India every year.
6. QUESTION (Q): *Why do you torment me so?*
7. IMPERATIVE (IMP): *Listen to this.*

An eighth class OTHER is assigned to clauses without an SE label, e.g. bylines or email headers.

Features that distinguish SE types are a combination of linguistic features of the clause and its main verb, and the nature of the main referent of the clause.[1] There is a correlation between the distribution of SE types in text passages and discourse modes, e.g., narrative, informative, or argumentative (Palmer and Friedrich, 2014; Mavridou et al., 2015; Becker et al., 2016a).

3 Related Work

Feature-based classification of situation entity types. The first robust system for SE type classification (Friedrich et al., 2016) combines task-specific syntactic and semantic features with distributional word features, as captured by Brown clusters (Brown et al., 1992). This system segments each text into a sequence of clauses and then predicts the best sequence of SE labels for the text using a linear chain conditional random field (CRF) with label bigram features.[2]

Although SE types are relevant across languages, their linguistic realization varies across languages. Accordingly, some of Friedrich et al. (2016)'s syntactic and semantic features are language-specific and are extracted using English-specific resources such as WordNet and Loaiciga et al. (2014)'s rules for extracting tense and voice information from POS tag sequences.

Friedrich et al. (2016)'s system is trained and evaluated on data sets from MASC and Wikipedia (Section 5), reaching accuracies of 76.4% (F1 71.2) with 10-fold cross-validation, and 74.7% (F1 69.3) on a held-out test set. To evaluate the contribution of sequence information, Friedrich et al. (2016) compare the CRF model to a Maximum Entropy baseline, with the result that the sequential model significantly outperforms the model which classifies clauses in isolation, particularly for the less-frequent SE types of GENERIC SENTENCE and GENERALIZING SENTENCE.

When trained and tested within a single genre (of the 13 genres represented in the data sets), Friedrich et al. (2016)'s system performance ranges from 26.6 F1 (for government documents) to 66.2 F1 (for jokes). Training on all genres levels out this performance difference, with a range of F1 scores from 58.1-69.8.

Neural approaches to sentence classification, sequence and context modeling. Inspired by research in vision, sentence classification tasks have initially been modeled using Convolutional Neural Networks (Kim, 2014; Kalchbrenner et al.,

[1]The main referent of a clause is roughly the person/thing/situation the clause is about, often realized as its grammatical subject.

[2]Code and data: https://github.com/annefried/sitent

2014). RNN variations – with Gated Recurrent Units (GRU) (Cho et al., 2014) or Long Short-Term Memory units (LSTM) (Hochreiter and Schmidhuber, 1997) – have since achieved state of the art performance in both sequence modeling and classification tasks. Recent work applies bi-LSTM models in sequence modeling (PoS tagging, Plank et al. (2016), NER Lample et al. (2016)) and structure prediction tasks (Semantic Role Labeling, Zhou and Xu (2015) or semantic parsing into logical forms Dong and Lapata (2016)). Tree-based LSTM models have been shown to often perform better than purely sequential bi-LSTMs (Tai et al., 2015; Miwa and Bansal, 2016), but depend on parsed input.

Attention. Attention has been established as an effective mechanism that allows models to focus on specific words in the larger context. A model with attention learns what input tokens or token sequences to attend to and thus does not need to capture the complete input information in its hidden state. Attention has been used successfully e.g. in aspect-based sentiment classification (Wang et al., 2016), for modeling relations between words or phrases in encoder-decoder models for translation (Bahdanau et al., 2015), or bi-clausal classification tasks such as textual entailment (Rocktäschel et al., 2016). We use attention to larger context windows and previous labeling decisions to capture sequential information relevant for our classification task. We investigate the learned weights to gain information about what the models learn, and we start to explore how they can be used to provide features for a sequential labeling approach.

4 Models

We aim for a system that can fine-tune input word embeddings to the task, and can process clauses as sequences of words from which to encode larger patterns that help our particular clause classification task. GRU RNNs are used because they can successfully process long sequences and capture long-term dependencies. Attention can encode which parts of the input contain relevant information. These modeling choices are described and justified in detail below. The performance of the models is reported in Section 6.

4.1 Basic Model: Gated Recurrent Unit

Recurrent Neural Networks (RNNs) are modifications of feed-forward neural networks with recurrent connections, which allow them to find patterns in – and thus model – sequences. Simple RNNs cannot capture long-term dependencies (Bengio et al., 1994) because the gradients tend to vanish or grow out of control with long sequences. Gated Recurrent Unit (GRU) RNNs, proposed by Cho et al. (2014), address this shortcoming. GRUs have fewer parameters and thus need less data to generalize (Zhou et al., 2016) than LSTM RNNs, and also outperform the LSTM in many cases (Yin et al., 2017), which makes them a good choice for our relatively small dataset.[3] The relevant equations for a GRU are below. x_t is the input at time t, r_t is a reset gate which determines how to combine the new input with the previous memory, and the update gate z_t defines how much of the previous memory to keep. h_t is the hidden state (memory) at time t, and $\tilde{h_t}$ is the candidate activation at time t. W_* and U_* are weights that are learned. $\odot$ denotes the element-wise multiplication of two vectors.

$$r_t = \sigma(W_r x_t + U_r h_{t-1})$$
$$\tilde{h_t} = tanh(W x_t + U(r_t \odot h_{t-1}))$$
$$z_t = \sigma(W_z x_t + U_z h_{t-1})$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h_t} \quad (1)$$

The last hidden vector h_t will be taken as the representation of the input clause. After compressing it into a vector whose length is equal to the number of class labels (=8) using a fully connected layer with sigmoid function, we apply $softmax$.

4.2 Attention Model

We extend our GRU model with a neural attention mechanism to capture the most relevant words in the input clauses for classifying SE types. Specifically, we adapt the implementation of attention used in Rocktäschel et al. (2016) for our clause classification task as follows:

$$M = tanh(W_h H + W_v h_t \otimes e_L)$$
$$\alpha = softmax(w^T M)$$
$$r = H\alpha^T$$

where H is a matrix consisting of the hidden vectors $[h_1, ..., h_t]$ produced by the GRU, h_t is the last output vector of the GRU, and e_L is a vector of 1s where L denotes the L words of the input clause. $\otimes$ denotes the outer product of the

[3] Comparison of GRUs, bi-GRUs, LSTMs and bi-LSTMs on our dataset for our classification task showed that GRUs outperform the latter three, confirming this assumption.

two vectors. α is a vector consisting of attention weights and r is a weighted representation of the input clause. W_h, W_v, and w are parameters to be learned during training.

The final clause representation is obtained from a combination of the attention-weighted representation r of the clause and the last output vector v.

$$h^* = tanh(W_p r + W_x h_t) \quad (2)$$

where W_p and W_x are trained projection matrices. We convert h^* to a real-valued vector with length 8 (the number of target classes) and apply $softmax$ to transform it to a probability distribution.

4.3 Modeling Context and Genre

Text types differ in their situation entity type distributions: Palmer et al. (2007) find that GENERIC SENTENCES and GENERALIZING SENTENCES play a predominant role for texts associated with the argument or commentary mode (such as essays), and EVENTS and STATES for texts associated with the report mode (such as news texts). (Becker et al., 2016a) find that argumentative texts are characterized by a high proportion of GENERIC and GENERALIZING SENTENCES and very few EVENTS, while reports and talks contain a high proportion of STATES, and fiction is characterized by a high number of EVENTS. N-gram analyses show that sequences of SE types differ among different genres: e.g. while ST-ST is the most frequent bigram within journal articles, the most frequent bigram in Wikipedia articles is GEN-GEN. The most frequent trigram in Jokes is EV-EV-EV, followed by ST-ST-ST, whereas in government documents the most frequent trigrams are ST-ST-ST and EV-ST-ST. These results show that n-grams cluster in texts (cf. (Friedrich and Pinkal, 2015)), and they differ among genres. This supports the choice of incorporating (sequential) context information for classification of SE types. Fig. 1 illustrates both the context and the genre information our models consider for classifying SE types, while Fig. 2 illustrates our model's architecture.

4.3.1 Context Modeling: Clauses and Labels

We develop two models that not only consider the local sentence for SE classification in model training, but also the previous clauses' token sequences or the labels of previous clauses. When attending to **tokens of previous clauses** we add one GRU model with attention mechanism for each previous clause (N denotes the number of previous clauses) and concatenate their final outputs with the final output of the GRU with attention for the current clause (cf. Fig. 2).

$$h^*_{con1} =< tanh(W_p r_1 + W_x v_1); ...; \\ tanh(W_p r_N + W_x v_N) >$$

We then transform the concatenated vector into a dense vector equal to the number of class labels and apply $softmax$.

For attending to **labels of the previous clauses**, we first transform the gold labels used during training into embeddings and apply attention as described in section 4.2 to these representations. We then concatenate the last output of the current clause with the embeddings for the labels of the previous clauses (here N denotes the number of previous labels):

$$h^*_{con2} =< tanh(W_p r + W_x v); y_{t-1}; ...; y_{t-N} >$$

where y_{t-i} is the embedding representation for the previous *t-i* label. At test time we use the predicted probability distribution vector as the labels of the previous clauses.

4.3.2 Feature Modeling: Textual Genres

The English corpus we use consists of texts from 13 genres; the German corpus covers 7 genres (Section 5).

Information about genre is encoded as dense embeddings g of size 10 initialized randomly, and we apply attention mechanism to these representations. Adding genre information produces three new versions of the model: (i) genre+basic model: $< g; h_t >$ (h_t from eq.1), (ii) genre+attention model $< g; h_* >$ (h_* from eq.2), (iii) genre+context in form of previous labels (cf. Fig.2). Results for all three combinations are reported in Section 6.

4.4 Word embeddings

Word embeddings have been shown to capture syntactic and semantic regularities (Mikolov et al., 2013b) and to benefit from fine tuning for specific tasks. The features used by Friedrich et al. (2016) cover a variety of linguistic features – such as tense, voice, number, POS, semantic clusters – some of which we expect to be encoded in pre-trained embeddings, while others will emerge through model training. We start with pre-trained embeddings for both English and German, because this leads to better results than random ini-

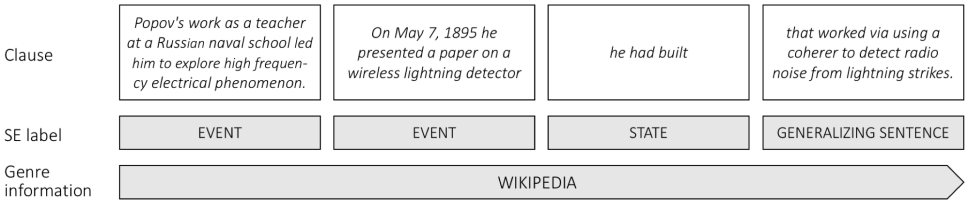

Figure 1: Context and genre information modeled in our system, example from Wikipedia

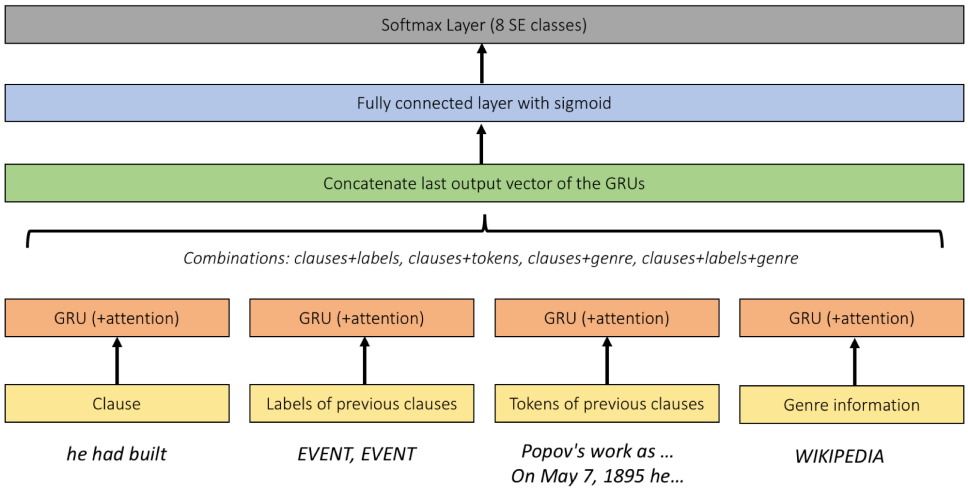

Figure 2: Model Architecture, illustrated with an example (cf. Fig. 1).

tialization. For German, we use 100-dimensional word2vec embeddings trained on a large German corpus of 116 million sentences (Reimers et al., 2014).[4] For English, we use 300-dimensional word2vec embeddings (Mikolov et al., 2013a) trained on part of the Google News dataset (about 100 billion words). The pre-trained embeddings are tuned during training.

4.5 Parameters and Tuning

Hyperparameter settings were determined through exhaustive random search using *optunity* (Bergstra and Bengio, 2012) on the development set, and we use the best setting for evaluating on the test set. We tune batch size, number of layers, GRU cell size, and regularization parameter (L2). For learning rate optimization we use AdaGrad (Duchi et al., 2011) and tune the initial learning rate. For the basic model (without attention), the best result on the development set is achieved for GRU with batch size 100, 2 layers, cell size 350, learning rate 0.05, and L2 regularization parameter (0.01). For the model using attention mechanism the parameters are identical except for L2 (0.0001).

[4]https://public.ukp.informatik.tu-darmstadt.de/reimers/2014_german_embeddings

Data set	# Clauses/SEs	# Tokens
English: MASC	30,333	357,078
English: Wiki	10,607	148,040
German: all	18,194	236,522

Table 1: Data sets with SE-labeled clauses

5 Data

We use the English dataset described in Friedrich and Palmer (2014).[5] The texts, drawn from Wikipedia and MASC (Ide et al., 2010), range across 13 genres, e.g. news texts, government documents, essays, fiction, jokes, emails. For German, we combine two data sets described in Mavridou et al. (2015) and Becker et al. (2016a) and additional data annotated by ourselves.[6] The German texts cover 7 genres: argumentative essays (Peldszus and Stede, 2015), Wikipedia, fiction, commentary, news texts, TED talks, and economic reports. Statistics appear in Table 1.

The distribution of SE types varies with the genre. For the selected English Wiki texts, 50% of the SE types are GENERIC SENTENCE clauses,

[5]Available at: https://github.com/annefried/sitent

[6]The data is available at http://www.cl.uni-heidelberg.de/english/research/downloads/resource_pages/GER_SET/GER_SET_data.shtml. This dataset only contains the German data that has been annotated within the Leibniz Science campus.

	Acc	F1
Palmer07, Brown dataset	53.1	-
Fried16, set A (CRF, test)	69.8	63.9
Fried16, set B (CRF, test)	71.4	65.5
Fried16, set A+B (CRF, test)	74.7	69.3
Fried16, set A+B (CRF, CV)	76.4	71.2
Fried16, set A+B (MaxEnt+CRF, CV, seq-oracle)	77.9	73.9

Table 2: Reported results of baseline models for English (accuracy and macro-average F1 score). CV=10-fold cross validation, test=eval. on test set.

with STATES second at 24.3%.[7] For the 12 MASC genres, STATE is the most frequent type (49.8%), with EVENTS second at 24.3%. GENERIC SENTENCES make up only 7.3% of the SE types in the MASC texts. In the German data, the distributions of SE types also differ according to genre: in argumentative texts, for example, GENERIC SENTENCES make up 48% of the SE types, followed by STATES with a frequency of 32%, while in most other genres the most frequent class is STATE.

The texts of the English dataset are split into clauses using SPADE (Soricut and Marcu, 2003). For segmenting the German dataset into clauses we use DiscourseSegmenter's rule-based segmenter (edseg, Sidarenka et al. (2015)), which employs German-specific rules. Because DiscourseSegmenter occasionally oversplit segments, we did a small amount of post-processing.

6 Experiments and Evaluation

For the English dataset, we use the same test-train split as Friedrich et al. (2016).[8] The German dataset was split into training and testing with a balanced distribution of genres (as is the case for the English dataset). Both datasets have a 80-20 split between training and testing (20% of training is used for development).

We report results in terms of accuracy and macro-average F1 score on the held-out test set.

Baseline systems. The feature-based system of *Palmer07* (Palmer et al., 2007) (Palmer07 in Table 2) simulates context through predicted labels from previous clauses. Friedrich et al. (2016) (*Fried16* in Table 2) report results for their CRF-based SE type labeler for different feature sets, with 10-fold cross validation and on a held-out test set. To test if the context is useful they extend their classifier with a CRF that includes the predicted label of the preceding clause. In the *oracle* setting it includes the gold label of the previous clause.

[7]The Wiki texts were selected by Friedrich et al. (2015) precisely in order to target GENERIC SENTENCE clauses.

[8]The cross validation splits of the data used by Friedrich et al. (2016) are not available.

	Acc	F1
Basic GRU	66.55	46.04
Basic GRU + genre	65.82	46.32
GRU + attention	68.99	68.87
GRU + attention + genre	71.12	67.95
GRU + att + clause (1)	69.06	59.39
GRU + att + clause (2)	**70.20**	**60.01**
GRU + att + clause (3)	69.64	37.29
GRU + att + pLab (1)	69.20	61.95
GRU + att + pLab (2)	**69.37**	**62.13**
GRU + att + pLab (3)	68.77	60.85
GRU + att + pLab (4)	68.05	59.31
GRU + att + pLab (5)	68.11	60.75
GRU + att + pLab + genre (1)	71.59	**64.94**
GRU + att + pLab + genre (2)	**71.61**	64.28
GRU + att + pLab + genre (3)	70.37	63.55
GRU + att + pLab + genre (4)	70.96	63.74
GRU + att + pLab + genre (5)	70.57	63.65

Table 3: SE-type classification on English test set, with context as predicted labels (pLab).

Feature set A consists of standard NLP features including POS tags and Brown clusters. Feature set B includes more detailed features such as tense, lemma, negation, modality, WordNet sense, WordNet supersense and WordNet hypernym sense. We presume that some of the information captured by feature set B, particularly sense and hypernym information, may not be captured in the word embeddings we use in our approach.

Evaluation of our neural systems. Our **local** system (cf. Section 4.1) achieves an accuracy of 66.55 (Table 3). Adding *genre information* does not help, but adding *attention* within the local clause yields an improvement of 2.44 percentage points (pp). Using both *attention and genre* information leads to a 2.13 pp increase over the model that uses only attention. Adding **context information** beyond the local clause – a window of up to three previous clauses – improves the word-based attention models slightly, but a wider window (four or more clauses) causes a major drop

	Acc	F1
GRU + att + gLab (1)	72.71	65.37
GRU + att + gLab (2)	72.68	**66.51**
GRU + att + gLab (3)	72.66	65.03
GRU + att + gLab (4)	72.61	64.33
GRU + att + gLab (5)	**73.40**	66.39
GRU + att + gLab + genre (1)	73.44	**66.76**
GRU + att + gLab + genre (2)	**73.45**	66.51
GRU + att + gLab + genre (3)	72.84	66.29
GRU + att + gLab + genre (4)	73.12	66.21
GRU + att + gLab + genre (5)	73.34	66.13

Table 4: SE-type classification on English test set, *sequence oracle model* using gold labels (gLab).

in accuracy.[9] Using context as predicted labels of previous clauses improves the model slightly (up to 0.38 pp), but adding genre on top of that improves the model by up to 2.62 pp compared to the basic model with attention. The oracle model (cf. Table 4), which uses the gold labels of previous clauses, gives an upper bound for the impact of sequence information: 73.40% accuracy for previous 5 gold labels. Combined with genre information, the upper bound reaches 73.45% accuracy when using the previous 2 gold labels.

The best accuracy on the English data (ignoring the oracle) is achieved by the model that uses 2 previous predicted labels plus genre information (71.61%). This model outperforms Friedrich et al. (2016)'s results when using standard NLP features (feature set A) and their model using feature set B separately. Our model comes close to Friedrich et al.'s best results obtained by applying their entire set of features, particularly considering that our system only uses generic word embeddings.

Window size as hyper-parameter? We achieve best results when incorporating two previous labels or two previous clauses (cf. Table 3). This is in line with Palmer et al. (2007) who report that in most cases performance starts to degrade as the model incorporates more than two previous labels. A window size of two does not always lead to best performance on the German dataset (cf. Section 7), where the model using predicted labels from the maximum window size (5) performs best. When adding genre information, we achieve best results with window size two (cf. Table 5 and 6). This inconsistency can possibly be traced back to the fact that we applied the best-performing variations of our system developed on English data to our German dataset without further hyperparameter tuning.

[9]We achieve 36.24 acc for 4 and 36.17 acc for 5 clauses.

Figure 3: Visualization of attention for ST, GS, GEN, and REP.

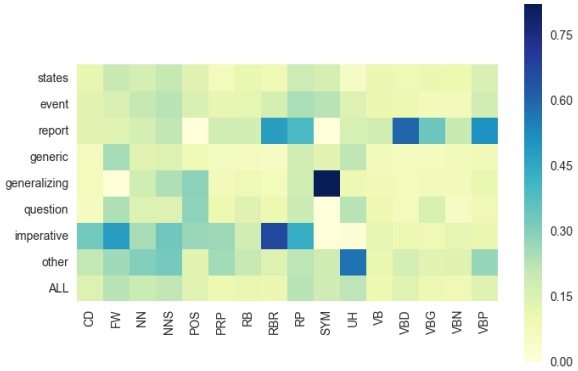

Figure 4: Mean attention scores per POS tags on English dataset. POS tags from PTB.

Results for single classes. Fig. 6 shows macro-average F1 scores of our best performing system for the single SE classes. The scores are very similar to the results of Friedrich et al. (2016). Scores for GENERALIZING SENTENCE are the lowest as this class is very infrequent in the data set, while scores for the classes STATE, EVENT, and REPORT are the highest. In addition, we explored our system's performance for binary classification (Fig. 6): here we classified STATE vs. the remaining classes, EVENT vs. the remaining classes etc. Binary classification achieves better performance and can be helpful for downstream applications which only need information about specific

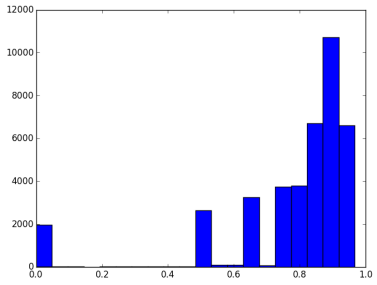

Figure 5: Position of words with maximum attention within clauses. x-axis represents the normalized position within the clause, y-axis the number of words with maximum attention at that position.

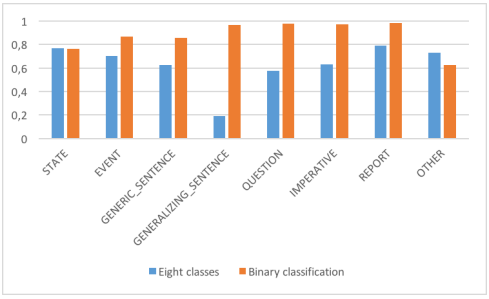

Figure 6: Macro-average F1 scores of our best performing system for single SE classes, multiclass vs. binary classification.

SE types, for example for distinguishing generic from non-generic sentences.

Analysis of attention. Attention is not only an effective mechanism that allows models to focus on specific parts of the input, but it may also enable interesting linguistic insights: (1) the attention to specific words or POS for specific SE types, (2) the overall distribution of attention weights among POS tag labels and SE types, and (3) the position of words with maximum/high attention scores within a clause.

Fig. 3 shows example clauses with their attention weights. In the first clause, a STATE, the model attends most to the nouns "China" and "Japan". In the next clause, a GENERALIZING SENTENCE, the noun "system" is assigned the highest attention weight. The highest weighted word in the GENERIC SENTENCE is the pronoun "their", and in REPORT it is the verb "answered".

Fig. 4 visualizes the mean attention score per POS tag for all SE types (gold labels).[10] Interestingly, attention seems to be especially important for classes that are rare, such as IMPERATIVE or REPORT, each less than 5% of the English dataset. The heat map indicates that the model especially attends to verbs when classifying the SE type REPORT. This is not surprising, since REPORT clauses are signaled by verbs of speech. GENERALIZING SENTENCE attend to symbols, mainly punctuation, and genitive markers such as "'s". The OTHER class, which includes clauses without an assigned SE type label, attends mostly to interjections. Indeed, OTHER is frequent in genres with fragmented sentences (emails, blogs), and numerous interjections such as "wow" or "um".

Fig. 5 shows the relative positions of words with maximum attention within clauses. The model mostly attends to words at the end of clauses and almost never to words in the first half of clauses. This distribution shifts to the left when considering more words with high attention scores instead of only the word with maximum attention – words with 2^{nd} (3^{rd}, 4^{th}, 5^{th}) highest attention score can often be found at the beginning of clauses. The model seems to draw information from a broad range of positions.

We explored the impact of the attention vectors as inputs to a sequence labeling model – each clause is described through the words with the highest attention weights and these weights, and used in a conditional random field system (CRF++[12]). The best performance was obtained when using the attention vector of the current clause (and no additional context) – 61.68% accuracy (47.18% F1 score). CRF++ maps the attention information to binary features, and as such cannot take advantage of information captured in the numerical values of the attention weights, or the embeddings of the given words.

7 Porting the System to German

One advantage for developing NN-based systems that do not rely on hand-crafted features is that they can be used with different language data. We use the system described above with German data, only adjusting the size of the input embeddings.[13] Compared to the English dataset, the German dataset is smaller (44% in size) and less diverse with respect to genre (7 genres). The genres in the German dataset (argumentative texts, wikipedia, commentary, news, fiction, report, talk)

[10] We post-process our data with POS tags using *spaCy*[11] with the PTB tagset (Marcus et al., 1993).

[12] https://taku910.github.io/crfpp/

[13] The different size of the embeddings (for English and German cf. section 4.4, may have an impact on the results.

	Acc	F1
Basic GRU	72.67	61.55
Basic GRU + genre	72.08	66.33
GRU + attention	72.31	72.23
GRU + attention + genre	73.75	65.69
GRU + att + clause (1)	**73.49**	**63.99**
GRU + att + clause (2)	70.21	58.66
GRU + att + clause (3)	49.31	47.01
GRU + att + pLab (1)	69.83	44.31
GRU + att + pLab (2)	70.12	44.33
GRU + att + pLab (3)	70.50	44.91
GRU + att + pLab (4)	72.16	45.12
GRU + att + pLab (5)	**72.85**	**45.52**
GRU + att + pLab + genre (1)	72.19	53.22
GRU + att + pLab + genre (2)	**73.98**	**54.78**
GRU + att + pLab + genre (3)	70.78	46.25
GRU + att + pLab + genre (4)	72.88	48.94
GRU + att + pLab + genre (5)	72.60	45.98

Table 5: SE-type classification on German test set.

	Acc	F1
GRU + att + gLab (1)	71.33	58.32
GRU + att + gLab (2)	72.23	59.43
GRU + att + gLab (3)	73.81	59.12
GRU + att + gLab (4)	75.74	60.39
GRU + att + gLab (5)	**76.32**	**61.01**
GRU + att + gLab + genre (1)	74.79	59.34
GRU + att + gLab + genre (2)	**77.97**	**61.47**
GRU + att + gLab + genre (3)	74.28	59.84
GRU + att + gLab + genre (4)	74.10	59.70
GRU + att + gLab + genre (5)	74.96	58.18

Table 6: SE-type classification on German test set, *sequence oracle model* .

are more similar to one another than the ones in the English dataset. The results comparing the effectiveness of integrating context and genre information are in Table 5. The results of the oracle model using gold labels for previous clauses are in Table 6. Compared to English, the models achieve higher performance, but attention by itself does not improve the results, and neither does the inclusion of genre information. Used jointly, attention and genre information yield a moderate increase of 1.06 pp. accuracy compared to the basic GRU. Attention may need more data and possibly more diversity to be learned effectively, and we will explore this in future work.

Modeling context seems to have a larger impact: compared to the basic GRU using attention, information about the current and the previous clauses improves the model by up to 1.67 pp. More contextual information leads to higher accuracy.

8 Conclusion

We presented an RNN-based approach to situation entity classification that bears clear advantages compared to previous classifier models that rely on carefully hand-engineered features and lexical semantic resources: it is easily transferable to other languages as it can tune pre-trained embeddings to encode semantic information relevant for the task, and can develop attention models to capture – and reveal – relevant information from the larger context. We designed and compared several GRU-based RNN models that jointly model *local and contextual* information in a unified architecture. Genre information was added to model common properties of specific textual genres. What makes our work interesting for linguistically informed semantic models is the exploration of different model variants that combine local classification with sequence information gained from the contextual history, and how these properties interact with genre characteristics. We specifically explore attention mechanisms that help our models focus on specific characteristics of the local and non-local contexts. Attention models jointly using genre and context information in the form of previous predicted labels perform best for our task, for both languages. The performance results of our best models outperform the state of the art models of *Fried16* for English when using either off-the-shelf NLP features (set A) or, separately, hand-crafted features based on lexical resources (set B). A small margin of ca. 3 pp accuracy is left to achieve in future work to compete with the knowledge-rich models of (Friedrich et al., 2016).

Acknowledgments. We thank Sabrina Effenberger, Jesper Klein, Sarina Meyer, and Rebekka Sons for the annotations, and the reviewers for their insightful comments. This research is funded by the Leibniz Science Campus Empirical Linguistics & Computational Language Modeling, supported by Leibniz Association grant no. SAS-2015-IDS-LWC and by the Ministry of Science, Research, and Art of Baden-Württemberg.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR*.

Maria Becker, Alexis Palmer, and Anette Frank. 2016a. Argumentative texts and clause types. In *Proceedings of the 3rd Workshop on Argument Mining*. pages 21–30.

Maria Becker, Alexis Palmer, and Anette Frank. 2016b. Clause Types and Modality in Argumentative Microtexts. In *Workshop on Foundations of the Language of Argumentation (in conjunction with COMMA)*. Potsdam, Germany, pages 1–9.

Y. Bengio, P. Simard, and P. Frasconi. 1994. Learning long-term dependencies with gradient descent is difficult. *Trans. Neur. Netw.* 5(2):157–166. https://doi.org/10.1109/72.279181.

James Bergstra and Yoshua Bengio. 2012. Random search for hyper-parameter optimization. *Journal of Machine Learning Research* 13(Feb):281–305.

Peter F. Brown, Peter V. Desouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. 1992. Class-based n-gram models of natural language. *Computational Linguistics* 18(4):467479.

Kyunghyun Cho, B van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. *On the properties of neural machine translation: Encoder-decoder approaches*.

Li Dong and Mirella Lapata. 2016. Language to logical form with neural attention. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 33–43. http://www.aclweb.org/anthology/P16-1004.

John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research* 12(Jul):2121–2159.

Annemarie Friedrich and Alexis Palmer. 2014. Situation entity annotation. In *Proceedings of the Linguistic Annotation Workshop VIII*.

Annemarie Friedrich, Alexis Palmer, and Manfred Pinkal. 2016. Situation entity types: automatic classification of clause-level aspect. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany, pages 1757–1768. http://www.aclweb.org/anthology/P16-1166.

Annemarie Friedrich, Alexis Palmer, Melissa Peate Sørensen, and Manfred Pinkal. 2015. Annotating genericity: a survey, a scheme, and a corpus. In *The 9th Linguistic Annotation Workshop held in conjuncion with NAACL 2015*. page 21.

Annemarie Friedrich and Manfred Pinkal. 2015. Discourse-sensitive Automatic Identification of Generic Expressions. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL)*. Beijing, China.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.

Nancy Ide, Christiane Fellbaum, Collin Baker, and Rebecca Passonneau. 2010. The Manually Annotated Sub-Corpus: A community resource for and by the people. In *Proceedings of the ACL2010 Conference Short Papers*. pages 68–73.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Baltimore, Maryland, pages 655–665.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar, page 17461751.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, San Diego, California, pages 260–270. http://www.aclweb.org/anthology/N16-1030.

Sharid Loaiciga, Thomas Meyer, and Andrei Popescu-Belis. 2014. English-french verb phrase alignment in europarl for tense translation modeling. In *Proceedings of The Ninth Language Resources and Evaluation Conference (LREC)*.

Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Kleio-Isidora Mavridou, Annemarie Friedrich, Melissa Peate Sorensen, Alexis Palmer, and Manfred Pinkal. 2015. Linking discourse modes and situation entities in a cross-linguistic corpus study. In *Proceedings of the EMNLP Workshop LSDSem 2015: Linking Models of Lexical, Sentential and Discourse-level Semantics*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013a. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. pages 3111–3119.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Atlanta, Georgia, pages 746–751. http://www.aclweb.org/anthology/N13-1090.

Makoto Miwa and Mohit Bansal. 2016. End-to-end relation extraction using lstms on sequences and tree structures. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany, pages 1105–1116. http://www.aclweb.org/anthology/P16-1105.

Alexis Palmer and Annemarie Friedrich. 2014. Genre distinctions and discourse modes: Text types differ in their situation type distributions. In *Proceedings of the Workshop on Frontiers and Connections between Argumentation Theory and NLP*.

Alexis Palmer, Elias Ponvert, Jason Baldridge, and Carlota Smith. 2007. A sequencing model for situation entity classification. In *Proceedings of ACL*.

Andreas Peldszus and Manfred Stede. 2015. An annotated corpus of argumentative microtexts. In *Proceedings of the First European Conference on Argumentation*.

Barbara Plank, Anders Søgaard, and Yoav Goldberg. 2016. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 412–418. http://anthology.aclweb.org/P16-2067.

Nils Reimers, Judith Eckle-Kohler, Carsten Schnober, Jungi Kim, and Iryna Gurevych. 2014. Germeval-2014: Nested Named Entity Recognition with neural networks. In *Proceedings of the 12th Edition of the KONVENS Conference*. page 117120.

Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomáš Kočiskỳ, and Phil Blunsom. 2016. Reasoning about entailment with neural attention. In *Proceedings of the 4th International Conference on Learning Representations (ICLR)*. San Juan, Puerto Rico.

Uladzimir Sidarenka, Andreas Peldszus, and Manfred Stede. 2015. Discourse Segmentation of German Texts. In *Journal for Language Technology and Computational Linguistics*. volume 30, pages 71–98.

Carlota S Smith. 2003. *Modes of discourse: The local structure of texts*, volume 103. Cambridge University Press.

Radu Soricut and Daniel Marcu. 2003. Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology.*.

Kai Sheng Tai, Richard Socher, and Christopher D. Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Beijing, China, pages 1556–1566. http://www.aclweb.org/anthology/P15-1150.

Ngoc Thang Vu, Heike Adel, Pankaj Gupta, and Hinrich Schütze. 2016. Combining recurrent and convolutional neural networks for relation classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California, pages 534–539. http://www.aclweb.org/anthology/N16-1065.

Yequan Wang, Minlie Huang, xiaoyan zhu, and Li Zhao. 2016. Attention-based LSTM for Aspect-level Sentiment Classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, pages 606–615. https://aclweb.org/anthology/D16-1058.

Wenpeng Yin, Katharina Kann, Mo Yu, and Hinrich Schütze. 2017. Comparative study of cnn and rnn for natural language processing. *CoRR* abs/1702.01923.

Jie Zhou, Ying Cao, Xuguang Wang, Peng Li, and Wei Xu. 2016. Deep recurrent models with fast-forward connections for neural machine translation. *Transactions of the Association for Computational Linguistics* pages 371–383.

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